Running head: SITUATION MODELS AND VALIDATION

Getting a picture that is both accurate and stable: Situation models and epistemic validation

Sascha Schroeder, Tobias Richter, & Inga Hoever

University of Cologne

accepted for publication in *Journal of Memory and Language*

Corresponding author:

Tobias Richter

University of Cologne, Department of Psychology

Bernhard-Feilchenfeld-Str. 11

50969 Köln, Germany

Phone: +49 (221) 470-2773, Fax: +49 (221) 470-5002

Email: tobias.richter@uni-koeln.de

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, grant RI1100/2-2 given to Tobias Richter and Norbert Groeben). We would like to thank Britta Wöhrmann for her help in preparing the text material and collecting data. The texts and item materials (in German) that we used in this study are available from the corresponding author upon request.

Abstract

Text comprehension entails the construction of a situation model that prepares individuals for situated action. In order to meet this function, situation model representations are required to be both accurate and stable. We propose a framework according to which comprehenders rely on epistemic validation to prevent inaccurate information from entering the situation model. Once information has been integrated in the current situation model, it serves as part of the epistemic background for validating new information, leading to a stable representation. We present evidence for this view from an experiment in which participants responded to paraphrase and inference items after reading expository texts. Multinomial model analyses of the responses and multilevel analyses of the response latencies revealed that plausible information is more likely to be integrated into the situation model while information that is part of the situation model is more likely to be judged as plausible. This pattern of results suggests a close bi-directional relationship between situation models and epistemic validation.

Key words: situation model, text comprehension, validation, verification

Getting a picture that is both accurate and stable: Situation models and epistemic validation

The assumption that readers construct a situation model as a referential representation of the state of affairs described in a text in a quick und effortless way is now a commonplace assumption in the psychology of language (Zwaan & Radvansky, 1998). However, the ultimate goal of situation model construction and its consequences for comprehension are still a point of contention. In this article, we start from the assumption that situation models serve the extra-linguistic purpose to enable comprehenders to interact with the world (Glenberg, 1997). In order to fulfill this function, situation models are required to be both accurate and stable representations of the actual state of affairs. We will suggest that comprehenders achieve accurate representations by using their world knowledge to validate text ideas before integrating them into their situation model of the text content. Stable representations are achieved by using the situation model itself to validate incoming text information. This framework implies that the likelihood for a particular piece of information to become part of the situation model depends on its plausibility. Moreover, integration of information into the situation model may be expected to increase its subjective plausibility. We will present results from an experiment that tested the hypothesized relationships of situation models and plausibility. Participants read extensive expository texts, and multinomial models were applied to comprehension and plausibility judgments collected with a modified version of the recognition method proposed by Schmalhofer and Glavanov (1986). The multinomial model results were cross-validated by an analysis of response times.

Situation Models Need to be Accurate and Stable

From the perspective of grounded language comprehension, it is plausible to assume that the ultimate goal of comprehension is not to acquire a coherent meaning representation (e.g. Kintsch, 1988) but to prepare individuals for situated action (e.g., Glenberg, 1997). According to this view, situation models are representations that enable comprehenders to use communicated

information to interact with the world (e.g, Zwaan, 1999). In order to meet this function, a situation model needs to have at least two important properties that seem to be partially incompatible at first sight. First, a situation model needs to represent the state of affairs described in a text as accurately as possible (criterion of truth). Second, situation models need to be sufficiently stable in order to allow successful interactions with the world (criterion of stability). We will discuss these two criteria in turn.

The relevance of the *criterion of truth* for comprehension can be illustrated by looking at everyday comprehension situations. In most of these situations, comprehension is not an end in itself, but part of more broadly defined actions that require adequate representations of real-world situations. Consider, for example, a situation model that an individual constructs while reading a user's manual of a technical appliance. This situation model will be useful to the extent that it adequately reflects the functionality of the appliance. More generally, informational or expository texts are usually read with the proximal goal of knowledge acquisition, i.e. the construction of internal representations that approximate the criterion of truth. Knowledge in terms of accurate representations, in turn, is an important prerequisite of goal-directed action.

The idea that comprehension critically depends on the acquisition of accurate representations may seem commonplace but it provides a perspective that is largely new to the psychology of language and text comprehension. Starting with the pioneering work of Kintsch and van Dijk (1978), most of the theories developed in the area of text comprehension, for example, have put their focus on coherence relations. Consequently, these theories have posited the construction of an internally coherent representation as the ultimate goal of comprehension. Compared to coherence, the correspondence of text ideas to states of affairs in the world and its role in comprehension have received relatively little attention (Long & Lea, 2005). A related problem is the way how theories of text comprehension have conceptualized the use of

knowledge in comprehension. Across the board, bottom-up, text-driven models such as the Construction-Integration model (Kintsch, 1988) as well as top-down models such as schema theory (Bransford & Johnson, 1972) or the constructionist theory of inferences (Graesser, Singer, & Trabasso, 1994) assume the primary function of knowledge to be a supplement to the information explicitly provided by a text. In particular, it has been suggested that knowledge aids the interpretation of incoming information and provides a scaffold for its integration or a knowledge base for inferences. In contrast, the idea that knowledge might be used to validate text information in order to construct a situational representation that approximates the criterion of truth is not covered by major theories of language and text comprehension. One notable exception is the theory of mental models. This theory assumes that new information is checked for consistency with other elements and relations in the current mental model before it is integrated into the model (Johnson-Laird, 1983, p. 249). So far, however, this aspect of the theory of mental models has not attracted any systematic research in the area of text comprehension. To conclude, the question whether situation models are accurate representations of the actual state of affairs and the related question of the role of knowledge-based validation are relevant in everyday comprehension situations. They are also reminiscent in many studies on text comprehension on inconsistency effects, which will be reviewed later. Nonetheless, these questions have largely been ignored by the dominant theoretical proposals in the area.

The relevance of the *criterion of stability* for comprehension also becomes apparent when the potential use of situation models for action is taken into account. If new information constantly prompted individuals to change their worldview, they would be unable to engage in goal-directed action (Dreisbach & Goschke, 2004). The stability of representations is not a new topic in the psychology of language. Traditionally, it has been solved either by assuming that stability is achieved by relatively inflexible knowledge structures such as schemata that guide

comprehension (Bransford & Johnson, 1972). Alternatively, it has been suggested that new propositions are being integrated into a large and therefore inert associative network with existing link strengths (Kintsch, 1988). Here, we take a different approach by assuming that both the criterion of truth and the criterion of stability are tied to the validation of incoming information.

A Framework for Epistemic Validation Processes

How do comprehenders manage to achieve both accurate and stable representations? We suggest that they carry out epistemic validation processes that monitor whether incoming information is consistent with other ideas provided in the text, with the current state of the situation model, and with general world knowledge. We assume that these validation processes are routinely carried out when situation models are updated and that they are a major determinant of whether a particular piece of information is integrated into the situation model, with the potential consequence of altering a comprehender's world view. It seems plausible to assume that epistemic validation rests on two component processes that may be termed epistemic monitoring and epistemic elaboration. These two types of processes are linked to the distinction of memory-based and explanation-based processes in comprehension.

Epistemic monitoring processes check for inconsistencies between incoming text information on the one hand and elements of the current situation model or world knowledge retrieved from long-term memory on the other hand. We expect that epistemic monitoring processes are carried out routinely and require relatively little cognitive effort. This is because they refer to information that is already part of working memory, such as elements of the currently active situation model, or to elements of long-term memory that can easily be made available by memory-based processes (e.g., McKoon & Ratcliff, 1995; Myers & O'Brien, 1998; O'Brien & Albrecht, 1992). Thus, in normal reading that is not governed by specific processing goals, inconsistencies of incoming text information with information that is active in working

memory or easily accessible, contextually relevant world knowledge are noticed whereas inconsistencies with information that is less salient may go unnoticed. *Epistemic elaboration* processes may become operative when an inconsistency has been detected. These processes elaborate hypothetical truth conditions of the incoming information that has caused the inconsistency. In contrast to epistemic monitoring, epistemic elaboration is based on processes that are resource-demanding and strategic, i.e. their scrutiny depends on specific processing goals of the reader (such as the goal to develop a justified point of view). Ultimately, epistemic elaboration may lead to a conscious decision about whether a particular piece of information is accepted as being valid or whether it is rejected as being invalid.

The exact nature of this decision and the criteria underlying the acceptance or rejection of incoming information are beyond the scope of this article (but see Johnson-Laird, Girotto, & Legrenzi, 2004, for a detailed proposal). At this point, it is sufficient to make the simplified assumption that plausible information has a higher likelihood of passing the routine validation check and, thus, a higher likelihood of becoming part of the situation model than implausible information. However, the extent and intensity of epistemic validation is limited by the necessity to uphold the criterion of stability. If every small inconsistency between the already existing situation model and incoming information brought about the possibility of its complete revision, the stability and coherence of the representation would be constantly endangered. For this reason, we assume that once new information has passed the epistemic gatekeeper and has been successfully integrated into the situation model, it is suspended from further validation unless drastic discrepancies occur. Moreover, information that has been integrated into the situation model now becomes part of the epistemic background that comprehenders use to validate new incoming information. Per default, this information is favored in case of a conflict between old and new information, with the consequence that the criterion of stability is maximized.

Evidence for a Bi-directional Relationship of Comprehension and Validation

The assumptions of a partial trade-off between maximizing the criteria of truth and stability and the bi-directional perspective on the relationship between epistemic validation and situation model construction go well beyond the traditional focus on coherence that dominates much of text comprehension research. Nevertheless, the framework suggested here may help to elucidate some seemingly contradictory findings from psycholinguistics, text comprehension research, and social psychology.

On the one hand, research on language comprehension provides ample evidence that individuals indeed engage in validation processes during situation model construction. Provided that they have access to relevant background knowledge, comprehenders routinely validate information even when working memory resources are depleted or the task does not require the validation of information (Richter, Schroeder, & Wöhrmann, 2006). In one experiment by Richter et al. (2006), for instance, participants performed an orthographical judgment task on individual words that were presented one by one on a computer screen. Sequences of words formed simple assertions. These assertions were either true (e.g., *Fire trucks are red*) or false (e.g., *Soft soap is edible*). The critical trials were those in which the target word was the last word of an assertion. In these trials, response latencies and error rates of the orthographical judgments were increased when the task required an affirmative response (i.e., the last word was spelled correctly) but the assertion was false. Thus, there was a Stroop-like interference effect suggesting that individuals automatically monitor the validity of information. Following a different approach, Singer (2006) has demonstrated with a reading time paradigm that people regularly verify incoming information against passively cued information during reading. In Singer's experiments, participants read stories that contained a target sentence (e.g., *The policeman knew/implied that the vehicle with the flat was/wasn't a truck*) that was either congruent or

incongruent with the state of affairs introduced by earlier story sentences (e.g., *Dan drove past a bus which was stopped with a flat tire*). Reading times for affirmative target sentences were prolonged when discourse context and pertinent knowledge rendered the target sentence false. Moreover, when the target sentence contained a factive verb (e.g., *knew*) as opposed to a non-factive verb (e.g., *implied*), the reading times for the target sentence followed an interaction of truth and negation that is typical for intentional sentence verification tasks. Thus, individuals validate information they encounter in a discourse context, and they seem to be sensitive to pragmatic cues that signal the epistemic status of the communicated information. In a similar vein, earlier research on causal and other types of bridging inferences has shown that these types of inferences do not only involve the activation but also the validation of inferred bridging information (Singer, Halldorson, Lear, & Andrusiak, 1992; Singer, 1993). For example, after reading the causal sequence *Dorothy poured the bucket of water into the fire – The fire went out*, participants were faster to verify questions such as *Does water extinguish fire?*, compared to a control condition in which temporal sequences were presented first.

Readers are also able to detect logical inconsistencies in a text when a logical relationship is signaled or the premise information is held active in working memory (Lea, 1995; Lea, Mulligan, & Walton, 2005). Lea et al. (2005) had participants read stories that contained two logical premises that were either separated by one sentence or by an intervening passage of ten sentences. At a later point in the story, a target sentence was presented that was locally coherent but inconsistent with the information provided by the two premises. When the two premises were close to one another in the story or when a contextual cue reactivated the distant premise, reading times for the inconsistent target sentence were prolonged, indicating that participants noticed and tried to fix the logical inconsistency. Similarly, inconsistencies of other types of situational information such as those between protagonists' actions and their previously described goals in a

narrative (Albrecht & Myers, 1995) or inconsistencies in information about a protagonist's location (O'Brien & Albrecht, 1992) are usually noticed when the antecedent information is still active in working memory. In sum, a large body of research on various aspects of comprehension provides evidence for the general assumption that readers regularly monitor the internal consistency and truthfulness of information in a text. The studies reviewed here also provide a (non-exhaustive) set of examples for the types of inconsistencies that are detected by epistemic monitoring processes. The inconsistencies that comprehenders seem to detect routinely include outright contradictions to elements of general world knowledge (Richter et al., 2006), false implicit premises (enthymemes) of causal relationships (Singer et al., 1992), violations of logical rules (Lea et al., 2005), and contradictions to various aspects of the current situation model (Albrecht & Myers, 1995; O'Brien & Albrecht, 1992; Singer, 2006).

In apparent contrast to these results, studies from various areas of research have demonstrated that individuals stick to false or insufficiently justified information even after it has been explicitly corrected. Experiments based on the debriefing or misinformation paradigm, for example, have shown that information given earlier in an experiment remains part of a comprehender's event representation even if this information has been explicitly contradicted or discredited by information provided later (continued influence of misinformation effect, Ross, Lepper, & Hubbard, 1975; Johnson & Seifert, 1994). The experiments by Johnson and Seifert (1994), for example, used fictional news reports as text material. The reports were continuously updated during the experiment. In the course of the updating, information that had been given earlier in the experiment was corrected by information that was given at a later point. Johnson and Seifert (1994) found effects of the outdated information on later inferences even if the correcting information was given immediately after the corrected information. Moreover, the corrected information maintained its influence only if it was embedded in the causal chain of the

news report but not if it was circumstantial information that was not part for the causal chain. Results from earlier research in social psychology point into a similar direction. Anderson, Lepper, and Ross (1980) demonstrated that the effects of discredited information on later judgments were particularly pronounced if participants had been instructed to infer causal explanations for this information before the discrediting information was given. In another study, the generation of knowledge-based inferences and explanations enhanced the subjective likelihood of events described in clinical case reports despite the fact that participants were told that the described events were purely fictitious (Ross, Lepper, Strack, & Steinmetz, 1977). These findings are particularly informative for the relationship of validation and situation model construction because being embedded in a causal chain, causal explanations, and knowledge-based inferences may all be assumed to increase the probability that a particular piece of information is integrated into a situation model representation.

In sum, despite the fact that comprehenders seem to validate incoming information in the course of situation model construction, they are also prone to stick to invalid information. The framework outlined here is able to resolve the contradiction between these two lines of research. The experiments suggesting a close link between comprehension and validation are consistent with the assumption that comprehenders perform epistemic monitoring processes on incoming information in order to arrive at an accurate situation model representation. However, once information has passed the epistemic gatekeeper it forms part of the background against which new information is validated. In order to ensure a stable situation model, information already integrated into the situation model is no longer subject to validation processes.

Situation Models and Epistemic Validation: A Multinomial Model

Rather than looking at epistemic validation processes directly, the present research was aimed at the representational outcomes of these processes. The framework outlined here implies

two predictions concerning the representational outcomes that result when individuals process texts that contain plausible as well as implausible information. First, the assumption that comprehenders validate incoming information in updating situation models implies that plausible information will be more likely to be incorporated in the situation model compared to implausible information (1). Second, the situation model itself is no longer subject to validation but serves as epistemic background for validating new information. Consequently, information that is part of the situation model should be more likely to be accepted as true (2). As an aspect of grounded language comprehension, epistemic validation is related specifically to comprehension on the level of the situation model. For this reason, we expected the hypothesized relationships between comprehension and validation to hold for situation model representation but not for the propositional textbase, i.e. the memory for explicit text information (e.g., Kintsch, 1988).

We tested these assumptions with an experiment based on naturalistic texts that contained plausible and implausible information. A modified version of the recognition method proposed by Schmalhofer and Glavanov (1986) was used to derive differential comprehension measures for situation model and propositional textbase and to determine their relationships with plausibility judgments. Participants were presented with test items that corresponded to either the textbase representation (paraphrases), the situation model representation (inferences), or neither the textbase nor the situation model (distracter items). Paraphrase and inference test items could be either plausible or implausible. In contrast to the original method proposed by Schmalhofer and Glavanov (1986), participants provided three different yes/no-responses to each test item rather than global recognition judgments. For each item, they judged (1) whether the information contained in an item was explicitly stated in the text (*textbase question*), (2) whether the information matched the contents of the text (*situation model question*), and (3) whether the information was plausible or not (*plausibility question*). Accordingly, there were eight (2X2X2)

possible patterns of responses for each test item.

These response patterns were analyzed by using multinomial models rather than the signal detection analysis proposed by Schmalhofer and Glavanov (1986). Multinomial models are a class of statistical tools for estimating and testing formal models of latent cognitive states that are assumed to have caused observed response patterns in cognitive tasks. These models are designed for testing detailed hypotheses about cognitive constructs and their relationships (Riefer & Batchelder, 1988).

The predictions concerning situation model construction and epistemic validation were tested on the basis of the models depicted in Figure 1. These models contain three types of theoretically relevant parameters that reflect different aspects of the text representation: One parameter (t) corresponds to the propositional textbase, two parameters ($s \mid t+$ and $s \mid t-$) correspond to the situation model, and four parameters ($e \mid t+s+$, $e \mid t+s-$, $e \mid t-s+$, and $e \mid t-s-$) correspond to the degree of epistemic understanding. These parameters were estimated separately for each item type. The model in Figure 1a was used for the plausible test items (plausible paraphrases and plausible inferences). The model in Figure 1b was used for the implausible test items (implausible paraphrases and implausible inferences). As will be explained later, the models for plausible and implausible test items are identical except for the way the e-parameters were assessed. In addition to the theoretically relevant parameters, both models contain three types of bias parameters, $g(t)$, $g(s)$, and $g(e)$. These parameters were estimated from responses to the distracter items in order to correct the estimates of the theoretically relevant parameters for guessing. The rationale of how the bias parameters were estimated will be clarified at the beginning of the results section. The remainder of this section focuses on describing the theoretically relevant parameters in Figure 1 and how these parameters relate to our hypotheses.

Textbase parameters. All parameters in the models depicted in Figure 1 reflect simple or

conditional probabilities for latent cognitive states. The textbase parameter *t*, for example, represents the probability that a particular type of information has been integrated into the propositional textbase, as indicated by yes-responses to the textbase question (*T*+). Accordingly, (1-*t*) represents the complementary probability that a particular type of information is not part of the propositional textbase. However, even if an item had not been recognized to belong to the propositional textbase it is possible that participants arrive at the correct response by guessing. This tendency is captured by the bias parameter *g*(*t*), which represents the probability of affirming the textbase question (*T*+) regardless of whether it is actually part of the textbase representation or not. Only if both textbase recognition and the corresponding guessing process yield a negative outcome, participants are assumed to give no-responses to the textbase question (T-). The textbase parameter *t* was expected to differ between paraphrase and inference items, with paraphrase items having a higher probability to be recognized as explicitly mentioned in the text. However, this parameter should be independent of the plausibility of a given test item because validation processes were not assumed to influence the construction of the propositional textbase. For this reason, we did not expect to find any differences in the *t*-parameters for plausible and implausible items.

Situation model parameters. The situation model parameters *s* reflect conditional probabilities that the information contained in a test item has been integrated into the situation model, as indicated by a yes-response to the situation model question (*S*+). There are two *s*-parameters because the model allows the probabilities of situation model integration to vary depending on whether a piece of information is part of the textbase representation or not ($s \mid t+$ and $s \mid t-$). Similar to the textbase parameter, both situation model parameters have corresponding bias parameters, $g(s \mid t+)$ and $g(s \mid t-)$, in order to correct for guessing. Again, only if both situation model recognition and situation model guessing yielded a negative outcome, no-

responses to the situation model question were expected (S-). The *s*-parameters are central to our first prediction. This prediction states that implausible information is less likely to be integrated into the situation model even if it matches the situation described in the text. Accordingly, the *s*-parameters for plausible items should be higher than those for implausible items. Moreover, this prediction should hold for paraphrase as well as inference items because both types of items contain information that may become part of the situation model representation.

Epistemic understanding parameters. The *e*-parameters capture participants' ability to judge plausible test items as being plausible (Figure 1a) or participants' ability to detect implausible test items as being implausible (Figure 1b). Thus, in the model for plausible test items, the *e*-parameters are conditional probabilities of yes-responses to the plausibility question (P+). In the model for implausible test items, in contrast, the *e*-parameters are conditional probabilities of no-responses to the plausibility question (P-) (cf. Klauer, Musch, & Naumer, 2000, for a similar approach to model belief bias in syllogistic reasoning). Consequently, the complementary probabilities (1-*e*) reflect the tendency of judging implausible information as being plausible. Due to the fact that the probabilities captured by the *e*-parameters may vary depending on the previous classification decisions, there are four *e*-parameters in each model, one for each combination of yes- and no-responses to the textbase and the situation model question ($e \mid t+s+$ to $e \mid t-s-$). Again, each of these parameters is accompanied by one corresponding bias parameter, $g(e \mid t+s+)$, $g(e \mid t+s-)$, $g(e \mid t-s+)$, and $g(e \mid t-s-)$ to correct for guessing. Our prediction concerning epistemic understanding states that information that has been integrated into the situation model is more likely to be accepted as true compared to information that has not been integrated into the situation model. Accordingly, we expected the *e*-parameters for plausible items that were part of the situation model ($e \mid s+$) to be higher than the e-parameters for items that were not part of the situation model ($e \mid s-$). Likewise, for the implausible items, the

complementary probabilities (1-*e*) in the models of the implausible items were expected to be higher for items that were part of the situation model (1-*e* | *s*+) compared to those items that were not part of the situation model (1-*e* | *s*-).

Cross-validation by response latencies. We cross-validated the results of the multinomial model analyses with an analysis of the latencies of responses to the textbase, situation model, and plausibility questions. For the latency data, we expected a pattern of results that should be strictly parallel to the response data. First, the assumption that the construction of the situation model partly rests on epistemic validation processes implies that the judgment about whether a particular item matches the situation described in a text would be facilitated for plausible items. Therefore, responses should be faster for plausible as compared to implausible items. Second, the assumption that contents of the situation model itself serve as epistemic background for the validation of incoming information implies that responses to the plausibility question should be facilitated for plausible information that had already been integrated into the situation model. Implausible information, in contrast, should be more easily rejected if it is not part of the situation model. Consequently, we predicted a crossover interaction of item plausibility and situation model integration for the response latencies to the plausibility question.

Method

Participants. Seventy psychology undergraduates (61 women and 9 men) took part in the experiment, with an average age of 22.4 years (*SD* = 6.1).

Text material. The experimental texts were two expository texts similar to those typically read by psychology undergraduates (see Appendix A for a text sample). One text was about social influences on interpersonal attraction (4,361 words; adapted from Thomas, 1991), and the other text dealt with theories on smoking behavior (3,714 words; adapted from Fuchs & Schwarzer, 1997). In each text, 30 sentences were modified by incorporating one of five common

argumentation errors (Dauer, 1989; see Table 1 for examples). These modifications rendered the component statements of the sentence implausible by weakening their justification while preserving their meaning and propositional content. In the text on smoking behavior, for instance, the sentence *If the father shows that his cigarette after dinner tastes wonderful the children develop a positive attitude towards smoking* was made implausible by exchanging cause and effect: *If the children develop a positive attitude towards smoking the father shows that his cigarette after dinner tastes wonderful*. It is important to note, however, that none of the implausible sentences was syntactically or semantically anomalous. In addition, both the plausible and the implausible versions of the focal sentences were coherent with previous discourse context and thus congruent with the current state of the situation model. The only difference between plausible and implausible sentences was that the justification of the argument presented in implausible sentences was weak or defective.

Plausible and implausible sentences were selected by a quasi-random procedure, with the constraints that (a) implausible sentences should not follow directly after another implausible sentence and (b) no more than three sentences on each text page should be implausible (there were approximately 15-20 sentences on each text page). Consequently, both types of sentences were comparable in features such as length or semantic complexity: Plausible and implausible sentences had a mean length of 2.8 clauses and had similar readability scores (38.8 for the plausible sentences vs. 33.2 for the implausible sentences as indexed by the German adaptation of Flesch's Reading Ease Index, cp. Amstad, 1978). In addition, the mean frequency class of content words did not differ between plausible and implausible sentences (plausible sentences: 11.1, implausible sentences: 11.4; frequencies based on the standard corpus of Deutscher Wortschatz Leipzig, 2008).

Test items. Experimental test items were constructed on the basis of the 30 implausible

sentences and 30 plausible sentences for both texts. For each of the 60 sentences, one paraphrase item and one inference item were constructed (see Table 1 for examples).

Paraphrase items were derived by replacing the key words of the sentence with synonyms and scrambling the syntactic constituents. For instance, the sentence *Smoking is a goal-directed, intentional activity that fulfills specific functions* was changed into *Smoking serves particular functions and is a purposeful and goal-directed activity*. This procedure changed the surface structure of the sentence but left its propositional representation intact.

Inference items, in contrast, did not correspond to information explicitly mentioned in the text. Rather, they reflected information that readers with average prior knowledge were likely to add to the text information. For example, when comprehenders read the sentence *The concept of nicotine sensitivity was introduced to explain why some people do not become addicted to nicotine even if they smoke a lot of cigarettes* it is likely that they construct the implicit premise *If someone reacts sensitive to nicotine than he or she is more likely to develop an addiction to cigarettes* as a bridging inference (cf. Singer et al., 1992). While this inference shares only minimal propositional content with the original sentence it is an important aspect of the state of affairs described in the text. As such, it is likely to be integrated into the situation model.

In addition to the experimental items, 30 distracter items for each text were constructed by selecting sentences from unrelated texts, which shared at least one key concept with the experimental texts. For example, whereas the text on interpersonal attraction mentioned Zajonc's mere exposure hypothesis but did not elaborate any further on it, one distracter item stated that the mere exposure hypothesis was tested in an experiment involving Chinese characters. This information was neither explicitly mentioned in the text nor did it correspond to a likely inference. Thus, the distracter items bore a superficial resemblance to the contents of the text but were not part of the textbase or the situation model.

For both texts, two test item lists were created that contained either the paraphrase or the inference item for one of the original sentences. As a consequence, every test item list contained 15 plausible paraphrase and 15 plausible inference items as well as 15 implausible paraphrase and inference items respectively. The same set of 30 distracter items was used in both lists.

Validation of text and item materials. We conducted two norming studies to validate text and item materials. The goal of the first norming study was to ensure that plausible and implausible assertions did indeed differ in their plausibility. Four graduate students of psychology were asked to rate the original sentences (30 plausible and 30 implausible sentences from each text) and the test items derived from these sentences (30 plausible and 30 implausible paraphrase items, and 30 plausible and 30 implausible inference items for each text) on a 6-point-scale (ranging from 0=*not convincing* to 5=*very convincing*). The ratings were very homogenous ($ICC(337,1011)=.86$, $p < .001$), making it possible to aggregate the ratings across the four experts. The aggregated ratings served as the dependent variable in a by-items ANOVA with intended item plausibility as independent variable. Implausible text passages and test items were judged as far less convincing ($M=0.74$, $SE_M=0.07$) than plausible text passages and items ($M=3.44$, $SE_M=0.07$), $F(1,358)=844.9$, $p<.001$, $\eta_p^2=.70$.

The goal of the second norming study was to determine whether paraphrase and inference items were distinctive. Three graduate students who were familiar with the theoretical distinction between propositional textbase and situation model were asked to compare the test items to the original sentences and (a) indicate whether this item is a paraphrase or an inference item and (b) rate the confidence in their judgment on a 6-point-scale (ranging from 0=*not certain* to 5=*very certain*). Inference responses were multiplied by -1 in order to create a combined scale for inference and paraphrase responses (ranging from -5=*certain inference* to +5=*certain paraphrase*). Again, the ratings were very homogenous ($ICC(238,476)=.83$, $p < .001$), making it

possible to aggregate the ratings across the three experts. The aggregated ratings served as the dependent variable in a by-items ANOVA with intended item type and item plausibility as independent variables. A significant main effect for item type ($F(1,236)=351.74$, $p<.001$, $\eta_p^2=.60$) indicated that paraphrase ($M=3.63$, $SE_M=0.21$) and inference items ($M=-1.92$, $SE_M=0.21$) were distinctive. Neither the main effect of item plausibility nor the interaction of item plausibility and item type reached significance (for both tests: $F\leq 1$).

Procedure. The experimental procedure consisted of a reading phase and a test phase for each text. Participants first read the text paragraph by paragraph on a computer screen in a self-paced fashion. They were instructed to read the text thoroughly for understanding. Participants were also told that they would receive questions about the text in a later phase of the experiment. After reading the text, the test phase started. The experimental items were presented to the participants one-by-one in black letters (height 2 cm, font type arial) in a white 27 X 6 cm square placed at the top of the screen against a green background. The viewing distance was approximately 60 cm. Participants were asked to read the item carefully and press a response key when they had understood it correctly. After that, three questions (textbase question, situation model question, and plausibility question) were displayed one after the other in the area below the item. The wordings of the three questions (and the wordings of the additional instructions given before the block of questions) were as follows:

> *1. Was the information that is included in the assertion also explicitly mentioned in the text?* By asking this question, we would like to know from you whether the assertion that is presented to you was mentioned in the text. For a positive answer, it is not necessary that the assertion has been taken verbatim from the text. However, it should correspond directly to the contents of one of the sentences of the text.
>
> *2. Does the assertion match the contents of the text?* By asking this question, we are

interested in whether you would associate the presented assertion with the contents of the text. For a positive answer, it is not necessary that the assertion was actually included in the text. However, the assertion should be in line with the contents of the text or it should be possible to infer the assertion from the contents of the text.

3. Is the assertion convincing? By asking this question, we would like to know whether you find the assertion plausible. Aspects that can render an assertion implausible include mistakes in the definition of concepts, faulty conclusions, or backing by weak arguments. Thus, in judging the plausibility of an assertion, you should also consider how convincing the assertion is in the light of pertinent background information.

The screens displaying the three questions were separated by blank screens that appeared for 500 ms each. Participants were instructed to respond to each of the three questions as quickly and accurately as possible by pressing one of two response keys (marked *yes* and *no*). We recorded the responses as well as the response latencies. Before the actual test trials, there was one practice item to familiarize participants with the task.

Design. The design was a 2(*item plausibility*: plausible vs. implausible) X 2(*item type*: paraphrase vs. inference) design with repeated measurements on both variables. The assignment of test item lists, the assignment of response keys to yes- and no-responses, and the order of the experimental texts was counterbalanced across participants. For each participant, the items were presented in random order.

Results and Discussion

Multinomial Model

Stability of response pattern across texts. The frequencies of responses to the textbase question, situation model question, and the plausibility question (aggregated across participants) were highly similar for both texts (Appendix B). Multinomial models estimated separately for

each text did not reveal any significant differences in hypothesis-relevant parameter estimates, which means that these estimates could be replicated across texts. For ease of presentation, we will report multinomial model analyses based on response frequencies aggregated across the two experimental texts.

Definition of bias parameters. The estimated multinomial model consisted of four sub-models corresponding to the four types of experimental items (plausible vs. implausible paraphrases and inferences, Figure 1a and 1b) and one additional model for the distracter items (Figure 2). In each sub-model, there were three types of theoretically relevant parameters, one for the textbase representation (t), two for the situation model representation (s), and four for epistemic understanding (e). Additionally, for each of these parameters a bias parameter was incorporated into the model to correct for guessing and other response strategies. These bias parameters were assumed to be equal in all sub-models and were derived from responses to the distracter items. For the guessing parameter for textbase base ($g(t)$) and the two guessing parameters for the situation model ($g(s \mid t+)$ and $g(s \mid t-)$), the proportions of positive responses to the textbase question ($T+$) and the situation model question ($S+$) in the corresponding branches of the distracter model, respectively, were used as estimates. This procedure is justified because the information contained in the distracter items was neither part of the textbase nor the situation model of the texts.

However, while the assumption of identical situation model guessing parameters for plausible and implausible items seems sensible for the situation model guessing parameter without previous textbase recognition ($g(s \mid t-)$), it is problematic for the situation model guessing parameter with (erroneous) textbase recognition ($g(s \mid t+)$). Theoretically, for plausible items there are strong reasons to assume that textbase recognition implies situation model integration (cf., for example, Kintsch, 1988). If someone recognizes an item as explicitly mentioned in a text,

he or she is very likely to agree that the information contained in this item also matches the content of the text. In contrast, for implausible items it is conceivable that participants realize that an item was explicitly mentioned in the text but that it does not fit into the situation model that they constructed during reading the text (in fact, this assertion is part of one of our predictions). In line with this assumption, the probability of positive responses to the situation model question after giving a positive response to the textbase question (as can be inferred from the raw frequencies given in Appendix B) was very large for plausible items. For implausible items, in contrast, a considerable number of items that received a positive response to the textbase question still received a negative response to the situation model question. Methodologically, the discrepancy in the pattern of results for plausible and implausible items speaks against posing the restriction that the same guessing parameter $g(s \mid t+)$ that is applied to plausible items should also be applied to implausible items. The main argument against posing this restriction is that the guessing parameter $g(s \mid t+)$ has been estimated from plausible distracter items and, as a consequence, might not be adequate for the sub-models of implausible paraphrase and inference items. Given that textbase and situation model judgments seem to be more or less uncoupled in implausible items, it seems more reasonable to assume identical situation model guessing parameters for the implausible items regardless of whether these items have received a positive or a negative response to the textbase question ($g(s \mid t+)=g(s \mid t-)$).

For the guessing parameters for epistemic understanding, a slightly different logic was applied. In order to account for the fact that the distracter items conveyed plausible information, the sub-model for the distracter items contained one common plausibility parameter e_D for all sub-branches of the model. Thus, we assumed that participants were able to (correctly) detect the plausibility of these items with probability e_D. Only if this process failed with probability $(1-e_D)$, participants were assumed to guess plausibility with the probabilities $g(e \mid t+s+)$, $g(e \mid t+s-)$, $g(e \mid$

t-*s*+) or *g*(e | *t*-*s*-), respectively.

Estimation of multinomial model parameters. Due to the additional plausibility parameter e_D in the distracter-model, the sub-branches for the plausibility bias parameters *g* in the distracter model (Figure 2) were not locally identified (Hu & Batchelder, 1994). Therefore, two additional constraints were needed to obtain unique estimates. In a first step, we introduced the equality restrictions *g*(*e* | *s*+)=*g*(*e* | *t*+*s*+)=*g*(*e* | *t*-*s*+) and *g*(*e* | *s*-)=*g*(*e* | *t*+s-)=*g*(*e* | *t*-*s*-) to reduce the number of free parameters and obtain more reliable estimates for the remaining parameters. Given that the responses to the plausibility question were unlikely to vary with previous textbase integration and differences between these parameters were irrelevant for our hypotheses, this restriction seems to be reasonably justified. The second restriction was the assumption that the plausibility bias after situation model integration (*g*(*e* | *s*+)) and the plausibility bias without situation model integration (*g*(*e* | *s*-)) were in opposite directions. In its strongest form, this implies the restriction *g*(*e* | *s*-)=1-*g*(*e* | *s*+), which states that the two bias parameters are complementary to each other. Thus, the restriction equates the probability of a positive response to the plausibility question for distracter items that have been assigned to the situation model with the probability of a negative response for distracter items that have not been assigned to the situation model. This equality constraint is in line with the idea that when responding to the plausibility question, participants use a metacognitive strategy to view those items that they have judged to be part of the situation model representation as more plausible, and to view those items that they have excluded from the situation model as implausible. It is important to note that both restrictions together essentially reformulate the pattern of effects predicted by our second hypothesis in terms of mere guessing. Therefore, introducing bias parameters that are estimated based on these restrictions into the models in Figure 1 yielded a rather strict test of our prediction concerning epistemic understanding. This is because all the remaining effects in the *e*-parameters

can be regarded as unconfounded with the metacognitive strategy captured by the bias parameters.

The resulting model was globally as well as locally identified and had two degrees of freedom. Accordingly, it was possible to assess its fit and test additional hypotheses. All parameters were estimated with the Maximum-Likelihood method as implemented in the Expectation-Maximization-algorithm of the HMMTree program (Stahl & Klauer, 2007).

Power analysis. Overall, there were 12,600 data points (70 participants X 90 items X 2 texts), yielding an extremely high power for our significance tests if conventional levels of type-I-error probability are chosen. As an undesirable consequence of the large data set, even very small differences between the experimental conditions are likely to reach significance even if they are theoretically meaningless (Riefer & Batchelder, 1988). For this reason, the α-level was set to .0001. Despite the low α-level, the power of all hypothesis tests remained high (> .99) even for small effect sizes (w=.10, Cohen, 1988). Power calculations were carried out with the GPower 3 program (Faul, Erdfelder, Lang, & Buchner, 2007).

Goodness of fit of the overall model. The fit of the global model was acceptable ($G^2(2)=6.07$, $p > .01$), given that the power of the test was very high. The estimates and standard errors for the different parameters of the model are displayed in Table 2. In addition, we estimated separate multinomial models for each participant in order to control for biased results due to aggregated data (Riefer & Batchelder, 1991). Because of low cell frequencies for each individual participant, we added a constant of one to all response categories to avoid cells with a frequency of zero. The average fit of the individual models was acceptable ($G^2(2)=5.61$; $p > .05$). Out of the 70 individual models, ten (14%) had to be rejected at an α -level of .01). Moreover, the means of the parameter estimates from the individual models were highly similar to the model based on aggregated data (mean absolute deviation < .05). In the following sections, we will only

report parameter estimates and hypothesis tests based on the aggregated data.

Hypothesis tests. On the basis of the multinomial model in Figure 1, it was possible to test the hypothesized relationships between situation model construction and validation by testing whether parameter estimates in different sub-models or different branches of one sub-model differ from one another. First, the assumption that comprehenders validate incoming information before integrating it into the situation model implies that the estimates for situational model parameters should be larger in the sub-models of plausible test items compared to the sub-models of implausible test items. For the textbase parameters, in contrast, we did not expect any effect of item plausibility but only an effect of item type, with larger textbase parameters for paraphrase items than inference items. Second, the assumption that information which is part of the situation model is more likely to be accepted as being true implies that the parameter estimates for epistemic understanding should be larger for information that has been integrated into the situation model ($e \mid s+$) compared to information that has not been integrated into the situation model ($e \mid s-$). In multinomial models, hypotheses of this kind can be tested by constraining corresponding parameters in different sub-models or different branches of the same sub-model to equal values. If the fit of the model (indicated by the fit statistic G^2) is significantly impaired by these constraints, it can be concluded that the population parameters are indeed different. We report significance tests based on differences (ΔG^2) between the fit of the global model and the fit of the model with the equality constraints corresponding to our hypotheses.

Situation model parameters (s). Our first prediction was that the situation model parameters for plausible information would be larger than those for implausible information. In line with this prediction, the situation model parameters differed significantly between plausible and implausible test items, $\Delta G^2(4)=250.62$, $p < .0001$. Overall, plausible items had a considerable higher probability of being integrated into the situation model (cf. Table 2). This effect was

mainly due to information that was not part of the textbase representation (s | t-), $\Delta G^2(2)= 248.19$, $p < .0001$), whereas the difference between parameters for information that was part of the textbase representation was not significant (s | t+), $\Delta G^2(2)=2.44$, $p = .71$). In sum, participants showed a general tendency to favor plausible over implausible information for inclusion in the situation model of the text content.

In contrast to item plausibility, item type did not have an effect on the situation model parameters (ΔG^2 (4)=19.75, p>.001), indicating that information in paraphrase and inference items had approximately the same probability of being included in the situation model. In addition, for plausible as well as implausible items, situation model parameters were higher for information that had previously been integrated into the textbase representation (ΔG^2 (4)=1162.62, p>.0001).

Moreover, there was an interesting discrepancy between the two guessing parameters that were based on the responses to the situation model question. For distracter items that had erroneously been judged as being part of the textbase, the probability of judging these items as being part of the situation model was very high ($g(s$ | t+)=.92), indicating that overall, participants followed the reasonable heuristic that textbase items should also belong to the situation model. In contrast, the guessing parameter for distracter items that had not been judged as being part of the textbase was below chance level ($g(s$ | t-)=.21).

Textbase parameters (t). For the textbase parameters, we conducted the same comparisons as for the situation model parameters as an additional validity check. The predicted relationships between comprehension and validation should hold on the situation model level but not on the textbase level. For this reason, we did not expect any effects of item plausibility on the textbase parameters. However, we expected an effect of item type. Information in paraphrase items should be more likely to be part of the textbase representation than information in inference items. In

line with these predictions, textbase parameters did not differ significantly between plausible paraphrases and implausible paraphrases ($\Delta G^2(1) = 4.08$, *n.s.*). Contrary to our expectations, however, textbase parameters were slightly higher for plausible inferences than for implausible inferences, $\Delta G^2(1)=39.62$, $p < .0001$. As expected, item type had a strong effect on textbase parameters ($\Delta G^2(2)=639.54$, $p < .0001$). For paraphrases, textbase parameters were higher than for inference items. Additionally, the textbase-guessing parameter $g(t)$ was very low, indicating that participants were unlikely to guess textbase-responses. Thus, participants were able to distinguish between paraphrases and inferences. By and large, the pattern of estimates for the textbase parameters supported the assumption that the content of the textbase representation does not depend on the validity of text information but on the overlap with explicit text information.

Epistemic understanding parameters (e). Our second prediction was that the probability of accepting information as being true would be larger if this information had been integrated into the situation model of the text content. This probability is captured by the epistemic understanding parameters which were defined differently in the models for plausible and implausible test items. In the model for plausible test items, the *e*-parameters represent conditional probabilities that participants judge an item as plausible. Therefore, we used the estimates for these parameters directly to test our hypothesis. In the model for implausible test items, in contrast, the *e*-parameters represent conditional probabilities that participants are able to detect the implausibility of a sentence. Consequently, for implausible items, we used the complementary probabilities (1-*e*) to test our predictions as they reflect the tendency of judging implausible information as being plausible. For ease of presentation, Table 2 provides estimates of the *e*-parameters for the plausible test items and the complementary probabilities (1-*e*) for the implausible test items.

First, consider the plausibility bias parameter $g(e \mid s+)$ (Table 2). This parameter differed

considerably from .50 ($\Delta G^2(1)=858.78$, $p < .0001$), indicating a general bias to judge items as being plausible if they had been integrated into the situation model. However, even if this metacognitive strategy is taken into account, the results were in line with our prediction. Overall, the estimates for the epistemic understanding parameters were considerably higher for information that was part of the situation model compared to information that was not part of the situation model. This tendency was very clear cut for plausible items ($\Delta G^2(4)=20.93$; $p < .0004$). As can be seen in Table 2, the *e* | .*s*+ parameters were generally higher than the *e* | .*s*- parameters, indicating that participants judged items previously incorporated into the situation model as being more plausible than items without situation model integration. This pattern of results did not depend on previous textbase integration. For implausible items, however, the corresponding effect was mainly due to items that had also been integrated into the textbase representation ($\Delta G^2(2)=22.18$; $p < .0001$). The conditional probabilities (1-*e* | *t*+*s*+)were higher than the conditional probabilities (1-*e* | *t*+*s*-), indicating that for information that was also part of their textbase representation, participants were less able to detect the implausibility of this information if it had been integrated into the situation model representation. In contrast, for information that had not been integrated into the textbase representation, the epistemic understanding parameters did not differ with situation model integration ($\Delta G^2(2)=7.06$; $p = .03$). With this one exception, which is probably due to ceiling effects, the results for the plausible as well for the implausible epistemic understanding parameters support the idea that situation model integration enhances the probability that information is not detected as being implausible but is judged as plausible.

Finally, the *e*-parameters for plausible and implausible items, i.e. the (conditional) probabilities of judging the item as plausible, were higher if the information contained in the item had also been integrated into the textbase representation. This unexpected effect was equally strong for plausible ($\Delta G^2(4)=27.44$; $p < .0001$) and implausible items ($\Delta G^2(4)=31.72$; $p < .0001$).

One possible interpretation of this effect is in terms of source credibility (Hovland & Weiss, 1951; Pornpitakpan, 2004). Our texts were typical chapters of undergraduate psychology textbooks which most students would not suspect of containing argumentation errors or faulty information. As a consequence, information that had been recognized as part of the textbase might have been more likely to be judged as being plausible.

In sum, the multinomial model results provide evidence for a strong and bi-directional relationship of situation model construction and the validation of text information. On the one hand, the situation model parameters were influenced by the plausibility of a test item, with plausible information being more likely to be integrated into a situation model representation. This result is not trivial because just as plausible test items, implausible paraphrases and inferences were well understandable and perfectly compatible with the situation described in the experimental texts. The only relevant difference was that implausible test items rested on weak arguments and, as a consequence, were more likely to be rejected by epistemic validation processes. On the other hand, the probability that information was accepted as being true increased considerably if it was part of the situation model.

In principle, off-line measures such as responses to the questions that we used to assess the representational outcomes of comprehending the two experimental texts may be biased by metacognitive and other response strategies (e.g. guessing) at the time of testing. In the present analyses, however, the impact of such strategies was controlled for by bias parameters that were estimated from responses to the distracter items. The predicted pattern of results was left intact when these bias parameters were included in the models. For this reason, an alternative explanation in terms of response strategies at the time of testing is not very likely. Rather, the results suggest that the situation models constructed during reading reflect a close relationship between comprehension and validation.

Response Latencies

We conducted multilevel analyses with random coefficients (Raudenbush & Bryk, 2002) to analyze latencies of responses to the three types of questions, textbase question, situation model question, and plausibility question. For the response latencies, we expected a pattern of effects analogous to the one revealed by the multinomial models analysis of the responses. Accordingly, the first prediction was that responses to the situation model question should be facilitated, i.e. be faster, for plausible items compared to implausible items. No such effect of item plausibility but an advantage of paraphrases over inferences was expected for the latencies of responses to the textbase question as dependent variable. Responses to the plausibility question were expected to be faster for plausible items that were part of an individual's situation model because these items should be easier to verify as being plausible. These responses should also be facilitated in implausible items that were not part of an individual's situation model because these items should be easier to reject as being implausible.

Multilevel models are required in this case because one variable that was used to predict latencies to the plausibility question (situation model integration) was not manipulated experimentally but measured and varied between participants, yielding an incompletely balanced design. As an additional asset, multilevel models of response times allow for including the actual responses as predictor variables. In this way, we were able to account for the general observation that yes-responses are usually faster than no-responses (Luce, 1986). Given that the same pattern of results was expected for responses and response latencies, it was important to control for this potentially confounding factor.

More generally, multilevel models adequately separate item-level variance (level 1) and person-level variance (level 2) in balanced as well as in unbalanced designs. In contrast to classical ANOVA techniques, they can account for the fact that not only participants but also

experimental items are sampled from a larger population (Maxwell & Delaney, 2004; Richter, 2006). Again, we will report results based on latency data combined from both texts. Control analyses that included the experimental text as an additional predictor variable did not reveal any differences relevant for the hypothesis tests.

Data Preparation. The original response latencies were heavily skewed to the right. We log-transformed the latencies to normalize their distributions and to linearize their relationships with the predictor variables (Ratcliff, 1993; see Appendix C for the original response latencies). Log-transformed latencies deviating more than 2.5 standard deviations from the grand mean (2.1 % of all latencies) were discarded from the analysis.

Response latencies for the situation model question. The multilevel model for the log-transformed latencies of the responses to the situation model question contained four item-level (level 1) predictors and one interaction term (see Appendix D for the model equations). First, the *position of an item* in the experimental procedure was included as predictor to control for practice effects (Newell & Rosenbloom, 1981). In addition, we included the actual *response* to each item (yes vs. no) to account for the fact that yes-responses are typically faster than no-responses. As theoretically relevant predictors, *item type* (paraphrase vs. inference items), *item plausibility* (plausible vs. implausible items) and the interaction of both variables were included. The model did not contain any person-level predictors but variance components for the intercept and all item-level slopes (random coefficient regression model, Raudenbush & Bryk, 2002). In this way, it was possible to test whether effects of item level predictors vary randomly between participants. Parameters were estimated with the Restricted Maximum Likelihood algorithm implemented in HLM 6 (Raudenbush, Bryk, & Congdon, 2004).

Table 3 (right columns) summarizes the parameter estimates for the multilevel model for the situation model question. Figure 2b displays the estimated mean latencies for responses to

plausible vs. implausible paraphrase and inference items (back-transformed into the original metric). As predicted, item plausibility had a strong effect on the response latencies. On average, responses to plausible items were 24 ms faster than responses to implausible items. As expected, this effect of item plausibility was not moderated by item type. The variance components for the intercept as well as the slopes for item position, response and item type were significant, indicating random variation of these effects between participants.

Response latencies for the textbase question. The multilevel model for the log-transformed latencies of the responses to the textbase question was identical to the model for the situation model question (Appendix D). The parameter estimates for this model are summarized in Table 3 (left columns). Figure 2a displays the estimated mean latencies for responses to plausible vs. implausible paraphrase and inference items. As expected, there was an effect of item type. Responses to paraphrase items were on average 30 ms faster than responses to inference items. However, this effect was only marginally significant. There was no effect of item plausibility on the responses to the textbase question. None of the variance components associated with these slopes was significantly different from zero.

Response latencies for the plausibility question. For the latencies of responses to the plausibility question, we expected effects of item plausibility that should depend on whether a particular piece of information had been integrated into an individual's situation model. For this reason, we included the predictor *situation model integration* and the interaction of this variable with item plausibility in the model. Apart from these extensions, the model was identical to the models for the textbase and the situation model questions (Appendix E).

Table 4 provides the parameter estimates for the model for the plausibility question. Figure 2c displays the estimated mean latencies for plausible vs. implausible information that was either part of an individual's situation model or not. As expected, there was a large crossover interaction

of item plausibility and situation model integration. Responses to plausible items were slightly faster (16 ms) than responses to implausible items if these items were already part of a participant's situation model. On the contrary, if an item had not been integrated into the situation model it took 63 ms longer to respond to plausible compared to implausible items. No other effects were significant.

In sum, the results for the response latencies were strictly parallel to the results of the multinomial model analyses of the response patterns. Judgments of whether an item was explicitly mentioned in the text showed a tendency to be facilitated for paraphrases compared to inferences but they were independent of the plausibility of the item. For judgments of whether an item was part of the situation model the reverse was true: These judgments were facilitated for plausible compared to implausible items but response latency did not differ between paraphrases and inferences. Again, this may be interpreted as an advantage of plausible information in situation model construction. Plausible information is more likely to be integrated into the situation model, and it is also easier to judge for plausible information whether it matches the situation described by a text or not. In contrast, the parameter estimates in the model for the plausibility question showed that situation model integration may also influence plausibility judgments. These judgments were facilitated for information that was already part of the situation model.

It is important to mention that the theoretically relevant facilitation effects cannot be attributed to the typical response time difference between affirmative and negative responses because this difference was controlled for in our models. Given that the multinomial models revealed effects that appear to be due to properties of the text representation rather than response strategies, it also seems reasonable to assume that the strictly parallel pattern that we found for the response latencies reflects the same representational properties. On the other hand, the

multilevel models of the response latencies are weaker than the multinomial models of the responses as they do not provide ways to dissociate effects that are due to the representation constructed during comprehension from those that are due to response strategies at the time of testing. Related to this is another potential limitation of the conclusions that may be drawn from the latency data. Due to the fact that the three questions were posed in a fixed order, it is likely that the response latencies for these questions were not completely independent from each other. From this perspective, it would probably be more appropriate to view the response latencies as component reaction times that reflect stages of one and the same question-answering process. In sum, the latency data provide additional support for the validity of the multinomial model results and, consequently, for the hypothesized bi-directional relationship of situation model construction and epistemic validation.

Conclusion

This study investigated the idea that situation model construction and epistemic validation reciprocally influence each other in order to create an accurate and stable representation of the state of affairs described in a text. After reading two expository texts that contained plausible and implausible information, participants responded to test items that corresponded either to a textbase or a situation model representation of the text contents. A multinomial model analysis of the responses and a multilevel analysis of the response latencies yielded converging evidence for the hypothesized bi-directional relationships of situation model construction and epistemic validation. In particular, plausible information was more likely to be integrated into the situation model representation than implausible information. In addition, judgments concerning the situation model status of plausible information were facilitated compared to implausible information. On the other hand, information was more likely to be judged as being plausible once it had been integrated into the situation model. In addition, rejection of implausible information

integrated into the situation model was slowed down compared to information not integrated into the situation model.

This very clear-cut pattern of results has methodological as well as theoretical implications. From a methodological perspective, the present study demonstrates that the recognition method proposed by Schmalhofer and Glavanov (1986) can be extended to assess aspects of the propositional and situation model representations in a more direct way. Moreover, the finding that paraphrases and inference items and item plausibility differentially influenced responses to the textbase question and the situation model question strongly suggests that both textbase and situation model should be assessed via different types of judgments rather than one uniform recognition judgment. Clearly, the procedure proposed here would have to be modified to serve the purpose that the original method is intended to serve, i.e. the detailed assessment of individual differences in the strengths of different levels of text representation. In principle, however, these modifications may be implemented by using multinomial models with a slightly different structure and estimating the parameters of these models for each individual participant.

From a more theoretical perspective, the observed pattern of results is consistent with the view that epistemic validation processes are operating during the construction of a situation model. These validation processes are used to monitor the incoming information and evaluate it against the current state of the situation model and background knowledge about the world. As such, they ensure that comprehenders achieve an understanding of communicated information that is a more or less accurate representation of the described state of affairs and at the same time relatively stable and coherent. Both constraints are important to prepare individuals for successful interactions with the world but they also stand in partial opposition to each other. For this reason, a trade-off has to be made which is likely to depend on the goals of the individual and situational demands. Under some circumstances, it may be adaptive to spend time and energy to evaluate the

correctness of information, while in another situation it may be more useful to hang on to the referential representation that has already been achieved or to accept incoming information without scrutinizing it. A good starting point for further research on this topic would be a more detailed investigation of the factors influencing the trade-off between the criterion of truth and the criterion of stability in comprehension. In addition to taking representational outcomes of comprehension processes into account, this research should also include on-line indicators of comprehension and validation processes.

It is important to note that the findings reported here are not covered by common theories of text comprehension such as the Construction-Integration Model (Kintsch, 1988) as these models concentrate on coherence relations and text-driven bottom-up processes (Long & Lea, 2005). With epistemic validation processes, the current study highlights one type of evaluative top-down processes that seems to play a major role in situation model construction. Generally, it would be worthwhile to examine the role of these processes and their interplay with comprehension processes across a wider range of comprehension situations. In many of these situations, for example, epistemic validation processes may be expected to trigger remedial activities in case implausible information is detected (Clark, 1996). These activities may include asking your conversational partner to elaborate further on a piece of troubling information, or to consult another text to cross-check the information.

References

Albrecht, J.E., & Myers, J.L. (1995). Role of context in accessing distant information during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1459-1468.

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen?* [How understandable are our newspapers?]. Unpublished doctoral dissertation, University of Zürich, Switzerland.

Anderson, C.A., Lepper, M.R., & Ross, L.R. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*, 1037-1049.

Bransford, J.D., & Johnson, M.K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Clark, H.H. (1996). *Using language.* Cambridge, UK: Cambridge University Press.

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

Dauer, F.W. (1989). *Critical thinking: An introduction to reasoning*. New York: Oxford University Press.

Deutscher Wortschatz Leipzig [German Online Lexicon Database] (2008). Retrieved April 15, 2008, from http://wortschatz.uni-leipzig.de

Dreisbach, G., & Goschke, T. (2004). How positive affect modulates cognitive control: Reduced perseveration at the cost of increased distractibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 343–353.

Faul, F., Erdfelder, E., Lang, A., & 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.

Fuchs, R., & Schwarzer, R. (1997). Tabakkonsum: Erklärungsmodelle und Interventionsansätze [Tobacco consumption: Explanations and interventions]. In R. Schwarzer (Ed.), *Gesundheitspsychologie: Ein Lehrbuch* (pp. 209-244). Göttingen, Germany: Hogrefe.

Gerrig, R.J., & McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes, 26*, 67-86.

Glenberg, A.M. (1997). What memory is for. *Behavioral &Brain Sciences, 20,* 1-19.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Hovland, C., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15,* 635-650.

Hu, X., & Batchelder, W. H. (1994b). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59,* 21-47.

Johnson, H.M., & Seifert, C.M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1420-1436.

Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inferences, and consciousness*. Cambridge: Cambridge University Press.

Johnson-Laird, P.N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111*, 640-661.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163-182.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Klauer, K.C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning.

Psychological Review, 107, 852-884.

Lea, R.B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 18-26.

Lea, R.B., Mulligan, E.J., & Walton, J.L. (2005). Accessing distant premise information: How memory feeds reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 387-395.

Long, D., & Lea, B. (2005). Have we been searching for meaning in all the wrong places? Defining the "search after meaning principle" in comprehension. *Discourse Processes, 39*, 279-298.

Luce, R.D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Maxwell, S.E., & Delaney, H.D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

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.

Myers, J.L., & O'Brien, E.J. (1998). Accessing the discourse representation during reading. *Discourse Processes, 26*, 131-157.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

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.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades of research. *Journal of Applied Social Psychology, 34,* 243-281.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin,*

114, 510-532.

Raudenbush, S.W., Bryk, A.S., & Congdon, R. (2004). *Hierarchical Linear & Nonlinear Modeling (Version 6.0).* Lincolnwood, IL: Scientific Software International.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41,* 221-250.

Richter, T., Schroeder, S., & Wöhrmann, B. (2006). You don't have to believe everything you read: Comprehension and validation are closely linked modes of information processing. *Manuscript submitted for publication*.

Riefer, D.M., & Batchelder, W.H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318-339.

Riefer, D.M., & Batchelder, W.H. (1991). Statistical inference for multinomial processing tree models. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 313-335). New York: Springer.

Ross, L., Lepper, M.R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of personality and Social Psychology, 32*, 880-892.

Ross, L., Lepper, M.R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology, 32*, 880-829.

Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language, 25,* 279-294.

Singer, M. (1993). Causal bridging inferences: Validating consistent and inconsistent sequences. *Canadian Journal of Experimental Psychology, 47*, 340-359.

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

Singer, M., Halldorson, M., Lear, J. C., & Andrusiak, P. (1992). Validation of causal bridging inferences. *Journal of Memory and Language, 31*, 507-524.

Stahl, C., & Klauer, K.C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods*, *39*, 267-273.

Thomas, A. (1991). *Grundriß Sozialpsychologie, Bd. 1* [Outline of social psychology, Vol. 1]. Göttingen, Germany: Hogrefe.

Zwaan, R.A. (1999). Embodied cognition, perceptual symbols, and situation models. *Discourse Processes, 28*, 81-88.

Zwaan, R.A., & Radvansky, G.A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

Appendix A: Sample Excerpt from the Experimental Text *Characteristics and Causes of Smoking Behavior*

The sample excerpt (296 words) is a translation of two passages taken from one of the experimental texts (adapted from Fuchs & Schwarzer, 1997; translation from German). Implausible sentences are printed in italics.

A person's development towards becoming a smoker (and later on maybe a non-smoker or ex-smoker again) is influenced by a complex interaction of a multitude of social, psychological, and biological factors. With regard to the biological factors, empirical support (especially from twin studies) underlines the decisive role of genetic influences on the initiation and continuation of smoking behavior. In this context, the concept of a hereditary sensitivity to nicotine plays an important role. *At the core of this concept is the observation that some people are more sensitive to nicotine, because they are more susceptible to nicotine than others.* The concept of nicotine sensitivity is used to explain why some people do not become addicted even though they have already smoked a considerable number of cigarettes.

However, the question of a genetic predisposition for smoking is not limited to the explanation of a differential sensitivity to nicotine. An additional influence is exerted by hereditary personality traits which are associated with smoking. On the basis of Eysenck's three-factor-theory of personality, a positive correlation between smoking behavior and the trait "extraversion" has been postulated and supported by a number of empirical studies (e.g., Lipkus, Barefoot, Williams & Siegler, 1994). The same is true for the correlation between smoking behavior and the trait "sensation seeking" sensu Zuckerman. *There is little substantial doubt or controversy about the assumption that there are intraindividual differences in the biological predisposition which covary with the probability of smoking, its initiation, its continuation, and*

possibly its termination. Despite this general consensus, it is largely unclear how these differential biological factors influence the cognitive and emotional processes important in explaining how a person starts smoking, why smoking becomes a continuous habit, and why under certain conditions or in a certain situation, a person quits smoking.

[…]

Appendix B: Frequencies of Responses to Textbase Question, Situation Model Question, and Plausibility Question

	Text on Interpersonal Attraction								*Text on Smoking Behavior*							
	Textbase Question								Textbase Question							
	T+				T–				*T*+				T–			
	Situation Model Question				Situation Model Question				Situation Model Question				Situation Model Question			
	S+		S–		*S*+		S–		*S*+		S–		*S*+		S–	
	Plausibility Question		Plausibility Question		Plausibility Question		Plausibility Question		Plausibility Question		Plausibility Question		Plausibility Question		Plausibility Question	
Item type	P+	P–	P+	P–	P+	P–	P+	P–	P+	P–	P+	P–	P+	P–	P+	P–
Paraphrase																
Plausible[a]	633	21	9	3	218	22	70	74	648	45	4	2	197	26	41	87
Implausible[a]	396	156	1	47	123	42	52	233	413	216	2	71	121	44	27	156
Inference																
Plausible[a]	332	11	6	0	391	26	110	174	471	35	1	5	294	29	76	139
Implausible[a]	213	92	3	44	192	63	102	341	178	95	5	35	150	86	85	416
Distracter[b]	190	10	11	1	472	27	875	514	59	9	3	4	305	26	1087	607
Total	1764	290	30	95	1396	180	1209	1336	1769	400	15	117	1067	211	1316	1405

Note. *T*+/T–: yes/no-response to the textbase question. *S*+/S–: yes/no-response to the situation model question. P+/P–: yes/no-response to the plausibility question. [a] 1050 responses per text (70 participants X 15 items), [b] 2100 responses per text (70 participants X 30 items).

Appendix C: Latencies of Responses to Textbase Question, Situation Model Question, and Plausibility Question for Experimental Items

	Textbase Question[a]		Situation Model Question[b]		Plausibility Question[c]	
Item type	*M*	SE_M	*M*	SE_M	*M*	SE_M
Paraphrase						
Plausible	1650	41	741	26	774	26
Implausible	1626	41	836	25	814	27
Inference						
Plausible	1566	42	694	25	791	26
Implausible	1560	41	781	25	822	27
Total	1601	18	763	8	800	9

Note. Response latencies that deviated more than 2.5 standard deviations from the grand mean (after log-transformation) were discarded from the analysis. [a] Based on 8240 responses, [b] based on 8210 responses, [c] based on 8217 responses.

.

Appendix D: Multilevel Model for the Response Latencies of the Textbase Question and the Situation Model Question

$Y_{ij} = \beta_{0ij} + \beta_{1ij} (X_{1ij} - \overline{X}_{1.j}) + \beta_{2ij} X_{2ij} + \beta_{3ij} X_{3ij} + \beta_{4ij} X_{4ij} + \beta_{5ij} X_{3ij} X_{4ij} + r_{ij}$. (Level 1, experimental items)

$\beta_{0ij} = \gamma_{00} + u_{0j}$. (Level 2, participants, intercept model)

$\beta_{1ij} = \gamma_{10} + u_{1j}$. (Level 2, participants, model for the slope of X_1)

$\beta_{2ij} = \gamma_{20} + u_{2j}$. (Level 2, participants, model for the slope of X_2)

$\beta_{3ij} = \gamma_{30} + u_{3j}$. (Level 2, participants, model for the slope of X_3)

$\beta_{4ij} = \gamma_{40} + u_{4j}$. (Level 2, participants, model for the slope of X_4)

$\beta_{5ij} = \gamma_{50} + u_{5j}$. (Level 2, participants, model for the slope of the interaction of X_3 and X_4)

Criterion and predictor variables:

Y_{ij}: Response latency (log-transformed) for item i and participant j

X_{1ij}: Position of item i in the experimental procedure of participant j

X_{2ij}: Response to question for item i of participant j (dummy coded: 1 = yes-response, 0 = no-response)

X_{3ij}: Item type of item i received by participant j (contrast coded: -0.5 = inference, 0.5 = paraphrase)

X_{4ij}: Item plausibility of item i received by participant j (contrast coded: -0.5 = implausible, 0.5 = plausible)

$X_{3ij} X_{4ij}$: Interaction of item type X item plausibility of item i received by participant j

Appendix E: Multilevel Model for the Response Latencies of the Plausibility Question

$$Y_{ij} = \beta_{0ij} + \beta_{1ij} (X_{1ij} - \overline{X}_{1.j}) + \beta_{2ij} X_{2ij} + \beta_{3ij} X_{3ij} + \beta_{4ij} X_{4ij} + \beta_{5ij} X_{5ij} + \beta_{6ij} X_{3ij} X_{4ij} + \beta_{7ij} X_{4ij} X_{5ij} + r_{ij} .$$ (Level 1, experimental items)

$\beta_{0ij} = \gamma_{00} + u_{0j}$. (Level 2, participants, intercept model)

$\beta_{1ij} = \gamma_{10} + u_{1j}$. (Level 2, model for the slope of X_1)

$\beta_{2ij} = \gamma_{20} + u_{2j}$. (Level 2, model for the slope of X_2)

$\beta_{3ij} = \gamma_{30} + u_{3j}$. (Level 2, model for the slope of X_3)

$\beta_{4ij} = \gamma_{40} + u_{4j}$. (Level 2, model for the slope of X_4)

$\beta_{5ij} = \gamma_{50} + u_{5j}$. (Level 2, model for the slope of X_5)

$\beta_{6ij} = \gamma_{60} + u_{6j}$. (Level 2, model for the slope of the interaction of X_3 and X_4)

$\beta_{7ij} = \gamma_{70} + u_{7j}$. (Level 2, model for the slope of the interaction of X_4 and X_5)

Criterion and predictor variables:

Y_{ij}: Response latency (log-transformed) for item i and participant j

X_{1ij}: Position of item i in the experimental procedure of participant j

X_{2ij}: Response to plausibility question for item i and participant j (dummy coded: 1 = yes-response, 0 = no-response)

X_{3ij}: Item type of item i received by participant j (contrast coded: -0.5 = inference, 0.5 = paraphrase)

X_{4ij}: Plausibility of item i received by participant j (contrast coded: -0.5 = implausible, 0.5 = plausible)

X_{5ij}: Situation model integration of item i received by participant j (dummy coded: 1 = yes-response to situation model question, 0 = no-response to situation model

question)

X_{3ij} X_{4ij}: Interaction of item type X item plausibility of item i received by participant j

X_{4ij} X_{5ij}: Interaction of item plausibility X situation model integration of item i received by participant j

Table 1

Examples for Plausible Sentences, Different Types of Implausible Sentences, and Corresponding Test Items

Sentence type	Original Sentence	Paraphrase Test Item	Inference Test Item
Plausible	The concept of nicotine sensitivity is used to explain why some people do not become addicted even though they have already smoked a considerable number of cigarettes.	The construct of nicotine sensitivity is employed in the explanation of why some people do not become addicted to nicotine.	If a person reacts more strongly to nicotine then this person has a higher risk of developing a nicotine addiction.
Implausible			
Contradiction	The fact that the proportion of teenagers consuming nicotine decreases from 80% to 50 % over the course of adolescence leads to the conclusion of an increasing interest in smoking throughout this phase.	The proportion of teenagers smoking occasionally or regularly decreases towards the end of adolescence from 80 % to 50 % which implies that there is an increasing interest in smoking over the course of youth.	The gradually decreasing proportion of smokers throughout the phase of adolescence is an indicator of an increasing interest in smoking.
Wrong example	Habitual smokers exhibit a number of impairments under conditions of nicotine deprivation: Reduced	There is a number of impairments observable in habitual smokers deprived of nicotine ranging from enhanced memory	The impairments associated with a deprivation from nicotine become apparent in the form of an improved

	restlessness, a good mood as well as an increased concentration and a higher memory capacity.	performance and concentration to a good mood and less agitation.	cognitive performance and a higher affective stability.
False dichotomy	Schachter responded to critics that stress caused a decrease in the urine ph factor and that therefore stress was no more than an unmediated effect whereas nicotine regulation directly causes a higher smoking frequency.	In reply to his critics, Schachter stated that the urine ph factor decreased under conditions of stress or anxiety, and that therefore, stress was only a direct but nicotine regulation an immediate cause of increased smoking activity.	When considering a number of different factors influencing smoking behavior, one should clearly distinguish between direct and immediate influences.
Circular reasoning	This result implies a strong impact of a cigarette's nicotine content on the smoking behavior as this aspect exerts strong influence on smoking.	Because the amount of nicotine per cigarette has a strong influence on smoking, it has a significant effect on the smoking behavior.	The effects of the amount of nicotine per cigarettes can be explained in terms of their influence on the smoking behavior.
Conversion of cause and consequence	If the children develop a positive attitude towards smoking the father displays that his cigarette after dinner tastes wonderful.	If children's positive attitude towards smoking is reinforced, then the father shows how much he enjoys his cigarette after dinner.	Reinforcing children's positive attitude towards smoking causes parents to smoke more often after meals.

Table 2

Parameter Estimates (with Standard Errors) for Experimental Test Items and Bias Parameters

Item type		Textbase	Situation Model		Epistemic Understanding			
		t	$s \mid t+$	$s \mid t-$	$e \mid t+s+$	$e \mid t+s-$	$e \mid t-s+$	$e \mid t-s-$
Paraphrase	Plausible	.62 (.01)	.80 (.06)	.53 (.02)	.73 (.04)	.66 (.13)	.43 (.08)	.28 (.04)
	Implausible	.59 (.01)	.88 (.01)	.26 (.02)	.84 (.02)[a]	.14 (.08)[a]	.90 (.03)[a]	.93 (.06)[a]
Inference	Plausible	.37 (.01)	.79 (.08)	.49 (.02)	.70 (.05)	.49 (.17)	.59 (.06)	.23 (.03)
	Implausible	.27 (.01)	.83 (.02)	.17 (.02)	.83 (.03) [a]	.51 (.17) [a]	.85 (.03) [a]	1.00 (.11)[a]

Bias parameters	$g(t)$	$g(s \mid t+)$	$g(s \mid t-)$	$g(e \mid s+)$	e_D
	.07 (.01)	.93 (.01)	.21 (.01)	.82 (.01)	.56 (.01)

Note. Bias parameters $g(t)$ [textbase question], $g(s \mid t+)$, $g(s \mid t+)$, $g(s \mid t-)$ [situation model question], and $g(e \mid s+)$ [plausibility question] were estimated from the responses to the distracter items (Figure 2). The parameter e_D was also estimated from the distracter items. It reflects the probability of correctly judging the distracter items as plausible.

[a] The parameter estimates for implausible items are provided as complementary probabilities (1-e) in order to facilitate comparisons with plausible items.

Table 3

Estimates for Fixed Effects and Variance Components in the Multilevel Models for the Log-transformed Latencies of Responses to the Textbase Question and the Situation Model Question

Parameter	Textbase Question					Situation Model Question				
	Fixed Effects			Variance Components		Fixed Effects			Variance Components	
	Estimate	*SE*	*t* (69)	Estimate	χ^2 (69)	Estimate	*SE*	*t* (69)	Estimate	χ^2 (69)
γ_o (intercept)	6.961	0.065	107.81***	.292	4674.91***	5.958	0.065	91.02***	.297	2742.93***
γ_1 (position)	-0.005	0.001	-9.67***	< .001	176.17***	-0.011	0.001	-20.83***	< .001	118.73***
γ_2 (response)[a]	-0.113	0.025	-4.58***	.024	145.44***	-0.142	0.034	-4.13***	.041	139.37***
γ_3 (item type)[b]	-0.027	0.016	-1.65[†]	.003	73.40	-0.010	0.024	-0.41	.012	93.99*
γ_4 (plausibility)[c]	0.014	0.017	0.84	.005	90.87	-0.062	0.022	-2.83**	.009	77.27
γ_5 (item type X plausibility)	-0.021	0.028	-0.75	.003	59.41	-0.052	0.039	-1.34	.008	63.53

Note. [a] dummy coded: 0 = no, 1 = yes, [b] contrast coded: -0.5 = inference, 0.5 = paraphrase, [c] contrast coded: -0.5 = implausible, 0.5 = plausible.

[†] $p = .10$* $p = .05$, ** $p = .01$, *** $p = .001$ (two-tailed).

Table 4

Estimates for Fixed Effects and Variance Components in the Multilevel Models for the Log-transformed Latencies of Responses to the Plausibility Question

Parameter	Fixed Effects			Variance Components	
	Estimate	*SE*	*t* (69)	Estimate	χ^2 (69)
γ_o (intercept)	5.979	0.066	90.33***	.303	2575.15***
γ_1 (position)	-0.009	0.001	-16.78***	< .001	112.11***
γ_2 (response)[a]	-0.188	0.054	-3.49***	.138	204.61***
γ_2 (item type)[b]	-0.023	0.018	-1.29	.002	41.59
γ_3 (plausibility)[c]	0.160	0.046	3.87***	.031	77.62
γ_4 (situation model integration)[d]	0.107	0.054	1.98*	.130	180.51***
γ_5 (item type X plausibility)	-0.020	0.034	-0.58	.010	43.52
γ_6 (situation model integration X plausibility)	-0.200	0.054	-3.74***	.034	79.19

Note. [a] dummy coded: 0= no, 1 = yes, [b] contrast coded: -0.5 = inference, 0.5 = paraphrase, [c] contrast coded: -0.5 = implausible, 0.5 = plausible, [d] dummy coded: 0 = no-response to the situation model question, 1 = yes-response to the situation model question.

* p = .05, ** p = .01, *** p = .001 (two-tailed).

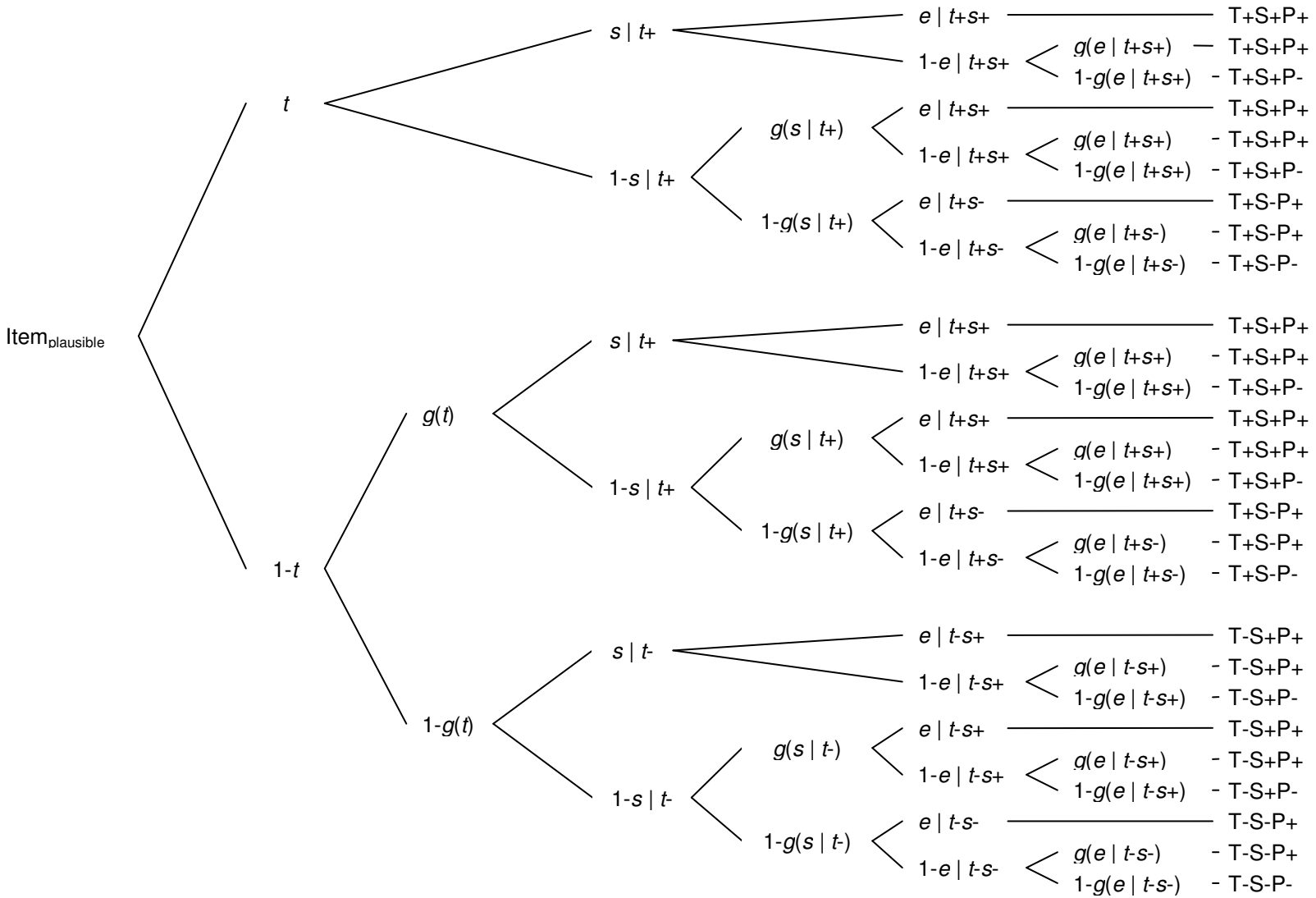
Figure Captions

Figure 1. Multinomial model of responses to plausible paraphrase and inference test items (a) and implausible paraphrase and inference test items (b). Yes- and no-responses to the textbase question (T+/T-), the situation model question (S+/S-), and the plausibility question (P+/P-) are modeled as function of a textbase parameter (*t*), situation model parameters (*s*), and parameters capturing epistemic understanding (*e*). In addition, the model contains bias parameters (*g*) estimated from the distracter items (cf. Figure 2) to correct for guessing.

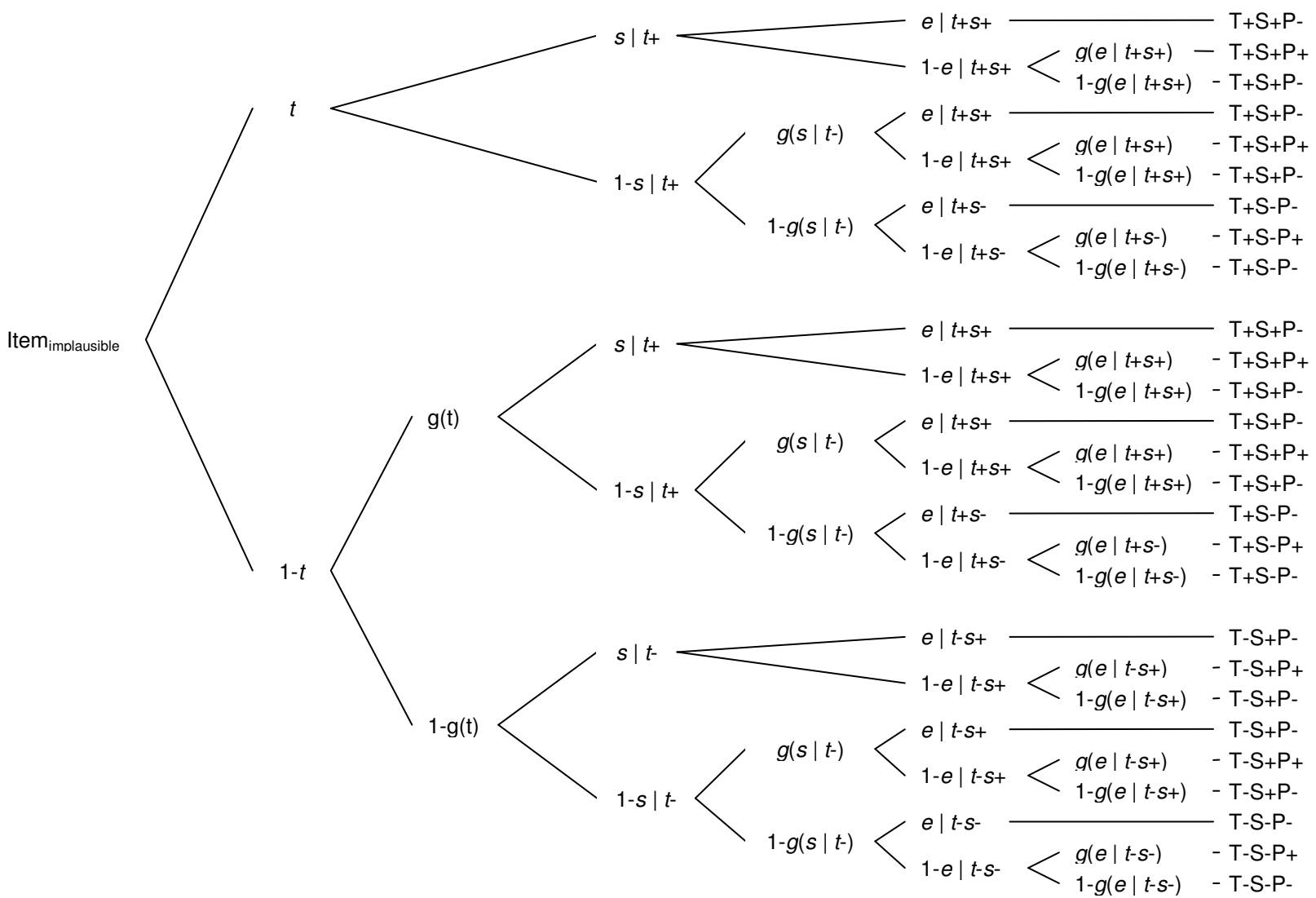
Figure 2. Multinomial model of responses to distracter items with estimates of bias parameters for guessing responses to the textbase question, *g*(*t*), the situation model question, g(s), and the plausibility question, *g*(*e*).

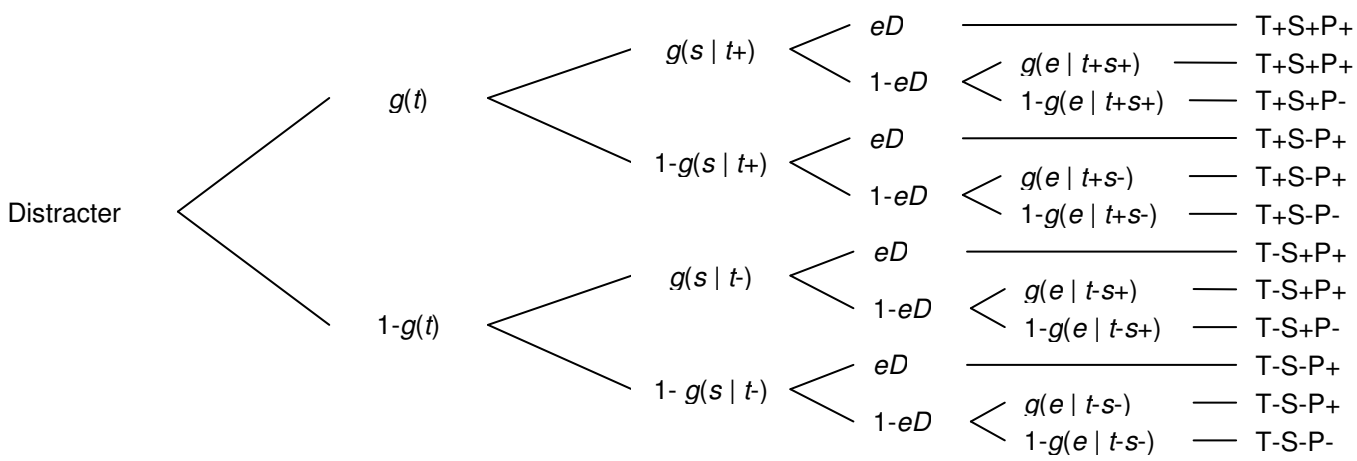
Figure 3. Response latencies for the textbase question (a), the situation model question (b) and the plausibility question (c) adjusted for all other variables in the multilevel models (see Appendices D and E). Error bars represent the standard error of the mean.

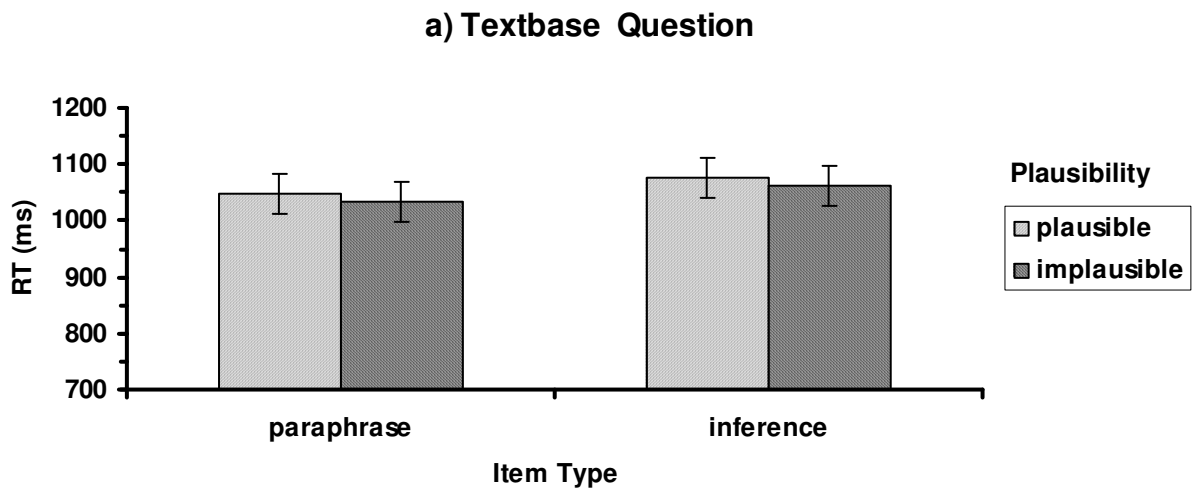
a) Plausible Test Items

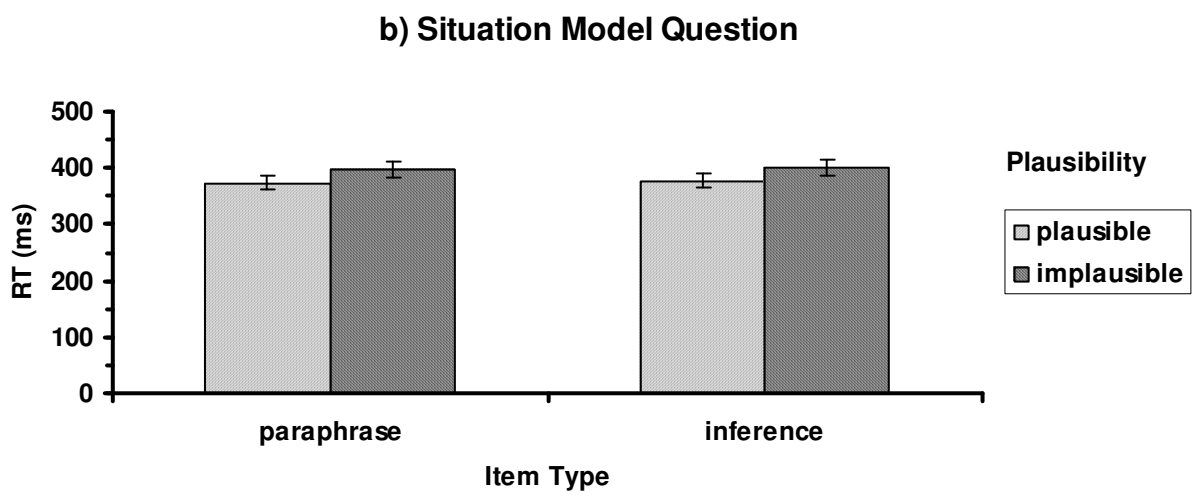


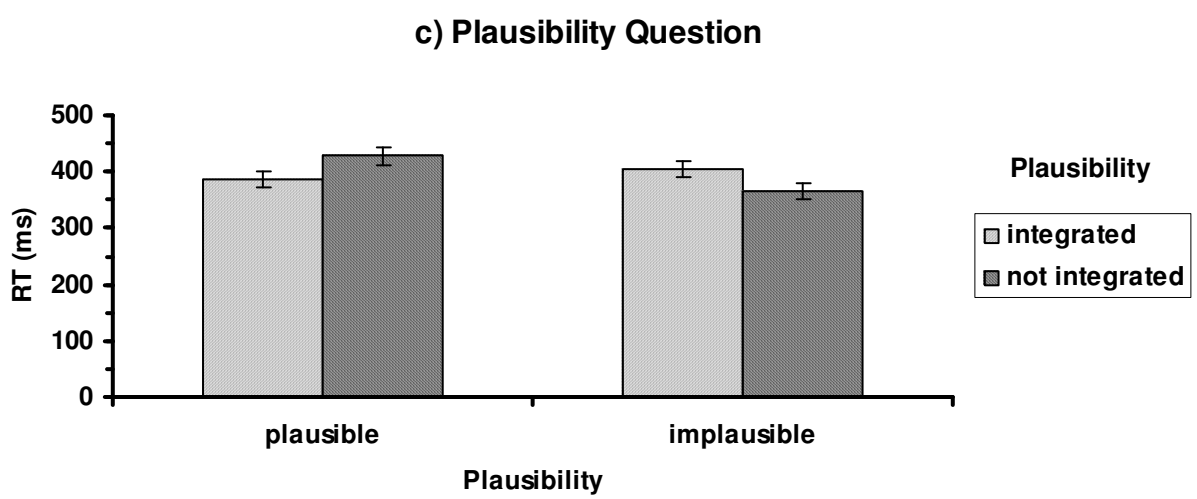
b) Implausible Test Items




Distracter
g(t)
1-g(t)
g(s | t+)
1-g(s | t+)
g(s | t-)
1- g(s | t-)
eD
1-eD
g(e | t+s+)
1-g(e | t+s+)
g(e | t+s-)
1-g(e | t+s-)
g(e | t-s+)
1-g(e | t-s+)
g(e | t-s-)
1-g(e | t-s-)
T+S+P+
T+S+P+
T+S+P-
T+S-P+
T+S-P+
T+S-P-
T-S+P+
T-S+P+
T-S+P-
T-S-P+
T-S-P+
T-S-P-


a) Textbase Question
1200
1100
1000
900
800
700
RT (ms)
paraphrase
inference
Item Type
Plausibility
plausible
implausible


b) Situation Model Question
500
400
300
200
100
0
RT (ms)
paraphrase
inference
Item Type
Plausibility
plausible
implausible


c) Plausibility Question
500
400
300
200
100
0
RT (ms)
plausible
implausible
Plausibility
Plausibility
integrated
not integrated