Darley, E. J., Kent, C., & Kazanina, N. (2020). A 'no' with a trace of 'yes': a mouse-tracking study of negative sentence processing. *Cognition*, *198*, [104084].

# A 'no' with a trace of 'yes': a mouse-tracking study of negative sentence processing

Emily J. Darley

Christopher Kent

Nina Kazanina

School of Psychological Science, University of Bristol

Word Count: 15,073 (excluding references & supplementary materials)

Corresponding Author:

Nina Kazanina

School of Psychological Science

University of Bristol

12a Priory Road

Bristol, BS8 1TU

United Kingdom

Email: nina.kazanina@bristol.ac.uk

Declarations of interest: none.

Funding: This work was supported by an ESRC Studentship to EJD. The funding source had no involvement in any aspect of the study.

## Abstract

There is strong evidence that comprehenders can parse sentences in an incremental fashion. However, when the sentence contains a negation, the evidence is less clear. Previous work has shown that increasing the pragmatic felicity of a negative sentence reduces or eliminates any processing overhead relative to affirmative sentences. However, in previous work felicity has gone hand-in-hand with the predictability of critical material in a sentence. In three experiments reported here, we presented equally felicitous sentences with critical material of varying predictability (operationalised as the number of possible completions) to test whether this might be a critical factor determining the ease with which partial sentences containing a negation are interpreted. Participants completed a truth-value judgement task (Experiment 1) or a sentence completion task (Experiments 2 and 3) after viewing a visual environment that provided the context for a test sentence, which could differ in truth value (in Experiment 1 only), polarity (affirmative or negative), and number of possible completions (one, two, or three). In all three experiments, we recorded response times and accuracy, but also response dynamics via participants' computer mouse trajectories, allowing us to test specific hypotheses about the time course of comprehension. Across all experiments, in conditions with one or two possible targets, we observed consistent detrimental effects of negative polarity, suggesting that the difficulty in processing negation cannot be reduced to effects relating to predictability or pragmatic felicity. We discuss this finding in relation to incremental and two-stage models of processing and outline a new account of the processing difficulty arising from negation in terms of a conflict between what is locally activated on the basis of individual words and phrases and the global meaning of a negative sentence.

**Keywords:** negation, mouse-tracking, sentence processing, prediction

**A ‘no’ with a trace of ‘yes’: a mouse-tracking study of negative sentence processing**

A linguistic signal of any nature, be it oral as in spoken language, visual as in sign language, or written as in text, unfolds over time: rather than being available to the comprehender in one go, the signal becomes available in a piecemeal fashion. At a minimum, this property requires that the parser aggregate linguistic input before its propositional content can be derived. In fact, the parser goes far beyond simple aggregation: it analyses and incorporates newly encoded input (roughly, every new word) with the previous sentence fragment as soon as it arrives, a process known as incremental parsing (Crocker, 1994; Kimball, 1975). Representation of the partial input at each timepoint as a connected syntactic structure (Sturt & Lombardo, 2005) significantly reduces the otherwise insurmountable cost of retaining unintegrated sequences of words in working memory. To achieve incrementality, the parser must actively build and constantly update its representations of partial utterances, and may engage in the formation of predictions that will help it to integrate upcoming material (Sturt & Lombardo, 2005). Whereas the exact nature of the predictive mechanisms involved is debated (Kazanina, 2017; Nieuwland et al., 2018), the idea that language processing generally proceeds in an active, incremental fashion, rather than passively waiting for a full proposition whose meaning can be represented in its entirety, is unequivocally accepted (Altmann & Mirković, 2009; Pickering & Garrod, 2007; but cf. Pritchett, 1991). However, this does not necessarily mean that complete interpretation of the structure occurs at every point during input; immediate, incremental processing of material may only mean that it is incorporated in some way into the partial structure built to incorporate material encountered so far (Garrod & Sanford, 1999).

Alongside this background of broad acceptance of the notion of incrementality in sentence processing, there are a small number of cases in which this does not conclusively

hold. In particular, there is evidence that negative utterances may constitute an exception to incrementality, or a case in which immediate incorporation does not include complete interpretation, presenting one scenario in which it is difficult for the parser to generate accurate predictions that take into account all the information available. In the philosophical and semantic literature (reviewed, e.g., by Horn, 1989), various positions on negation hold that there is an intrinsic asymmetry between affirmatives and negatives, backed by observations from the morphosyntactic markedness of negation, with negative markers ('not') being present in all languages but affirmative markers being rather rare (St Thomas Aquinas, cited in Oesterle, 1962:64) to the semantico-pragmatic observation that negative statements (*The sky is not green*) are generally less informative than affirmative ones (*The sky is blue*). Views on this asymmetry and their history are summarised by Horn and Wansing (2017).

This stance extends to several accounts of the online processing of negatives. In a pioneering study, Fischler, Bloom, Childers, Roucos, and Perry (1983) found that false sentences like *A robin is not a bird* elicit a reduced N400 event-related potential (ERP) component, generally taken to indicate conformance with semantic expectations or ease of integration (Kutas & Federmeier, 2011), in comparison to true sentences like *A robin is not a tree*. The authors interpreted their findings as supporting a sentence comprehension model whereby negation is processed in two steps: the embedded affirmative proposition, such as {is a | robin, bird} — corresponding to *A robin is a bird*, is processed first, while the negation is applied in a strictly subsequent step, {false | {is a | robin, bird}} (Clark & Clark, 1977; Kintsch, 1974). Similar results have been obtained in other ERP studies on negative sentences, supporting the view that their truth value is not always predictive of the magnitude of the N400 component (Kounios & Holcomb, 1992), especially if only a short processing window is available (Lüdtke, Friedrich, de Filippis, & Kaup, 2008). These findings add to

earlier psycholinguistic findings demonstrating more generally (without necessarily supporting a specific theory on why) that negative sentences are harder and slower to process (Just & Carpenter, 1971; Gough, 1965; Wason, 1959, 1961; Wason & Jones, 1963). Subsequent accounts of the mechanisms underlying this difficulty include the need to complete a processing step in which a negative 'tag' is incorporated into the representation (Carpenter & Just, 1975) and the need to engage in simulation of a model world representing the embedded proposition before identifying possible worlds corresponding to the full negative proposition (Kaup, Lüdtke, & Zwaan, 2006; Kaup & Zwaan, 2003).

More recent evidence, however, has suggested that the incremental processing difficulties associated with negation are in fact limited to specific circumstances, largely those in which missing contextual information makes the use of negation irrational and inefficient. Theoretical positions of this nature have argued that negation must normally be licensed by the pragmatics of the context and is used to reject what plausibly may have been true. For example, assertions may be felicitously formulated using negation (*A whale is not a fish*) in the context of previous assertions or existing misconceptions — in this case, that whales are fish (Wason, 1965), or when there is a prominent question under discussion (*It was Mike who didn't iron his shirt*; Tian, Breheny, & Ferguson, 2010). Nieuwland and Kuperberg (2008) provide experimental support for this claim by showing that the N400 pattern observed when negative sentences are used out of context as in Fischler et al. (1983) (e.g., a larger N400 at the adjective in *Bulletproof vests aren't very dangerous* than in *Bulletproof vests are very safe*) disappears when pragmatically licensed negatives and affirmatives are compared (e.g., *With proper equipment, scuba diving is / isn't very safe / dangerous and often good fun*).

A similar distinction in processing between context-lacking and context-embedded negative sentences has been observed by Dale and Duran (2011) using a computer mouse-

tracking methodology that is also employed in the current study. Their participants' task was to report their judgements on the truth value of sentences using a computer mouse; when participants judged sentences presented out of context as true or false, they followed more direct trajectories in giving their responses for affirmative statements (*Elephants are large*) than for negative statements (*Elephants are not small*). This was consistent with the two-stage processing view: before choosing the correct response to negative sentences, participants were momentarily attracted towards the incorrect one. Critically, however, the difference in trajectories between negatives and affirmatives disappeared when a rich discourse preamble was presented (e.g., *'You want to lift an elephant?' the mother said to her child, 'but elephants are not small.'* in their Experiment 3). These findings suggest that negative sentences are processed in a fully incremental manner as long as they are used in an appropriate pragmatic context, supporting the view that negation is inherently contextual (e.g., Glenberg, Robertson, Jansen, & Johnson-Glenberg, 1999) and that negatives are not necessarily any more difficult to process than affirmatives (e.g., Johnson-Laird & Tridgell, 1972; Wason 1965). Developmental research has identified similar patterns in young children: for example, negative sentences are more readily understood and accepted by 3- and 4-year-olds when presented in a pragmatically supportive context (Nordmeyer & Frank, 2018), although 2-year-olds appear to have more fundamental difficulties with semantic processing of truth-functional negation (Reuter, Feiman, & Snedeker, 2018).

In sum, it is plausible that the apparently non-incremental processing of negative sentences stems from the fact that studies observing this employed negative sentences that were not pragmatically licensed by context. Pragmatically licensed negative sentences are processed as incrementally as affirmative sentences. However, the same findings also led us to another observation that highlights the role of predictability of communicative intent and (thereby) of upcoming material in processing of negative sentences. In particular, enriching

the pragmatics of the situational context (as in Dale & Duran, 2011; Nieuwland & Kuperberg, 2008) narrows the scope of possible communicative meanings and enables more accurate predictions about upcoming material in a negative sentence. Without context, the sentence *Elephants are not …* could be continued in a multitude of ways: besides *small*, these include *yellow*, *extinct*, *from America*, etc. Addition of the preceding context *'You want to lift an elephant?'* singles out a relevant dimension, i.e. weight or size, and makes the continuation *small* (or similar) predictable. Similarly, Mayo, Schul, and Burnstein (2004) showed that negatives involving bipolar predicates (those with a clear opposite, such as *tidy – messy*) are easier to process and recall (compared to those without a clear opposite, such as *creative*, which might have antonyms on various axes, including *dull*, *unproductive*, or *untalented*), perhaps because the relevant dimension along which the negation applies is clear. A similar result was obtained by Orenes, Beltrán, and Santamaría (2014) in a comparison of negation processing in binary vs. 'multary' contexts. In other words, negation is easier to process in the presence of a clear 'contrast frame' (Löbner, 2000), and one way of supplying this is through contextual enrichment. The impact of predictability, operationalised as cloze probability, on the processing of positive and negative quantifiers (e.g. *most/few*) has also yielded comparable results in a recent study by Nieuwland (2016): N400 responses to low-cloze sentences indicated a similar interaction between truth value and quantifier polarity to that observed in previous studies for low-felicity affirmative and negative sentences, whereas positive and negative quantifiers produced much more similar patterns when they were presented in high-cloze sentences. This pattern suggests that quantifiers can be incrementally incorporated and used to make specific predictions for upcoming material, where such predictions are available, regardless of whether they are positive or negative.

Following these observations on the potentially crucial role of predictability, the current study investigates whether predictability governs the extent to which comprehenders

can update their representations for negative sentences as readily as for affirmative sentences, i.e., whether in less predictable contexts comprehenders have more difficulty taking into account all available information from partial input for negatives and interpreting it strictly incrementally compared to affirmatives. To do so, we used the number of possible completions to sentences within a particular context as a manipulable operationalisation of predictability, while ensuring that all affirmative and negative statements were pragmatically licensed in the experimental context. We achieved this by using visual displays to set up episodic contexts in which participants were to interpret each sentence, forming temporary associations between certain objects and their locations inside a clearly visible grid. This approach avoided the introduction of any confounding effects from long-term memory associations (such as prior world knowledge or familiarity), which are difficult to avoid in sentence processing studies relying on real-world knowledge, while enabling tight control of the manipulation. Furthermore, the game-like nature of the experimental set-up, with its constrained 'world' and limited number of sentence forms, meant that participants came to anticipate hearing both affirmative and negative sentences used to describe the visual scenes encountered episodically. In other words, the pragmatics of the experiment licensed the use of all sentences to an equal extent. (This use of a constrained and game-like scenario across the experiment in its entirety does not imply that participants resorted to special or non-naturalistic strategies during processing of each sentence itself. Individually, the descriptive nature of the linguistic stimuli means that their presentation may be compared to such real-world language uses as describing an event after observing it, or retelling a story.)

We employed a computer mouse-tracking methodology to tap into the time-course of participants' sentence processing and associated prediction-making or integration of new information. As exemplified in Dale and Duran's (2011) study reviewed above, this approach exploits the idea that when participants use a mouse or similar pointing device to respond to

stimuli, the trajectory followed by the cursor captures aspects of their cognition as they formulate and execute successive stages of the response, providing potentially rich information on initial and intermediate processing (see, e.g., Fischer & Hartman, 2014; Freeman & Ambady, 2009; Song & Nakayama 2009; Spivey & Dale, 2004, for reviews). Use of a computer mouse is not only routine and intuitive for participants, but also requires very little cognitive overhead, and participants perceive almost no task demands relating to the method of response. This is in contrast to some other behavioural measures that purport to provide similarly detailed levels of information, such as the signal-to-respond paradigm (e.g. Meyer, Irwin, Osman, & Kounios, 1988), which requires extensive training and imposes a high level of additional task demand. Additionally, mouse movements are very cheap and easy to record. A review of the theory underlying the mouse-tracking methodology and its application to various topics in memory and language is provided by Kent, Taylor, Taylor, and Darley (2017).

We present a series of three experiments in which we used this method to investigate the comparative effects of the number of possible sentence completions on incremental processing of affirmative and negative sentences. Experiment 1 used a truth-value judgement paradigm similar to that employed by Dale and Duran (2011) and in other mouse-tracking studies (e.g., Tomlinson, Bailey, & Bott, 2013). In Experiments 2 and 3, we sought to access participants' predictions more directly by using a sentence completion task; the findings of Experiment 2 were replicated in Experiment 3 using a speeded version of the task.

## Experiment 1

In Experiment 1, we aimed to examine the effects of polarity and number of possible true completions on incrementality of sentence processing using a truth-value judgement task. This methodological approach follows previous mouse-tracking studies (e.g., Dale &

Duran, 2011; Tomlinson et al., 2013), in which it has been found that participants' predictions for critical material in a sentence can be accessed by having them judge the sentence to be true or false. If the participant's initial prediction is accurate and the critical material matches it, they can readily judge the sentence to be true; conversely, if their initial prediction is accurate and the critical material mismatches it, they can readily judge it to be false. However, if the participant's prediction is inaccurate (e.g., because they have not incrementally updated it on the basis of the presence of negation), they may initially identify a true sentence as false, and vice versa, before updating their judgement on the basis of an evaluation of the whole sentence. For example, in the Dale and Duran (2011) study, participants may have initially predicted *large* as a likely completion for the sentence *Elephants are not…* as a result of failing to incorporate the negating element *not* into their online interpretation, hence their initial attraction towards a *True* response for the completion *large*. Using mouse-tracking, we aimed to capture this type of early cognition to identify cases in which participants were more likely to make mistaken predictions and thus to produce mouse trajectories exhibiting initial attraction towards the wrong answer. If certain conditions produced this type of attraction, we expected to observe two distinct subtypes of mouse trajectory: one consisting of a rather direct path from the starting point to the target response, arising from trials on which the correct response was immediately clear to the participant; and one consisting of a path initially deviating towards the incorrect response, followed by a course-correction after updating of this initial, mistaken attraction.

It is worth noting that, although we refer here and throughout this article to participants' *predictions*, the data does not necessarily reflect predictive pre-activation of a particular lexical item or concept. It may instead (or in a subset of cases, perhaps those in which prediction is more difficult) relate to participants' early processing in response to new information that has not been pre-activated. However, even in this case, such processing

occurs in the context of the participant's processing of immediately preceding information, and therefore reflects the extent to which they have been able to take the latter into account, in the same way that a prediction would.

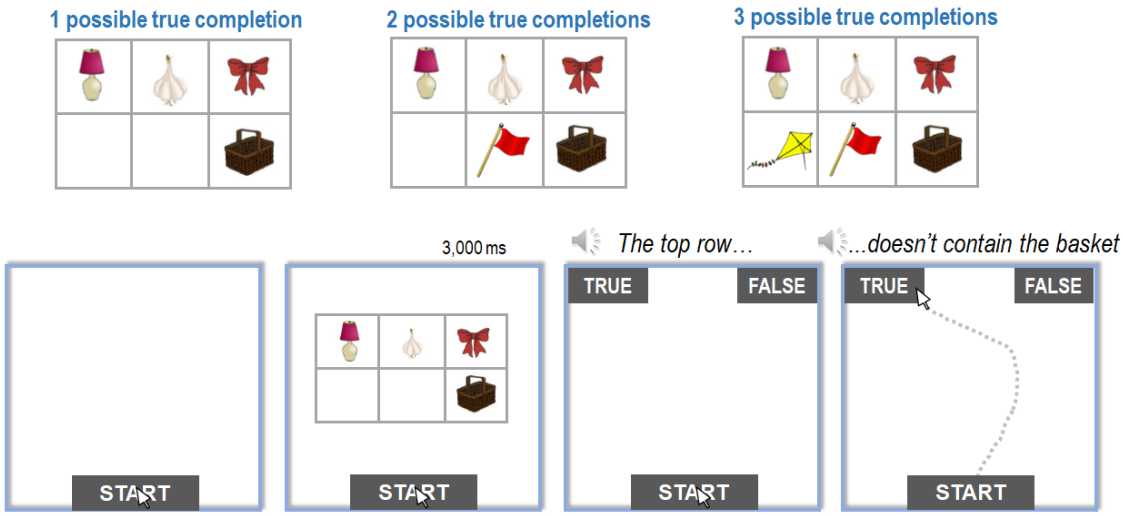


**Figure 1.** Example stimuli and trial structure in Experiment 1. *Top panels*: grids that could be shown during the visual context phase (exemplified in bottom panel 2); the condition presented here dictates the number of possible true completions for the subsequent sentence (basket; basket or flag; basket, flag, or kite). *Bottom panels*: the procedure during a trial. After the grid disappeared, the participant heard a true or false sentence and provided a truth-value judgment by moving the cursor from the START button to their chosen response. Images and text are for illustration and not shown to scale.

We provided an episodic visual context by presenting a visual stimulus (an array of objects in a 2 × 3 grid; see Figure 1) followed by an auditory sentence (e.g., *The bottom row contains the basket*). We then asked participants to indicate whether the sentence was true or false with respect to the visual context. In such sentences, the participant's earliest theoretical

opportunity to make a prediction about the critical word (the name of the object) is at the onset of either *contains* or *doesn't*; however, if the negative element of *doesn't* is not fully processed at this point, they may not be able to make an accurate prediction on negative trials, which may affect their initial truth value judgment of the sentence once the critical word is subsequently heard. In addition to the polarity of the sentence, the predictability of the critical word was manipulated by varying the number of objects in the grid whose names could complete the sentence to make it true; this variable was crossed with the polarity of the accompanying sentence (affirmative or negative, i.e. *The bottom row contains/doesn't contain the basket*). Because of the nature of the task, it was also necessary to manipulate the truth value of the sentences.

We hypothesised that participants would have more difficulty making incrementally-updated predictions for sentences with more possible true completions, and, if a two-stage mechanism were at work, that they would specifically generate mistaken predictions in the case of negatives with more possible true completions. This effect would apply specifically in conditions involving more possible true completions if it is the case that predictability is a critical factor governing incrementality of processing for negative sentences. Thus, on the basis of this hypothesis, we expected (i) lower response accuracy, slower initiation and completion of responses, and greater attraction in mouse trajectories to the foil response option across the board as the number of possible true completions increased (i.e., performance would be poorer on trials with two possible true completions than on those with one, and poorer again on those with three); and (ii) this detriment in performance to be disproportionately large (especially in regard to the degree of attraction to the foil response option) for negatives compared to affirmatives.

## Method

### *Participants*

Twenty-four native English speakers (21 female, aged 18–24 years [*M* = 19.5, *SD* = 1.7]) from in and around the University of Bristol community volunteered to participate in exchange for course credit or a payment of £6. Most were monolingual, but four had a second language of a good, advanced, or fluent standard. Ethical approval for all three experiments was granted by the University of Bristol's Faculty of Science Research Ethics Committee, and written informed consent was obtained from all participants. The sample size of *N* = 24 was set in advance.

### *Materials*

Visual stimuli consisted of images of highly-identifiable everyday inanimate objects (see Figure 1 for examples), selected for consistent naming by a small sample of British English speakers. On each trial, several images were presented in a 2 × 3 grid, accompanied by a sentence presented auditorily, of the form *The top / bottom row contains / doesn't contain the basket*. One row of the grid always contained three images (in Figure 1, the lamp, garlic, and bow on the top row), while the other row could contain one, two, or three images (in the Figure 1 example, it always contained the basket, and could additionally contain the flag, or the flag and the kite). In this way, we manipulated the predictability of the critical object name that could appear at the end of the sentence to make it true. Although, in principle, an infinite number of completions are generally hypothetically available to render a negative sentence true, in practice only a small number of possible completions are pragmatically relevant, and in the case of this experiment, possible completions were restricted to objects that had actually appeared in the preceding grid (a fact which we assume participants rapidly became accustomed to). The use of the definite article *the* also strongly

implied that a new object would not be introduced into the discourse in this part of the sentence. Therefore, we refer to this variable as reflecting the *number of possible true completions* even though it does not always enumerate all theoretical possibilities.

We constructed 72 sets of visual stimuli, each consisting of 18 versions: six with four objects, six with five objects, and six with six objects (see Figure 1). Each of the 18 versions within each set could be associated with four possible sentences: affirmative or negative, and true or false. Thus, each set of images was associated with 18 × 4 = 72 possible trials, of which half were critical trials and half were trials used for counterbalancing purposes. Critical trials were equally distributed across conditions in a 2 (polarity: affirmative or negative sentence) × 3 (number of possible true completions: one, two, or three) × 2 (truth value: true or false sentence) design. Further details on the design and counterbalancing are provided in the Supplementary Materials.

Audio recordings of the sentences were prepared using the NaturalReader software package (NaturalReader Version 11, NaturalSoft Ltd, 2015), employing a female British English speaker's voice with natural-sounding prosody. The audio files were manipulated using the Audacity recording and editing software (Audacity 2.12, Audacity Team, 1999–2016) to homogenise the time elapsing from the onset of the sentence fragment to the onset of the critical word that would allow the participant to distinguish an affirmative and a negative sentence (i.e., *contains* or *doesn't*).

*Procedure*

Trials were presented using the MouseTracker software package, Version 2.82 (Freeman & Ambady, 2010). Figure 2 shows an example trial. Each trial began with a 'START' button presented in the centre at the bottom of the screen; participants were required to click here to initiate each trial, so that the mouse and cursor were reset to the same

starting position. The cursor was then temporarily held at this location while the visual stimulus (i.e., the grid of objects) was displayed for 3,000 ms, plus a further 1,500 ms for every additional object beyond four (i.e., 4,500 ms for five-object trials and 6,000 ms for six-object trials). Next, the grid disappeared and, simultaneously, auditory presentation of the sentence began. Also at the same time as the onset of the audio, the response options ('TRUE' and 'FALSE') appeared in the top-left and top-right corners of the screen, respectively, and the cursor was released to allow the participant to complete their response when ready. The participant was required to indicate whether the sentence they had heard was true or false with respect to the grid displayed on that trial by clicking on the corresponding word, and to initiate their response (by beginning a mouse movement) within 5,000 ms of the cursor release; if they failed to do so, a warning message appeared.

The grid did not remain onscreen once the auditory stimulus had begun, because our design required that participants consult a pre-formed episodic association in memory when interpreting the linguistic input. For example, viewing of the grid on a particular trial may require them to form a temporary association between the concepts of the *bottom row* and the *basket*, which would then cause activation of *basket* when the *bottom row* subsequently featured in the sentence. If the grid had remained onscreen during presentation of the sentence, participants would have had little reason to formulate this type of episodic association in advance, instead waiting to hear the sentence before consulting the grid to evaluate it; therefore, they would have been unable to make predictions about upcoming material based on these associations, and their processing of further incoming material would have been unaffected by them, thus nullifying the effects of the experimental manipulation.

Mouse coordinates were sampled online every 32 ms during the response phase of each trial, based on a virtual coordinate space ranging from –1 to 1 in the *x* axis and 0 to 1.5 in the *y* axis, with the origin at the horizontal centre of the bottom of the screen. Response

initiation and completion times and accuracies were also collected. After completing their response, the participant received feedback in the form of a green O (correct) or red X (incorrect) displayed in the centre of the screen for 300 ms.

For each participant, 30 trials (selected pseudo-randomly, with 10 for each number of possible true completions) were followed by a memory test, as a strategy to encourage participants to pay careful attention to every object in the grids and to explore the extent of variation in the difficulty of the task according to the number of objects present. On such trials, a prompt appeared on-screen after the feedback stage asking the participant to recall the objects that had been presented in the grid and write their names in the appropriate locations on a pre-printed grid provided. If a participant could remember that an object appeared at a particular location, but not its identity, they could indicate this with an 'X'. There was no time limit for responses to the memory test.

Participants were tested individually (although in some cases simultaneously) in sessions that lasted approximately 60 minutes. The participant was seated at a comfortable distance from the screen, wore headphones, and used a standard USB laser mouse to give responses. The experiment began with a verbal explanation of the task from the experimenter, accompanied by four practice trials (two with memory tests). Participants were asked to look carefully at every item in the grid on each trial and to focus on clicking on the correct response button ('TRUE' or 'FALSE'). After the practice trials, each participant completed eight blocks of 27 randomly-ordered trials. They were encouraged to take a break between blocks.

*Data preparation and analysis*

Data and analysis scripts for all three experiments are available from the first author upon request.

Across all participants, the range of error rates was 3% to 22%. Counterbalancing trials and trials with incorrect responses (9% of all critical trials) were discarded from the analyses involving response time and trajectory measures. Those with response times longer than 6,000 ms or initiation times longer than 4,000 ms (a further 3% of the remaining trials) were also excluded. These thresholds were selected (in a pre-specified procedure) based on visual inspection of the positively-skewed distributions of response and initiation times, in order to trim the data at a point that would produce an approximately symmetrical distribution; the overall pattern of findings did not materially change when the same analyses were conducted without any trimming (see Supplementary Materials for further details). For ease of comparison across all trials, data from trials with the correct answer 'TRUE' (i.e., for which the target response was located on the left side of the screen) were reflected (all *x* coordinates were sign-reversed) using the MouseTracker software package's inbuilt analysis features. All subsequent cleaning and analysis procedures were carried out using R (R Core Team, 2018), and figures were produced using the ggplot2 package (Wickham, 2016).

In addition to response times and accuracy, we analysed mouse trajectory shapes. To characterise the shape of a path followed by the cursor from the starting position to the response position, generally useful measures are the area under the curve (AUC) and the maximum deviation (MD). The former represents the total area of the screen lying between an ideal straight-line trajectory drawn between the starting and finishing points and the actual trajectory taken, and the latter the distance at the furthest point between this ideal straight-line trajectory and the actual trajectory. However, neither the AUC nor the MD necessarily captures all the most pertinent characteristics of a given trajectory, given that they each collapse a rich set of information into a single value and do not give an indication of the specific manner in which a trajectory deviates. Furthermore, averaging these values across trials may result in the loss of distinct subtypes of trajectories. Therefore, we used a

clustering approach as a data-driven method of assessing whether a subset of trials exhibiting attraction to the foil response was present. Trajectories were assigned to clusters based on their *x* and *y* coordinates at each time point, using the Hartigan and Wong (1979) implementation of the k-means algorithm with $k = 2$, stopping after 1,000 iterations (or at convergence if this came first), and selecting the best outcome from the output of 1,000 randomly-generated starting points. Clustering was carried out by participant, across all conditions. After this process was complete for every participant's data, the data were re-integrated for analysis of cluster allocation, so that a given participant's cluster of trials represented by the cluster centre most similar to a straight line was treated as equivalent to that of other participants.

Response accuracy, response time, and trajectory cluster allocation were each modelled separately to analyse the effects of number of possible true completions, polarity, and truth value on these measures, using the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015); degrees of freedom and *p* values for statistical tests on coefficients were computed using lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017). For proportional data (i.e., response accuracy and cluster allocation), the glmer function was used to construct a mixed effects logistic regression model with a binomial family error distribution and logit link function; for continuous data (i.e., initiation and response time), the lmer function was used to construct a mixed effects linear regression model.

Although true and false sentences were evenly distributed across the other conditions and truth value was not of theoretical interest, we included truth value in the analysis because we expected it to interact with the independent variables of interest. In particular, there is reason to expect an interaction between truth value and polarity: false affirmative sentences are more difficult for participants to verify than true affirmatives, but this difference is less extreme (or even reversed) for judgements of negatives (as observed, for example, by Gough,

1965; Wason, 1959, 1961). Therefore, terms representing the main effects of the number of possible true completions (*n-completions*), polarity, truth value, and the interactions among these variables were entered into each model.

For each dependent variable, a full model was constructed including fixed factors for all the above effects, as well as the maximal random effects structure that allowed the model to converge, for example:

DV ~ polarity * n-completions * truth-value + (polarity|participant) (1)

Next, model comparisons were conducted to investigate the presence of main effects and interactions, using the following procedure. First, the maximal model was compared to models including only the fixed effects of, and interactions between, each of only two factors, e.g. polarity × n-completions:

DV ~ polarity * n-completions + (polarity|participant) (2)

If model (1) represented a significantly better fit to the data, this was taken as an indication of either a main effect of the omitted factor, or an its involvement in an interaction. Second, for each factor whose omission meant that the maximal model provided a significantly better fit, its involvement in an interaction specifically was tested by comparing the maximal model to a model including the fixed effects of, and interaction between, each of the other two factors, plus a fixed effect of the factor in question, e.g.:

DV ~ polarity * n-completions + truth-value + (polarity|participant) (3)

If the maximal model represented a significantly better fit to the data than the more restricted model, this was taken as an indication that the relevant factor was involved in an interaction with at least one other factor. Finally, if the above tests indicated the presence of interactions, the maximal model was compared to a model including fixed effects of all three factors, plus each of the possible two-way interactions, e.g.:

DV ~ (polarity * n-completions) + (polarity * truth-value) + (n-completions * truth-value) + (polarity|participant) (4)

If the former represented a significantly better fit to the data, this was taken as an indication of the presence of a three-way interaction.

Following these tests, the maximal model (as recommended by Barr, Levy, Scheepers, & Tily, 2013) was used to estimate coefficients for the simple effects of each factor at each level of the other factors. Associated 95% confidence intervals (using the Wald method) and *p* values (using the Satterthwaite approximation) were computed.

**Results**

*Response accuracy.* Response accuracies (Figure 2, left panel) were modelled across all conditions with a full set of fixed factors and a random factor of polarity by participant. Model comparisons testing for the presence of main effects or interactions were significant for polarity, $\chi^2(6) = 36.9$, $p < .001$, number of possible true completions, $\chi^2(8) = 95.8$, $p < .001$, and truth value, $\chi^2(6) = 36.3$, $p < .001$. Model comparisons testing specifically for involvement in an interaction were also significant for polarity, $\chi^2(5) = 11.4$, $p = .044$, number of possible true completions, $\chi^2(6) = 16.2$, $p = .013$, and truth value, $\chi^2(5) = 18.6$, $p = .002$. Finally, the model comparison testing for the full three-way interaction indicated that this was present, $\chi^2(2) = 6.2$, $p = .045$. Coefficients estimated for the maximal model indicating the simple effects of each factor are given in Supplementary Table 1. Overall, the model parameters indicated that negation, more possible true completions, and falsity generally had a detrimental effect on accuracy, but there were simple interactions indicating that the detrimental effect of falsity was reduced for negative sentences, especially those with a single possible true completion (ß = 1.21, $z = 3.03$, $p = .002$, 95% CI = [0.43, 2.00]); and that the effect of increasing the number of possible true completions from two to three

combined with falsity was enhanced for negative sentences (ß = –0.67, $z$ = –2.49, $p$ = .013, 95% CI = [–1.19, –0.143]).

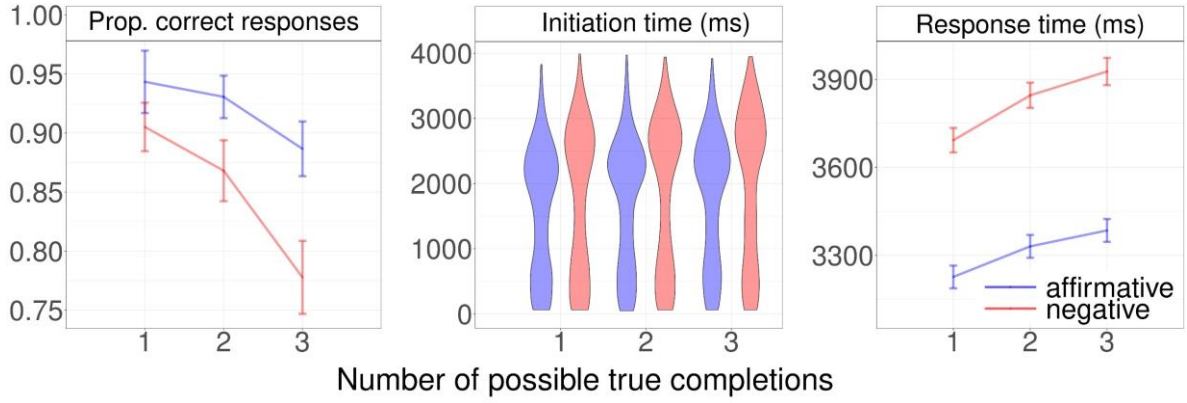


**Figure 2.** The results of Experiment 1 across conditions for three dependent variables. Left panel: mean proportion of correct responses; middle panel: the distribution of response initiation times; right panel: mean response time. Error bars in the left and right panels represent standard errors around the mean.

*Initiation time.* Across the full dataset, response initiation times were distributed bimodally, as confirmed by a significant Hartigan's dip statistic ($D$ = .06, $p$ < .001), with a main peak at 2,500 ms and secondary peak at around 578 ms, precluding the use of inferential statistical tests. The distribution of initiation times in each condition is shown in Figure 2 (middle panel).

*Response time.* After trimming as described above, the distribution of all response times still exhibited a positive skew around a peak located at 3,347 ms. However, despite a significant Hartigan's dip statistic ($D$ = .001, $p$ = .001), there was no clear second peak. Mean response times (Figure 2, right panel) were modelled across all conditions with a full set of fixed factors and a random factor of polarity by participant. Model comparisons indicated main effects or involvement in interactions for polarity, $\chi^2(6) = 102.4$, $p$ < .001, number of possible true completions, $\chi^2(8) = 118.6$, $p$ < .001, and truth value, $\chi^2(6) = 85.7$, $p$ < .001.

Further comparisons testing specifically for involvement in an interaction were significant for polarity, $\chi^2(5) = 32.4$, $p < .001$, and truth value, $\chi^2(5) = 27.2$, $p < .001$, but not for the number of possible true completions, $\chi^2(6) = 6.0$, $p = .421$. Coefficients estimated for the full model indicating the simple effects of each factor are given in Supplementary Table 2. Overall, the model parameters showed that falsity, negative polarity, and increasing the number of possible true completions all generally slowed down response completion. However, there were simple interactions indicating inconsistent effects of falsity, with a reduced effect of falsity for negative sentences in conditions with one (ß = –204.3, $t(4589.6) = –3.89$, $p < .001$, 95% CI = [–307.2, –101.4]), two (ß = –129.2, $t(4575.9) = –2.42$, $p = .016$, 95% CI = [–233.8, –24.6]), and three (ß = –138.2, $t(4374.2) = –2.48$, $p = .013$, 95% CI = [–247.2, –29.1]) possible true completions.

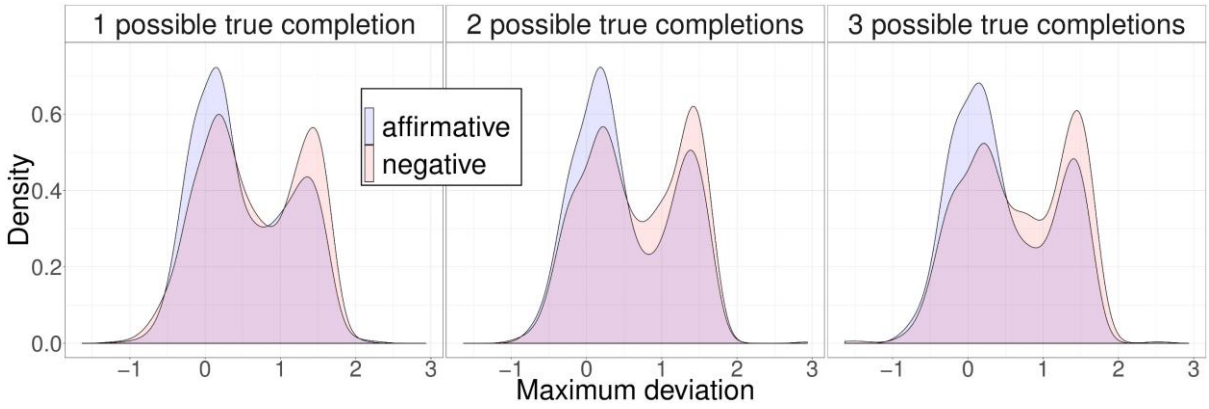


**Figure 3.** Distribution of the maximum deviation across trials in Experiment 1, for trials with different numbers of possible true completions. Maximum deviation is calculated as the distance, at the furthest point, between a given trajectory and the straight line between its starting and finishing points, in arbitrary units representing screen coordinates. Densities are kernel density estimates.

*Mouse trajectories.* In terms of trajectory shapes, the distribution of all MDs exhibited strong bimodality across all conditions (Figure 3). The use of trajectory clustering, with the number of clusters fixed at 2, was therefore deemed an appropriate way of analysing the comparative proportion of trials in each condition falling into a cluster characterised by more direct trajectories vs. trajectories exhibiting significant attraction toward the incorrect response. Clustering was carried out as described in the data preparation and analysis section; Figure 4 (left panel) illustrates the resulting allocation of trials to each cluster. The proportion of trials falling into the direct vs. the foil-skewed cluster was modelled across all conditions with a full set of fixed factors and a random effect of polarity by participant. Model comparisons indicated main effects or involvement in interactions for polarity, $\chi^2(6) = 23.0$, $p < .001$, number of possible true completions, $\chi^2(8) = 77.5$, $p < .001$, and truth value, $\chi^2(6) = 58.5$, $p < .001$. Model comparisons indicating interactions specifically, however, were not significant for polarity $\chi^2(5) = 5.1$, $p = .404$, number of possible true completions, $\chi^2(6) = 9.80$, $p = .133$, or truth value, $\chi^2(5) = 8.2$, $p = .147$, indicating that only main effects were present. Coefficients estimated for the full model indicating the simple effects of each factor are given in Supplementary Table 1, showing that falsity, negative polarity, and increasing the number of possible true completions all broadly increased the likelihood of trials exhibiting significant deviation towards the foil object; this pattern is illustrated in Figure 4 (right panel). The only significant simple interaction was an enhanced effect of falsity when the number of possible true completions was increased from two to three for affirmative sentences (ß = –0.48, $z = –2.08$, $p = .038$, 95% CI = [–.094, –0.03]).

A concern for the interpretation of the trajectory data using a clustering approach might be that trials falling into the foil-skewed cluster, rather than representing any particular attraction to the foil response, might simply represent trials where the participant happened to initiate their response more quickly by making an initial guess at the correct answer and

updating their response part-way through the trajectory. This would mean that trials falling into the direct cluster would disproportionately represent those on which the participant hesitated before initiating their response, masking any attraction to the foil. To rule out this interpretation, initiation times for trials falling into each cluster were compared. A paired samples *t* test showed that there was no significant difference in initiation time between clusters, $t(23) = -1.81$, $p = .08$, and the mean initiation time across all conditions was in fact greater for trials falling into the foil-skewed cluster.

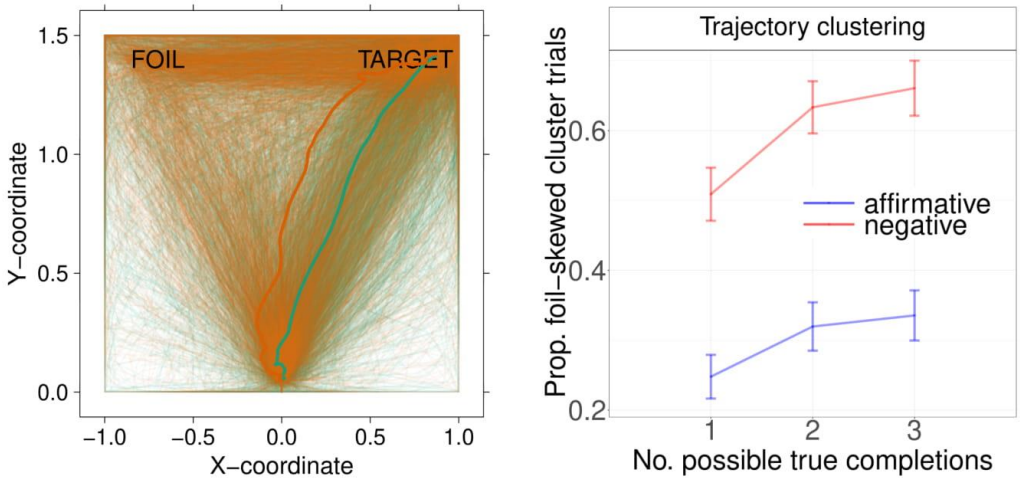


**Figure 4.** Results of the trajectory analysis in Experiment 1. The left panel shows a representation of the trajectories taken on all included trials, colour-coded according to cluster. Thick lines represent the cluster centres. The right panel shows the mean proportion of trials allocated to the cluster representing greater deviation from a straight line (stronger attraction to the foil stimulus), in each condition. Error bars represent standard errors.

*Memory test.* Participants' performance on the memory test accompanying a subset of trials was scored by awarding 4 points for each correct item in the correct location, 2 points for each correct item positioned in the wrong location, and 1 point for each incorrect item or 'X' positioned in a location that had contained an item. The total number of points was then

divided by the maximum available for that trial (16, 20, or 24 for 4-object, 5-object, and 6-object trials, respectively). As a percentage of the maximum points available, participants on average scored 76% on 4-object trials, 71% on 5-object trials, and 65% on 6-object trials. A one-way analysis of variance showed that the effect of number of objects was significant, $F(2, 23) = 18.3$, $p < .001$. This indicates that, despite the differences in display duration, the number of objects did affect participants' memory for the display, as expected. However, even when six objects were present, it is evident that participants were still able to recall the majority of them and were clearly focused on the task. Nevertheless, the effects of the number of possible true completions observed in Experiment 1 remained confounded with any effects arising from the number of objects present.

**Discussion**

Two findings were strongly in evidence in Experiment 1: the main effects of polarity and of the number of possible true completions. Both negative polarity and increasing the number of possible true completions had detrimental effects on all dependent variables, causing participants to respond less accurately, take longer to initiate and complete their responses, and produce trajectories that were more likely to veer towards the foil answer. The effect of truth value on their responses was less clear: falsity had an inconsistent, detrimental main effect only on response time and the proportion of foil-skewed trajectories.

In spite of the strength of the principal main effects, the pattern of interactions between variables was less clear. There was some evidence of a two-way truth value × polarity interaction, as hypothesised (affecting response time, with a reduced effect of negation for false sentences) and the suggestion of a three-way interaction among all the independent variables (affecting response accuracy). However, the lack of a clear pattern makes it difficult to interpret these findings as reliable evidence for any particular conclusion.

Based on previous findings by Nieuwland and Kuperberg (2008) and Dale and Duran (2011), alongside the notion that predictability might function as the 'active ingredient' of pragmatic felicity in terms of its interaction with polarity, the hypotheses were that no main effect of polarity would be observed, but that an interaction between polarity and the number of true completions (i.e., our operationalisation of predictability) would show that low predictability hinders incremental processing of negation, generating initially mistaken predictions to a much greater extent than these would appear for affirmatives. Neither of these hypotheses was supported by the evidence.

First, the main effects of polarity and the number of possible true completions suggest that neither pragmatic felicity nor predictability can account for all the processing difficulties associated with negation. All sentences presented in Experiment 1 were equally pragmatically felicitous (within the episodic context provided by the visual image and the experimental task). Despite this, negation consistently rendered participants slower, less accurate, and more attracted to the foil response, showing that even these licensed negatives were not as easy to respond to as affirmatives.

Second, the inconsistent pattern of interactions means that it is difficult to say whether there was an underlying interaction between polarity and the number of possible true completions in any aspect of processing, as hypothesised. If such an interaction was present, it was very weak (although in the same direction as the hypothesis, i.e., with a reduced impact of negation in sentences with fewer possible true completions), and arguably masked by two limitations of the experimental design and data analysis: first, the confounding effects of truth value, and second, the weakness of the cluster analysis.

Regarding the confounding effect of truth value, the interaction between truth value and polarity was problematic for the interpretation of Experiment 1, because truth value was manipulated not as a variable of theoretical interest, but purely as a mechanism to produce a

task that would tap into participants' ability to process the sentences incrementally and update their predictions for (or otherwise respond more readily to) upcoming material. However, it was not possible to collapse across truth value in analysing the results because of its involvement in interactions with the other independent variables, meaning that it was difficult to disentangle the effects of the variables of interest. Tian and Breheny (2015, p. 33) have recently highlighted the asymmetry of true and false answers in the truth value judgment task: 'Generally, language is processed on the assumption that a statement is true, or at least relevant. […] Verification is a metalinguistic task that normally requires establishing what would be the case if the sentence were true, and comparing that to evidence. […] Based on this idea, […] we should expect that true statements, whether positive or negative, take less time to verify.'

Regarding the weakness of the cluster analysis, Figure 4 shows that the clustering approach did not distinguish clearly between large deviations and relatively direct trajectories. This outcome can probably be attributed to two main factors: first, the relatively slow overall response times, meaning that not all stages of initial processing may have been captured by the trajectories; and second, an inadequate number of datapoints given the approach to clustering employed. Our data-driven approach was rather conservative, in that a classification of trials by hand would likely have produced a clearer differentiation between the two clusters, thereby strengthening the effects observed; this choice traded off power to detect effects in favour of objectiveness and replicability. Both these factors imply that the cluster analysis of trajectories is likely to have underestimated the size of any relevant effects.

As a result of the above limitations and the overall lack of clarity provided by the results, we sought in Experiments 2 and 3 to employ a task that would avoid the confounding effects of manipulating truth value, elicit faster responses, and allow the collection of more

datapoints, thus improving the power of the experimental design to capture and detect underlying effects in the shape of the response trajectories.

## Experiments 2 and 3

In Experiments 2 and 3, we aimed to avoid the problems for the experimental design of Experiment 1 arising from the use of the truth value judgement paradigm. In order to tap more directly into participants' predictions for and responses to upcoming material, we presented auditory sentence fragments instead of full sentences, leaving out the critical material, i.e., the name of the object. The task was then to select the response option that would best complete the sentence. In a similar way to Experiment 1, we expected participants to exhibit initial attraction towards the foil response option, correcting their course mid-way through the trial, in cases where they had made an erroneous prediction or otherwise engaged in processing that made the foil initially easier to integrate. This approach offered several advantages over the truth-value judgement methodology, in that the design was simpler; participants' predictions could be measured more directly; there was no task-specific reason for participants to calibrate their predictions for a mixture of true and false sentences; the interaction between truth value and polarity was eliminated; and the allocation of the target and foil to the left- and right-hand side of the screen could be randomised.

In Experiment 1, predictability and memory load were confounded in that images associated with trials with more possible true completions always contained more objects and thus imposed a higher memory load than trials with fewer possible true completions. The design of Experiments 2 and 3 allowed us to decouple memory load from predictability, and thereby to test the effects of memory load, through the inclusion of control trials on which the number of objects in the non-mentioned row was varied. On such trials, the number of

possible completions was always low (because the mentioned row always contained three items, as in critical trials with three possible completions), but the memory load incurred by storing all the objects in the visual stimulus varied. This allowed us to disentangle the effect of the number of possible completions from that of a simple increase in memory load. The design was identical for Experiments 2 and 3, except that in the latter case, we increased the pressure on participants to respond quickly by reducing the amount of time available for their response before a warning message was presented and by presenting an auditory response signal that created an impression of urgency. We expected that this would allow us to capture more of participants' early processing and cognition within their mouse trajectories and aimed to investigate whether the same pattern as observed in Experiment 2 would hold under these conditions.

In both experiments, the hypothesis was the same as for Experiment 1: namely, that we would observe lower response accuracy, slower initiation and completion of responses, and greater attraction in mouse trajectories to the foil response option as the number of possible completions increased, with a disproportionately large effect (especially in the latter dependent measure) for negatives compared to affirmatives.

## Method

### *Participants*

Participants in Experiments 2 ($N$ = 24;18 female, aged 18–22 years [$M$ = 19.7, $SD$ = 1.1]) and Experiment 3 ($N$ = 32 after exclusion of one participant with an anomalously high error rate in responses; 23 female, aged 18–29 years [$M$ = 23.5, $SD$ = 4.5]) were recruited from in and around the University of Bristol community. None had participated in Experiment 1, and none participated in both Experiments 2 and 3. Most were monolingual,

but 21 (11 in Experiment 2, 10 in Experiment 3) had a second language of a good, advanced, or fluent standard. Both sample sizes were set in advance.

*Materials*

Visual stimuli in both experiments were of the same type as in Experiment 1 and were constructed in a similar way. For Experiments 2 and 3, 216 highly-identifiable objects were distributed with equal frequency across 216 different 2 × 3 grids, each containing four, five, or six objects; as before, one row contained three objects and the number in the other row was manipulated (Figure 5). Further details on the design and counterbalancing scheme are provided in the Supplementary Materials.

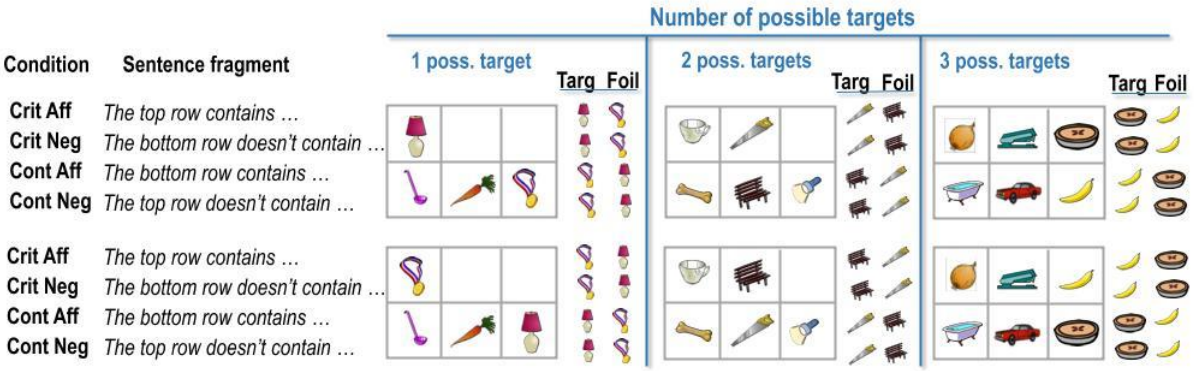


**Figure 5.** The experimental design and counterbalancing scheme used in Experiments 2 and 3, with a set of example trials at each level of predictability (i.e., the number of possible targets).

Each visual stimulus could be accompanied by one of four sentence fragments, of the form: 'The top / bottom row contains / doesn't contain…'. The number of possible objects whose name would complete the sentence truthfully was one (the target), two (the target or a single alternative), or three (the target or two alternatives), depending on the number of images in the grid. Thus, predictability of the sentence fragment (i.e., the number of possible

true completions) was manipulated as in Experiment 1; in this case, possible true completions were equivalent to possible targets that could appear as correct response options. Each pair of grids was therefore associated with 2 × 4 = 8 possible trials, of which half were critical trials and half were control trials in which the row referred to always contained three objects across all conditions (rather than a varying number), meaning that the number of possible targets was always three. No participant was presented with more than one trial from this set of eight. Across sets of visual stimuli, critical trials were equally distributed across conditions following a 2 (polarity: affirmative or negative sentence fragment) × 3 (number of possible targets: one, two, or three) factorial design, yielding six critical conditions. In total, each participant completed 216 of the 3,456 trials constructed; half of these were distributed evenly across the six critical conditions, and the other half were controls as described above, also distributed evenly across the equivalent conditions. Each participant additionally completed 108 filler trials, in which four, five, or six objects were distributed randomly in the grid and the sentence fragment took the form: "The left / right side contains / doesn't contain…". Audio recordings of the sentence fragments were prepared as in Experiment 1. For use in Experiment 3 only, an auditory response signal lasting 2,000 ms was also prepared (see Procedure).

*Procedure*

Trials in both experiments were presented in the same way as in Experiment 1, except for the following modifications (Figure 6). The response options consisted of images of two objects that would complete the sentence fragment to make it either true (target) or false (foil), each randomly presented in either the top-left or the top-right corner of the screen. The participant was required to click on the object that would complete the sentence fragment truthfully. The response options appeared at the same time as the onset of the critical word

(*contains* in the affirmative condition or *doesn't* in the negative condition), which occurred between 690 and 902 ms after the start of the audio, depending on its contents, and the mouse cursor was also released at this time. No memory test trials were included in Experiment 2 or 3. The location of the cursor was sampled every 20 ms during the response phase.

In Experiment 3, the participant was required to complete their response to each trial within 2,000 ms of the onset of the critical word; following trials on which they failed to do so, a warning message was displayed. Additionally, an auditory response signal lasting 2,000 ms was superimposed onto each sentence fragment. This began with a brief tone coinciding with the onset of the critical word, followed by ten tones increasing in frequency and amplitude.

The testing procedure in both cases was the same as for Experiment 1, except for in the instructions (the participant was asked to look carefully at every item in the grid on each trial and to focus on selecting the correct completion for the sentence fragment as quickly and accurately as possible) and the number of practice trials (two). After the practice trials, each participant completed nine blocks of 36 randomly-ordered trials. The experiment took approximately 60 minutes to complete.

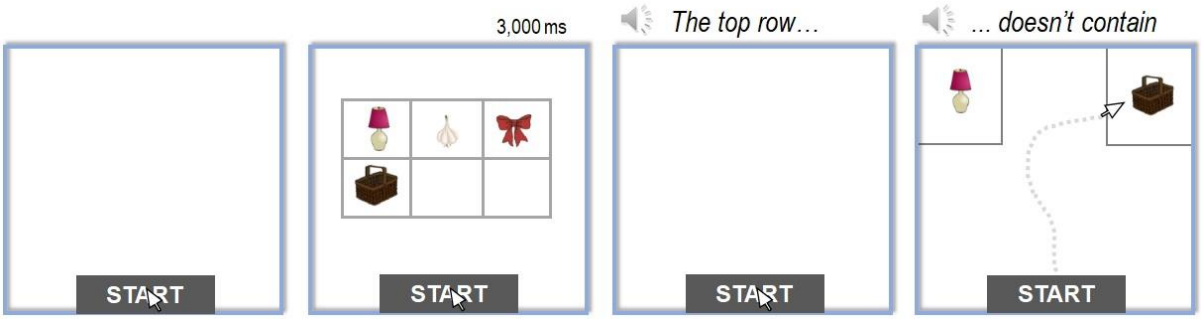


**Figure 6.** Example trial structure for Experiments 2 and 3. The participant clicks on the START button to begin the trial and on the target or foil image to give a response based on which would complete the sentence fragment accurately. Images and text are for illustration and not shown to scale.

*Data preparation and analysis*

Filler trials were excluded from all analyses, as were trials with incorrect responses (Experiment 2: 15% of all trials; Experiment 3: 12% of all trials). Control trials were excluded from the main analyses, but included in a secondary analysis testing the effects of total number of objects in the display vs. number of possible targets specifically. In Experiment 2, trials with initiation times longer than 2,000 ms or response times longer than 4,000 ms (a further 6% of the remaining trials) were also excluded; in Experiment 3, these thresholds were 1,000 ms and 2,000 ms, respectively, meaning that 8% of the remaining trials were discarded. These thresholds were selected in the same way as for Experiment 1; again, running the analyses without any such trimming did not materially affect the overall results patterns (see Supplementary Materials for further details). In both cases, X-coordinate data from trials in which the target response was located on the left side of the screen were reflected in the same way as in Experiment 1. All other aspects of data preparation and computation of dependent measures were identical to those used in Experiment 1.

The same approach to modelling each of the dependent measures and estimating simple effects was also employed as in Experiment 1, except that because truth value was not manipulated in Experiment 2 or 3, it was not included as a factor in any analyses. The procedure for model comparison was therefore simplified due to the presence of only two factors.

**Results**

*Response accuracy.* Response accuracies (Figure 7, left panels) were modelled across all critical conditions with a full set of fixed factors. The model for Experiment 2 also included a random intercept by participants, and the model for Experiment 3 a random factor of polarity by participant. For Experiment 2, model comparisons indicated significant main

effects of polarity, $\chi^2(3) = 17.6$, $p < .001$, and number of possible targets, $\chi^2(4) = 32.2$, $p < .001$, but no interaction between them, $\chi^2(2) = 3.8$, $p = .15$. For Experiment 3, the same main effects were present, $\chi^2(3) = 54.9$, $p < .001$, and $\chi^2(4) = 32.9$, $p < .001$; in addition, the interaction between polarity and number of possible targets was also significant, $\chi^2(2) = 16.6$, $p < .001$. Coefficients estimated for the full model in each experiment, indicating the simple effects of each factor, are given in Supplementary Tables 3 and 4. Overall, the model parameters showed that there were consistent detrimental effects of negative polarity and of increasing the number of possible targets in both cases, with a stronger effect of the latter on affirmative than on negative sentence fragments in the case of Experiment 3.

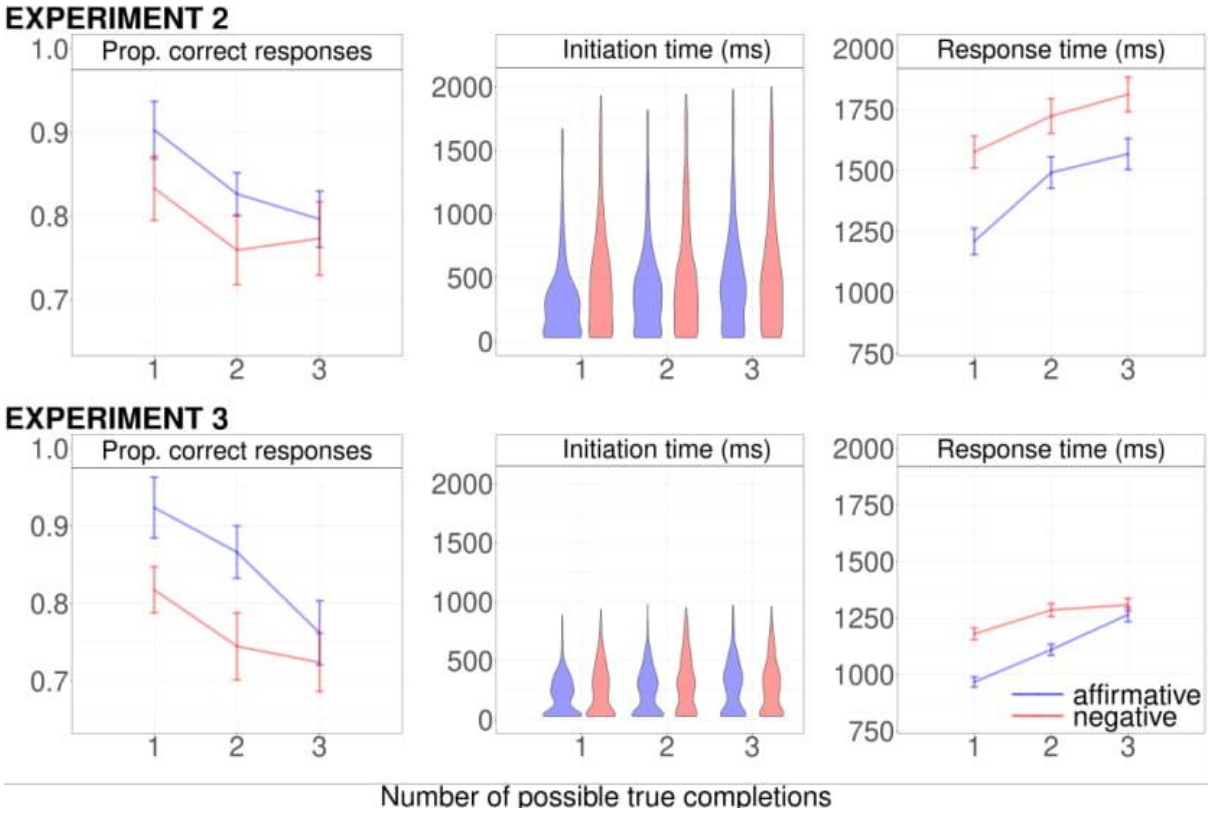


**Figure 7**. The results of Experiments 2 and 3 across critical conditions for three dependent variables. Left panel: mean proportion of correct responses; middle panel: the distribution of response initiation times; right panel: mean response time. Error bars in the left and right panels represent standard errors around the mean.

*Initiation time.* Response initiation times were distributed bimodally, as confirmed by a significant Hartigan's dip statistic; Experiment 2: $D = .046$, $p < .001$, with peaks at 333 and 63 ms; Experiment 3: $D = .05$, $p < .001$, with peaks at 39 ms and 306 ms. The distribution of initiation times in each condition is shown in Figure 7 (middle panel).

*Response time.* In both experiments, the distribution of response times was somewhat skewed around a peak (Experiment 2: at 1,118 ms; Experiment 3: at 1,023 ms); despite a significant Hartigan's dip statistic in both datasets (Experiment 2: $D = .025$, $p = .03$; Experiment 3: $D = .003$, $p = .015$), there was no second peak in either case. Response times (Figure 8, right panels) were modelled across all critical conditions with a full set of fixed factors. The model for Experiment 2 also included random effects of polarity and number of possible targets by participant; for Experiment 3, the model included only a random effect by participant of the number of possible targets. For Experiment 2, model comparisons again indicated significant main effects of polarity, $\chi^2(3) = 36.1$, $p < .001$, and number of possible targets, $\chi^2(4) = 40.9$, $p < .001$, but no interaction, $\chi^2(2) = 4.7$, $p = .09$. For Experiment 3, the same main effects were present, $\chi^2(3) = 60.0$, $p < .001$, and $\chi^2(4) = 62.3$, $p < .001$, but a significant interaction between polarity and number of possible targets was also observed, $\chi^2(2) = 22.8$, $p < .001$. Coefficients estimated for the full model indicating the simple effects of each factor are given in Supplementary Tables 5 and 6; overall model parameters followed the same pattern as for initiation time.

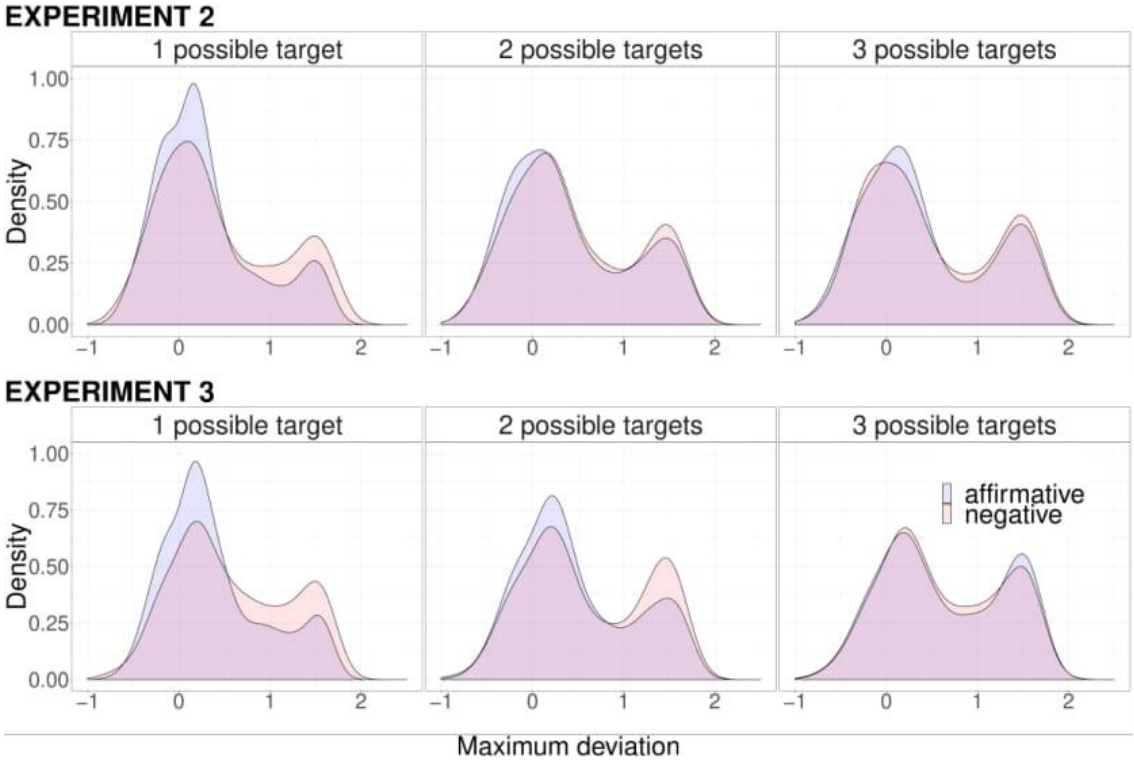


**Figure 8**. Distribution of the maximum deviation across trials in Experiments 2 and 3, for trials with one, two, and three possible targets. Maximum deviation is calculated as the distance, at the furthest point, between a given trajectory and the straight line between its starting and finishing points, in arbitrary units representing screen coordinates. Densities are kernel density estimates.

*Trajectories.* As in Experiment 1, the distribution of MDs exhibited strong bimodality across all critical conditions in both experiments (see Figure 8). The same approach to trajectory clustering was therefore employed; Figure 9 illustrates the resulting allocation of trials to each cluster. For both experiments, the proportion of trials falling into the direct vs. the foil-skewed cluster was modelled across all conditions with a full set of fixed factors and a random effect of polarity by participant. For Experiment 2, model comparisons indicated significant main effects of polarity, $\chi^2(3) = 30.0$, $p < .001$, and number of possible targets, $\chi^2(4) = 88.7$, $p < .001$, and an interaction between them, $\chi^2(2) = 10.5$, $p = .005$. The same main effects were present in Experiment 3, $\chi^2(3) = 61.7$, $p < .001$, and $\chi^2(4) = 162.0$, $p < .001$,

and the same interaction was also observed, $\chi^2(2) = 41.3$, $p < .001$. Coefficients estimated for the full model in each experiment, indicating the simple effects of each factor, are given in Supplementary Tables 3 and 4. Overall, the model parameters showed that in both experiments, both negation and increasing the number of possible targets were associated with an increased proportion of trials exhibiting significant deviation towards the foil image, with a stronger effect of predictability in affirmative sentences. This pattern is also illustrated in Figure 9.

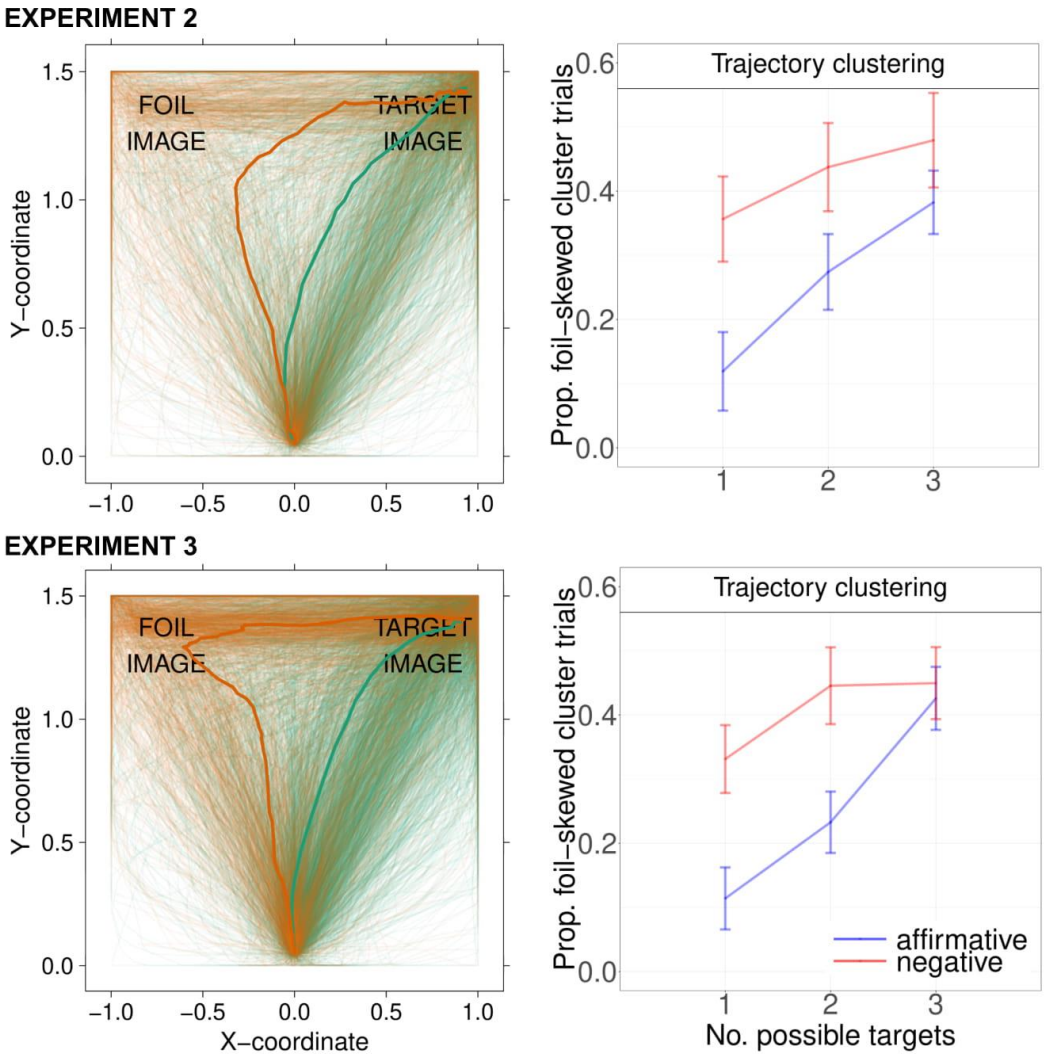


**Figure 9**. Trajectory analysis in Experiments 2 and 3. Left panels show a representation of the trajectories taken on all included trials, colour-coded according to cluster. Thick lines represent the cluster centres. Right panels show the mean proportion of trials allocated to the

cluster representing greater deviation from a straight line (stronger attraction to the foil stimulus), in each condition. Error bars represent standard errors.

As in Experiment 1, to rule out the interpretation that the foil-skewed trajectory cluster arose primarily as a result of participants making a rapid initial guess on certain trials, initiation times for trials falling into each cluster were compared. In fact, the mean initiation time was greater for trials falling into the foil-skewed cluster in both Experiments 2 and 3, and paired samples *t* tests showed that this difference was significant in both cases, Experiment 2: $t(23) = -6.69$, $p < .001$, Experiment 3: $t(31) = -6.09$, $p < .001$. Hence, foil-skewed trajectories could not be interpreted as having arisen disproportionately from trials with initial guessing.

Another concern in interpreting the data is the relative roles of the number of possible targets and memory load. In Experiment 1, we found that participants' memory for the objects in the grid grew poorer as more objects were added. Storing multiple objects for recall in later prediction or other processing represents a possible drain on cognitive resources, and the total number of objects in the grid was confounded with the number of possible targets on critical trials, meaning that the former factor could be responsible for some or all of the apparent main effects of the latter. To test this for Experiments 2 and 3, we examined the control trials, which were counterbalanced using the equivalent scheme as for critical trials, except that the possible targets were always drawn from the row with three objects. Figure 6 illustrates this: compare the critical and control conditions with reference to the images in each column. In the case of the rightmost column, there is no difference. Thus, all these control trials were effectively equivalent to critical trials with three possible targets, as the correct answer could be one of three options; however, they differed in the potential number of foils, and thereby in the total number of objects in the grid. Any degradation in performance with an increased number of objects in control trials can be attributed solely to

the increase in memory load. Thus, the difference in degradation as the number of objects increases between control trials (always three possible targets) and critical trials (varying possible target set size) can be taken as representing the effects of the number of possible targets (i.e., predictability) above and beyond memory load.

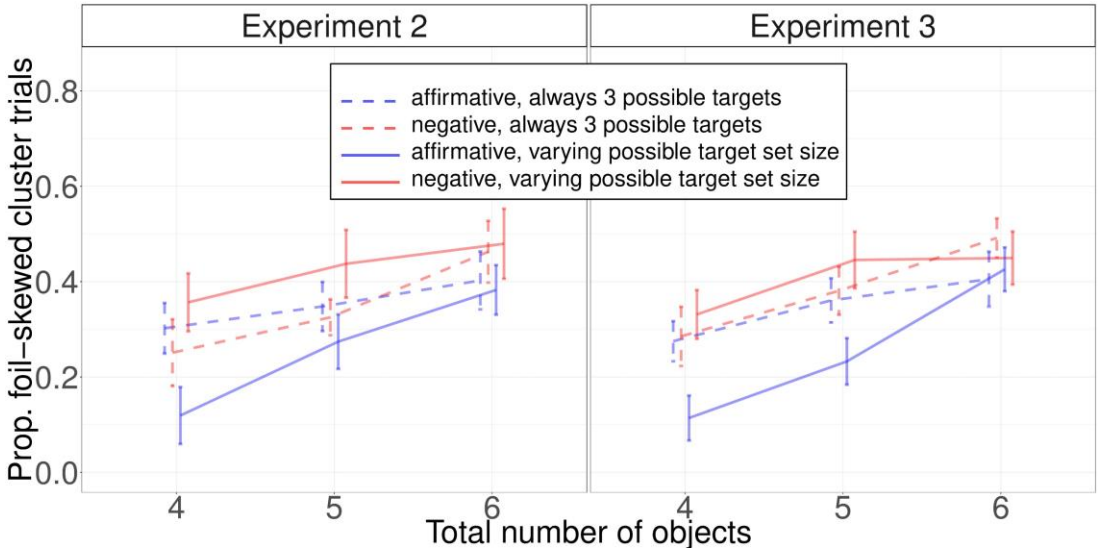


**Figure 10**. A comparison of the proportion of trajectories falling into the more foil-skewed cluster between critical trials (with varying possible target set size) and control trials (on which there were always three possible targets) in Experiment 2 (left) and Experiment 3 (right). Solid lines (representing varying possible target set size, i.e. critical trials) represent the same data as that presented in Figure 9 (right panels). The slopes of the dotted lines represent the effects of introducing additional objects, while keeping the number of possible targets at 3; the difference between these slopes and those of the solid lines represents the effect of changing the number of possible targets, over and beyond the increased memory load resulting from the presence of additional objects.

To make this comparison, we constructed mixed effects logistic regression models over the proportion of trials falling into each trajectory cluster in Experiments 2 and 3,

separately for affirmatives and negatives, including fixed factors of the number of possible targets (constant or varying), number of objects present in the display (4, 5, or 6), and the interaction between these two factors. A random effect of the number of possible targets by participant was also included in both cases. (Models with a more complex random effect structure failed to converge.) In Experiment 2, for affirmative trials, model comparisons indicated significant main effects of number of possible targets, $\chi^2(2) = 23.0$, $p < .001$, and number of objects, $\chi^2(2) = 75.7$, $p < .001$, and a significant interaction, $\chi^2(1) = 15.7$, $p < .001$. For negative trials, model comparisons indicated significant main effects of number of possible targets, $\chi^2(2) = 8.9$, $p = .009$, and number of objects, $\chi^2(2) = 58.1$, $p < .001$, but no significant interaction, $\chi^2(1) = 2.4$, $p = .13$. The pattern was very similar in Experiment 3: for affirmative trials, there were significant main effects of number of possible targets, $\chi^2(2) = 48.3$, $p < .001$, and number of objects, $\chi^2(2) = 163.3$, $p < .001$, and a significant interaction, $\chi^2(1) = 33.9$, $p < .001$; for negative trials, there was a significant main effect of number of objects, $\chi^2(2) = 64.0$, $p < .001$, but not of number of possible targets, $\chi^2(2) = 4.9$, $p = .088$, and a significant interaction, $\chi^2(1) = 4.3$, $p = .038$. Specifically, the proportion of foil-skewed trials increased at a similar rate to the increase in the number of objects in the display for affirmatives and negatives in the case of control trials (those with a fixed number of possible targets), whereas for critical trials, there was a much steeper increase for affirmatives than for negatives, as illustrated in Figure 10. This comparison supports the view that predictability specifically, not just memory load resulting from the increased number of objects, disproportionately affected processing of impacted affirmative sentences.

**Discussion**

By using a completion task in Experiments 2 and 3, we eliminated the confounding effects of manipulating truth value that were seen in Experiment 1 and tapped more directly

into participants' predictions about and responses to upcoming material. The results revealed a clear pattern, in which the allocation to trajectory clusters captured a pronounced distinction between direct and foil-skewed trajectories; in both experiments, the main effects of polarity and the number of possible true completions seen in Experiment 1 were replicated across all dependent variables, and a clear interaction arose between them in response accuracy and trajectory cluster allocation (Experiment 2) and subsequently (under pressure to respond more quickly, in Experiment 3) in all dependent variables. Interestingly, this interaction consistently fell in the opposite direction to the hypothesised direction, and to the weak interaction seen in Experiment 1. That is, the detrimental effect of increasing the number of possible targets on response accuracy and participants' ability to move directly to the target without attraction from the foil was stronger for affirmative than for negative sentences. This pattern is discussed in detail in the General Discussion.

The differences between Experiments 2 and 3 are likely to have arisen from the fact that participants were still rather slow to initiate and complete their responses in Experiment 2, meaning that switching to the completion task did not provide access to the early stages of processing to the extent hoped for. It has recently been demonstrated that delaying initial mouse movements might lead to smaller effects than encouraging rapid initiation of the mouse (Scherbaum & Kieslich, 2017). Experiment 3 applied more pressure to participants to initiate and complete their responses quickly, to ensure that as much information about early processing as possible was captured in their mouse trajectories. This enabled the clustering model to distinguish more clearly between direct and indirect trajectories; thus, although the number of participants was only 1.33 times the sample size of Experiment 2, the power of the design to detect trajectory-related effects was improved.

Before turning in the General Discussion to a more detailed interpretation of the overall set of results, it is worth discussing a potential concern that in Experiments 2 and 3,

certain negative conditions may be indistinguishable due to chance-level responding. That is, perhaps performance (as measured by trajectory cluster) was already so poor even on the easiest trials (those with only one possible target) for negatives that there was little room for it to fall further: in other words, initial responses were effectively guesses, which had to be remedied later in the trial. The lack of difference between the negative conditions with two and three possible targets in Experiment 3 supports this view. However, note that this interpretation is only reasonable if it is the case that no more than approximately 50% of trials fall into the foil-skewed cluster under any conditions. The proximity of the proportions in the aforementioned conditions to 50% supports this view, but there are two reasons to suppose that this pattern does not necessarily reflect a floor effect. First, it can be argued that a foil-skewed cluster proportion of 50% does not theoretically represent a floor in performance, as the hypothesis is that participants may experience active attraction towards the foil under some conditions. If this occurs, it would not be impossible for the proportion of foil-skewed trials to exceed 50%, as a result of a combination of trials representing initial guessing and those representing initial attraction to the foil for principled reasons; in fact, Experiment 1 provides evidence that the proportion of foil-skewed cluster trials may in reality exceed 50% under some conditions in a similar paradigm. However, note that this reasoning only applies to negative conditions, as there is no reason to expect active attraction to the foil to operate in any of the affirmative conditions. Second, as described in the Results sections for Experiments 2 and 3, trials falling into the foil-skewed cluster were, on average, initiated more slowly than those falling into the direct cluster. The opposite pattern would be expected if highly deviant trials arose from time-pressured participants initiating their mouse movements quickly with a random initial guess. In summary, if any foil-skewed trajectories must be driven by random initial guessing (as must be the case in the affirmative conditions, where there is no reason for participants to be actively attracted to the foil), there is a natural

limit for the proportion of direct cluster trials at 50%; but if there is an effect beyond this operating in the case of negative conditions, no such limit applies. Although the possibility cannot be ruled out absolutely that participants' initial trajectories reflected chance-level responding, these characteristics of the data support the interpretation of at least a subset of foil-skewed trials in the negative conditions as representing not simply a mistaken initial guess, but active attraction towards the foil response option, as might be observed in the case of a mistaken prediction. This interpretation, however, remains unlikely in the case of the condition with three possible targets, where the affirmative and negative conditions were very similar. Because it is unlikely that *less* guessing occurred in the negative than in the affirmative condition, and all affirmative foil-skewed trials must be attributable to guessing, there was probably not a meaningful fraction of trials exhibiting active attraction to the foil in this case.

## General Discussion

We investigated processing of negative sentences in three mouse-tracking experiments by manipulating the number of possible completions to sentences (and thus, their predictability) using preceding episodic contexts. Across all three experiments, consistent main effects of sentence polarity and the number of possible completions were observed, affecting every dependent variable. Both negative polarity and increasing the number of possible completions consistently reduced the accuracy and speed of participants' responses, and made them more likely to exhibit attraction to the foil response. However, in terms of the interaction between these two independent variables, the findings of Experiments 2 and 3 differed greatly from Experiment 1. In Experiment 2, and even more clearly in Experiment 3, the findings suggested that the number of possible targets had a stronger impact on processing for affirmatives than for negatives; furthermore, the difference between

affirmatives and negatives was larger in the case of trials with a single possible completion relative to those with two or three.

We begin the interpretation of these findings by considering the implications of a consistent and robust effect of polarity in all three experiments and across all measures, i.e., the fact that negative sentences were overall processed less accurately and more slowly than affirmative sentences, and yielded fewer direct mouse trajectories. This pattern obtained despite the fact that both sentence types were equally pragmatically licensed in the experimental context. This result has important implications, as it provides counter-evidence to the view that pragmatic licensing (Dale & Duran, 2011; Nieuwland & Kuperberg, 2008) or indeed contextual predictability (Nieuwland, 2016) is sufficient to observe symmetry of processing between affirmatives and negatives. Furthermore, using mouse trajectories, it is possible to explore the nature of the asymmetry more precisely, i.e., the difficulty arising during processing of negative sentences specifically stems from the interference from the incorrect response, as demonstrated by the comprehenders' initial movement often demonstrating attraction to the foil answer (such attraction also occurred with affirmatives, but with significantly lower frequency). In other words, initial processing of a negative sentence does not always account for the presence of negation. In principle, this result is compatible with the claim that negative sentences are not processed fully incrementally, as proposed by two-stage models of processing of negatives. According to this model, when the fragment *The bottom row doesn't contain* ... is available, the parser puts the interpretation of the negating element on hold in order to first process the embedded affirmative proposition (*The bottom row contains …*), only subsequently computing the meaning of its negation. The larger proportion of mouse trajectories exhibiting deviation towards the foil object in the negative condition reflects the first stage, at which the initial choice of completion (the foil) reflects the meaning of the embedded affirmative. However, we note that an alternative

explanation for these findings is available, as discussed in the section on local activation vs. global processing below.

Why did controlling for pragmatic felicity eliminate differences between negatives and affirmatives in Nieuwland and Kuperberg (2008) and Dale and Duran (2011) but not in our experiments? A possible explanation for the consistent main effect of polarity, and the apparent disparity between these and previous findings, might be located in the specific details of the pragmatics of the sentences presented. Dale and Duran (2011) observed that the main effect of negation only disappeared entirely in the case where they provided the richest versions of their contextual 'preambles' (their Experiment 3); these sentences were also the most similar to those presented by Nieuwland and Kuperberg (2008) in their licensed condition. In the case of the present experiment, although the visual contexts licensed all the sentences presented, it is possible that their episodic nature interfered with the mechanism facilitating incremental processing of negation in more felicitous contexts. There may be a meaningful distinction, in this case, between the long-term, stable associations in memory between everyday objects and their properties (for example), and the weak, temporary associations formed between locations and object names for a task such as this one. In this way, the sentences presented here may have been more comparable to those presented by Dale and Duran (2011) in their Experiment 2, or to the low cloze sentences in Nieuwland (2016): pragmatically licensed, but only weakly contextually enriched. Note, however, that even if this explanation is correct, it remains the case that pragmatic licensing alone is insufficient to account for any relative difficulties in processing of negative vs. affirmative sentences. Although our results are entirely compatible with the existence of important effects of pragmatics, they suggest that predictability also interacts (in a very distinct way) with negation.

With respect to the interaction between polarity and number of possible completions, where they contradict one another (Experiment 1 indicated that increasing the number of possible completions had a more detrimental impact on negatives, whereas Experiment 2 and 3 indicated that doing so had a more detrimental impact on affirmatives), we consider the design and combined findings of Experiments 2 and 3 to be more reliable than those of Experiment 1. As discussed above, the truth value judgement paradigm employed in Experiment 1 interacted in undesirable ways with the independent variables of interest, meaning that the results were difficult to interpret. Although they fell partly in line with our initial hypotheses, this pattern is not necessarily explained entirely by the interaction between polarity and number of possible true completions, but rather is confounded by the sentence truth value; furthermore, the data and clustering approach captured only relatively impoverished information about participants' early cognition during the task. Therefore, we take the more surprising pattern of results observed in Experiments 2 and 3 to be more indicative of the effects of predictability on participants' incremental processing of negatives.

In Experiments 2 and 3, the interaction between polarity and the number of possible targets reflected a larger detrimental effect of the presence two or three possible targets (relative to the presence of only one) on affirmative sentences, compared to negative sentences. This is a reversal of the initial hypothesis that negative sentences would pose more problems relative to affirmatives as the number of possible targets increased. Recall that this initial hypothesis was based on the demonstration that the introduction of context rendering negatives pragmatically felicitous makes them no more difficult to process than affirmatives (Dale & Duran, 2011; Nieuwland & Kuperberg, 2008) and on the assumption that this effect arose from the relationship between pragmatic felicity and predictability. Specifically, in contexts where experimental manipulations rely on participants' prior world knowledge, it is almost inevitably the case that pragmatic felicity and predictability go hand-in-hand: negative

sentences that have low predictability seem out-of-the-blue and odd (*A robin is not a tree*), whereas the contextual enrichment that makes others sound more appropriate also makes them more predictable (*With proper equipment, scuba diving isn't very dangerous*). The use of episodic contexts, on the other hand, provides a way to dissociate these dimensions. In particular, in our study, affirmative and negative descriptions were equally expected and natural (participants knew that they would see grids of objects which would be described using affirmative or negative sentences); our manipulation of predictability, via the number of possible completions, was orthogonal and did not interfere with the pragmatic felicity of the sentences. We hypothesised that increasing the number of possible completions (i.e., decreasing predictability) would be more detrimental for processing of negative sentences. Yet this hypothesis was not borne out in Experiments 2 and 3: negative sentences were more difficult than affirmatives, despite their equal felicity, even in the case of trials with a single possible target, and this difference between conditions in fact decreased when the number of possible targets was increased, representing a disproportionately large effect on affirmatives, rather than negatives.

How can this interaction pattern be accounted for? We believe that the interaction reflects the special status of affirmative trials with certainty of a unique completion. Indeed, it is reasonable to imagine that prediction-making is an easy and appealing strategy for this type of sentence. Use of such a strategy makes it possible to complete the sentence and the task easily (as reflected by rapid responses and direct mouse trajectories); this ability deteriorates sharply, even in the affirmative conditions, as the number of possible targets increases. Because there is no underlying reason other than a simple mistake or wrong initial guess for the participant to experience any attraction towards the foil in any of the affirmative conditions (unlike in the case of negative conditions, where such attraction could be actively and specifically driven by *systematically* erroneous predictions), the proportion of trials

falling into the foil-skewed cluster in the affirmative conditions reflects a minimum amount of initial wrong guesses, subsequently corrected prior to completing the response, that must be made at that level of difficulty (e.g., around 10% in the easiest case). Thus, the difference between the affirmative and negative conditions for trials with each number of possible targets reflects the addition of foil-skewed responses arising from a combination of both 1) any increase in initial guessing or mistakes attributable to general processing difficulties, which happen in this case to be imposed by negation, and 2) active attraction to the foil caused by some effect in which an initial prediction does not take the negation into account (e.g. two-stage processing or a related alternative). Although it is difficult to disentangle these factors, it is clear that the effect of both of them combined can no longer be detected in the case of trials with three possible targets, because the affirmative case shows that with a six-object grid, the task has become difficult enough to force participants to make an initial guess on the majority of trials (meaning that the proportion of initial motions towards the foil approaches 50%). Thus, although effects specific to negation may still be in operation, they are difficult to detect given that prediction is already very difficult. Conceivably, participants make trial-by-trial decisions (whether consciously or unconsciously), based on the global complexity or resource consumption of the overall trial conditions, as to whether to attempt to make a prediction or to incorporate all the available information into their intermediate representation of the input at all.

Processing in the specific context of the experimental task may have differed in some important ways from ordinary processing of language in a non-laboratory context. This is true of all laboratory experiments, but arguably it is particularly true in the present case because of the especially repetitive nature of the sentences (not in their exact content, but in their construction and content more generally) and the fact that the visual contexts were similarly repetitive and constrained. However, we took measures to ensure that participants

could not learn any special strategies for completing the task (for example, the inclusion of control sentences meant that they could not assume that the sentence would be about the 'less full' part of the grid). Conceivably, participants adopted a strategy for processing the repetitive linguistic input that differed from the strategy they would use in a more naturalistic setting: for example, perhaps they were more motivated than usual to activate potential completions to the sentence. It is therefore possible that our experiment overestimates the extent to which comprehenders engage in incremental processing; however, we would expect any such overestimation to occur equally in the affirmative and negative conditions. Therefore, if participants were impaired in their incremental processing even in the easiest negative condition (i.e., when there is a single possible target) relative to affirmative conditions, this suggests that similarly predictable sentences in a more naturalistic context might, if anything, give rise to an even larger difference between affirmatives and negatives (since it is clear from previous work that incremental processing does occur for affirmatives).

The comparison between critical and control trials presented for Experiments 2 and 3 shows that the effects of possible target set size can be attributed to the increased complexity of the visual display (indicating effects of memory load rather than effects specific to linguistic processing), but only partially. The remainder of the effect must reflect effects of manipulating the target set size that go above and beyond the increase in visual complexity or working memory demands. Furthermore, this latter component of the effect differed between affirmative and negative sentences, meaning that linguistic processing was also specifically modulated by the target set size.

It is also worth considering how the experimental paradigm and task used here fit into the wider methodological picture. As reviewed in the introduction, two other methodologies have been among those most commonly used to explore processing of negative sentences. ERP studies of negation have focused on the N400 component, and have typically presented

more varied sentences reliant on long-term 'world knowledge' (e.g., Fischler et al., 1983; Kounios & Holcomb, 1992; Nieuwland & Kuperberg, 2008). Eye-tracking studies of negation (e.g., Orenes et al., 2014; Orenes, Moxey, Scheepers, & Santamaría, 2016) make use of the visual world paradigm; this approach requires visual information to be presented concurrently with linguistic input, and hence does not directly involve memory. Our study sits in between these methodologies, as the tasks we have employed rely on the application of newly-acquired, contextually-relevant knowledge stored in short-term memory. The need to tap into such information is a common occurrence in natural settings (e.g., interlocutors frequently discuss an event that they have just observed). Furthermore, the use of mouse-tracking in the present study lends itself to the capture of relatively early stages of cognition (via participants' mouse trajectories that occur prior to the eventual response), and in the case of the completion task, this paradigm avoids the reliance on presenting both true and false sentences that is inherent in the use of the N400 as a dependent measure.

Whereas the paradigm employed here is well suited to exploring how comprehenders draw on associations arising from episodically-presented information, it did considerably limit the range of sentences used (similarly to eye-tracking studies, and unlike those studies reliant on 'world knowledge'). Thus, it is not entirely clear to what extent our findings would generalise to the processing of negation in other contexts or environments. Despite this, we would argue that we have used a context that strongly favours the use of incremental processing and prediction-making as a strategy for completing the experimental task. Therefore, in conditions in which participants fail to do this successfully, it is unlikely that they would do so when encountering sentences with similar characteristics in the course of naturalistic language processing. Specifically, given the lack of equivalence between affirmatives and negatives even when the number of possible targets is small and prediction-

making thus relatively easy, the results strongly suggest that there is an asymmetry in incremental processing between affirmative and negative sentences more generally.

**Local activation vs. global processing**

The most robust finding across all three experiments was that negative sentences required more processing effort than affirmative sentences. Most significantly, in Experiments 2 and 3, there was more often an initial attraction towards the foil answer in the case of negatives. As discussed above, this result could be taken as evidence that negation is initially not incorporated into the sentence representation, as held by two-stage processing models of negation. However, we propose an alternative explanation that highlights an inherent conflict between local and global sentence processing mechanisms that arises during real-time processing of negative sentences.

Consider the negative fragment *The top row doesn't contain…* as heard by the participant in the context of the grid shown in Figure 6, which should be correctly completed with *basket*, the alternative *lamp* being an incorrect completion. As the sentence unfolds in real time, the listener first hears *The top row*, which activates the object(s) located in that row, i.e. *lamp*, *garlic* and *bow* through processes involved in simple lexical access (e.g. Swinney, 1979). In order for the negative sentence to be processed fully incrementally, this set of objects, having been locally activated by the initial fragment, should be supressed as soon as the negated verb (*doesn't contain*) appears. The participant's attention should shift away from the top row, i.e. to the bottom row and the elements therein. Hence, there is a mismatch between local and global representations (to use terminology employed by Tabor, Galantucci, & Richardson, 2004, in discussing the incremental construction of syntactic structure), because the early local activation that occurs on the basis of the partial input (*lamp* and other objects in the top row) needs to be subsequently supressed once the final sentence interpretation is derived. For comparison, such a mismatch does not occur in the affirmative

condition *The bottom row contains…*: the initial noun phrase *the bottom row* locally activates *basket*, which is also the globally correct output.

There are various situations in which this type of mismatch occurs and must be resolved. For example, the concept of *waiter* is lexically activated by local elements of both the fragments *The customer was served…* and *The customer served*…, although it forms part of the likely continuation in only one of these cases; however, the parser is able to update on the more global structural roles of portions of the input in generating predictions (Chow, Smith, Lau, & Phillips, 2016). This mismatch-resolving process is an inherent feature of many types of negative sentences. For example, compare affirmative and negative sentences in Fischler et al (1983). In the affirmative sentence *A robin is a bird*, the local activation of the category *bird* on the basis of the initial phrase *a robin* is consistent with the sentence globally and makes the final word *bird* expected. Its negative counterpart *A robin is not a tree*, on the other hand, creates a local-global mismatch between activation of the concept *bird*, primed by *robin* early on, and the later need to suppress it once the negation appears.

Although this view differs from a two-stage model of negation processing in that it does not specifically require considering a full affirmative proposition before applying negation, it makes equivalent predictions in some cases. Under both views, negation is an operation that requires considering a representation other than (and in the easiest cases, complementary to) the one that is initially under consideration (for the two stage model: the set opposite to the one denoted by the affirmative proposition; here: supressing what has been initially activated). This interpretation is also in line with Garrod and Sanford's (1999) presentation of the notion that information incrementally incorporated into a comprehender's structural representation of a sentence does not necessarily induce a full update to the interpretation of the input at every stage.

In the examples above, the local-global mismatch is unavoidable even if negation is processed fully incrementally, simply for the reason that in the linear input, negation follows other relevant linguistic material, i.e. the noun phrase *The top row* or *The robin*. This type of mismatch should be expected to be considerably attenuated if negation appears early in the linear input. For illustration, consider the ill-formed English sentence fragment *Doesn't contain the top row …* If this word order were possible in English, then the presence of negation early in the left-to-right input, if processed fully incrementally, could significantly reduce — although perhaps not completely eliminate — low-level priming from *the top row* by immediately shifting the focus of attention to the complement location. Although such word order is illicit in English, it is licit in other languages (e.g., Russian).

**Conclusion**

Overall, this set of results suggests that predictability as operationalised here (using the number of possible sentence completions based on episodic contextual associations) cannot explain the variation in how incremental processing of negative sentences differs from that of affirmative sentences under certain sets of circumstances. The strong main effects of polarity as well as the number of possible completions observed throughout indicate that negation might impose specific processing difficulties in a broader set of contexts than previously thought, including in some cases when fully licensed by the episodic context. In particular, the main effects of the number of possible sentence completions observed here constitute evidence that episodic associations may be less conducive to the rapid and incremental incorporation of information and associated prediction-making that is made possible by a rich pragmatic context (perhaps specifically relying on long-term semantic associations or world knowledge). While the latter may allow felicitous negative sentences to be processed as readily as equivalent affirmatives, the type of licensing based on episodic context that is used

here makes incremental processing and prediction generally more difficult, but also makes it disproportionately difficult for the comprehender to over-ride unhelpful associations arising from the local content of the negated material.

**Acknowledgements**

Hugo Hammond, Jamie Mcevoy, Shanaz Pottinger and Alesi Rowland assisted with data collection. Thanks also to Michele Gubian for useful input on the design and to Evgenii Kalenkovich for discussion of the data analyses.

## References


Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33,* 583–609. doi: 1.1111/j.1551-6709.2009.01022.x

Barr, J. B., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. doi: 10.18647/jss.v067.i01

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review, 82*, 45–73. doi: 10.1037/h0076248

Chow, W.-Y., Smith, C. Lau, E. & Phillips, C. (2016). A 'bag-of-arguments' mechanism for initial verb predictions. *Language, Cognition and Neuroscience, 31,* 187–276.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language*. New York: Harcourt Brace Jovanovich by Stanford University Libraries.

Crocker, M. W. (1994). On the nature of the principle-based sentence processor. In C. Clifton, Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 245–266). Hillsdale, NJ: Erlbaum.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science, 35*, 983–996. doi: 10.1111/j.1551-6709.2010.01164.x

Fischer, M. H., & Hartmann, M. (2014). Pushing forward in embodied cognition: May we mouse the mathematical mind? *Frontiers in Psychology: Cognition, 5*, 1315. doi: 10.3389/fpsyg.2014.01315

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W., Jr. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology, 20*, 400–409.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science, 20*, 1183–1188. doi: 10.1111/j.1467-9280.2009.02422.x

Freeman, J. B., & Ambady, N. (2010). Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods, 1*, 226–241. doi: 10.37358/BRM.42.1.226

Garrod, S., & Sanford, A. (1999). Incrementality in discourse understanding. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 3–27). Mahwah, NJ: Lawrence Erlbaum.

Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C. (1999). Not propositions. *Cognitive Systems Research, 1*, 19–33.

Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior, 4*(2), 107–111.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistics Society, Series C (Applied Statistics), 28*(1), 100–108. doi: 10.2307/2346830

Horn, L. R. (1989). *A natural history of negation.* Chicago: The University of Chicago Press.

Horn, L. R., & Wansing, H. (2017). Negation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 edition). Retrieved from https://plato.stanford.edu/archives/spr2017/entries/negation

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*, 151–171. doi: 10.1016/j.actpsy.2010.11.003

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10,* 244–253. doi: 10.1016/S0022-5371(71)80051-8

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics, 38,* 1033–1050. doi: 10.1016/j.pragma.2005.09.012

Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29,* 439–446. doi:10.1037/0278-7393.29.3.439

Kazanina, N. (2017). Predicting complex syntactic structure in real time: Processing of negative sentences in Russian. *Quarterly Journal of Experimental Psychology, 70,* 2200–2218. doi: 10.1080/17470218.2016.1228684

Kent, C., Taylor, S., Taylor, N., and Darley, E. (2017). Tracking trajectories: A brief review for researchers. *The Cognitive Psychology Bulletin, 2*(1), 12–16.

Kimball, J. (1975). Predictive analysis and over-the-top parsing. In J. Kimball (Ed.), *Syntax and semantics, Vol. 4* (pp. 155–179). New York: Academic Press.

Kintsch, W. (1974). *The representation of meaning in memory*. Oxford: Lawrence Erlbaum.

Kounios, J., & Holcomb, P. J. (1992). Structure and process in semantic memory: Evidence from event-related brain potentials and reaction times. *Journal of Experimental Psychology: General, 121,* 459–479. doi: 10.1037/0096-3445.121.4.459

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of* Psychology, 62, 14.1–14.27. doi: 10.1146/annurev.psych.093008.131123

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. doi: 10.18637/jss.v082.i13

Löbner, S. (2000). Polarity in natural language: Prediction, quantification and negation in particular and characterizing sentences. *Linguistics and Philosophy, 23*(3), 213–308.

Lüdtke, J., Friedrich, C. K., de Filippis, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence-picture verification paradigm. *Journal of Cognitive Neuroscience, 20*, 1355–1370. doi: 10.1162/jocn.2008.20093

Mayo, R., Schul, Y., & Burnstein, E. (2004). 'I am not guilty' vs 'I am innocent': Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology, 40*, 433–449. doi: 10.1016/j.jesp.2003.07.008

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review, 95*, 183–237. doi: 10.1037/0033-295X.95.2.183

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Leraning, Memory, and Cognition, 42*, 316–334. doi: 10.1037/xlm0000173

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science, 19*, 1213–1218. doi: 10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife, 7*, e33468. doi: 10.7554/eLife.33468

Nordmeyer, A. E., & Frank, M. C. (2018). Early understanding of pragmatic principles in children's judgments of negative sentences. *Language Learning and Development, 14*, 262–278. doi: 10.1080/15475441.2018.1463850

Oesterle, J. (1962). *Aristotle: On interpretation.* Commentary by St. Thomas and Cajetan. Milwaukee: Marquette University Press.

Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language, 74*, 36–45. doi: 10.1016/j.jml.2014.04.001

Orenes, I., Moxey, L., Scheepers, C., & Santamaría, C. (2016). Negation in context: Evidence from the visual world paradigm. *The Quarterly Journal of Experimental Psychology, 69*, 1082–1092. doi: 10.1080/17470218.2015.1063675

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*, 105–110. doi: 10.1016/j.tics.2006.12.002

Pritchett, B. L. (1991). Head position and parsing ambiguity. *Journal of Psycholinguistic Research, 20*, 251–270. doi: 10.1007/BF01067218

Reuter, T., Feiman, R., & Snedeker, J. (2018). Getting to *no*: Pragmatic and semantic factors in two- and three-year-olds' understanding of negation. *Child Development, 89*, e364–381. doi: 10.1111/cdev.12858

Scherbaum, S., & Kieslich, P. J. (2017). Stuck at the starting line: How the starting procedure influences mouse-tracking data. *Behavior Research Methods, 50*, 2097–2110. doi: 10.3758/s13428-017-0977-4

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences, 13*, 360–366. doi: 10.1016/j.tics.2009.04.009

Spivey, M. J., & Dale, R. (2004). On the continuity of mind: Toward a dynamical account of cognition. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 43), 87–142.

Sturt, P., & Lombardo, V. (2005). Processing coordinated structures: Incrementality and connectedness. *Cognitive Science, 29*, 291–305. doi: 10.1207/s15516709cog0000_8

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645–659. doi: 10.1016/S0022-5371(79)90355-4

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50*, 355–370. doi: 10.1016/j.jml.2004.01.001

Tian, Y., & Breheny, R. (2015). Dynamic pragmatic view of negation processing. In P. Larrivée & C. Lee (Eds.), *Negation and polarity: Experimental perspectives* (pp. 21–43). Switzerland: Springer.

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology, 63*, 2305–2312. doi: 10.1080/17470218.2010.525712

Tomlinson, J., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language, 69*, 18–35. doi: 10.1016/j.jml.2013.02.003

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology, 11*, 92–107.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology, 52*, 133–142.

Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior, 4*, 7–11.

Wason, P. C., & Jones, S. (1963). Negatives: Denotation and connotation. *British Journal of Psychology, 54*, 299–307.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* New York: Springer-Verlag.

## Supplementary Materials

### Further Design and Counterbalancing Details

*Experiment 1*

No participant was presented with more than one trial from any set of 72 associated with a particular set of visual stimuli. In total, each participant completed 216 of the 5,184 trials constructed; half of these were distributed evenly across the 12 critical conditions, and the other half were counterbalancing trials, in which the row of the grid containing a fixed number of objects was referred to; the latter were also distributed evenly across the equivalent non-critical conditions, and were included so that participants would not be able to predict which row would be referred to immediately upon seeing a grid containing four or five objects.

Within each of the 18 subsets in each set of possible trials, the critical object exchanged roles with one of the objects in the other row in half the versions; and within in each of these three subsets, the critical object appeared in the left, centre, or middle column of the grid. The variable number of objects appeared on the top row in half the image sets and on the bottom row in the other half. These further counterbalancing measures ensured that effects specific to particular visual items, their names, or their locations in the grid applied equally across all conditions.

*Experiments 2 and 3*

In both sentence completion experiments, the target object appeared equally frequently in the left, middle, and centre columns on critical trials, as did the foil. There were two versions of each grid, in which the identities of the target and foil objects were exchanged. For example, in the leftmost grids presented in Figure 5, the lamp and the medal exchanged roles as the target and foil. The horizontal location of the target was randomised.

In the same way as for Experiment 1, no participant completed more than one trial from the set associated with a particular combination of images.

**Analyses Without Data Trimming**

Prior to carrying out analyses, the data were trimmed as described in the main text in order to discard trials on which participants were anomalously slow to initiate or complete their responses. To check that our findings were generally robust to the exact thresholds (for initiation time and response time) used for this trimming, we repeated the analyses for each dependent variable in the same way as reported in the main text, but with no trimming (i.e., with all correct trials included). The results of the model comparisons carried out is these analyses are summarised in Supplementary Tables 7, 8 and 9.

## Supplementary Table 1

Simple effects of each independent variable on response accuracy and proportion of trajectories falling into each cluster (Experiment 1).

| Variable | | Response accuracy (proportion correct) | | | | | Proportion of direct trajectories | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ß | *z* | *p* | 95% CI | | ß | *z* | *p* | 95% CI | |
| | | | | | Lower | Upper | | | | Lower | Upper |
| **Falsity** | | | | | | | | | | | |
| Aff. | | | | | | | | | | | |
| | 1 compl. | –0.99* | –3.05 | .002 | –1.62 | –0.35 | –0.35* | –2.01 | .044 | –0.69 | –0.01 |
| | 2 compl. | –0.22 | 0.82 | .412 | –0.75 | 0.31 | –0.35* | –2.17 | .030 | –0.67 | –0.03 |
| | 3 compl. | –0.69* | –3.05 | .002 | –1.13 | –0.25 | –0.84* | –5.01 | < .001 | –1.17 | –0.51 |
| Neg. | | | | | | | | | | | |
| | 1 compl. | –0.22 | –0.95 | .343 | –0.69 | 0.24 | 0.26 | 1.73 | .084 | –0.04 | 0.56 |
| | 2 compl. | 0.00 | 0.00 | .999 | –0.40 | 0.40 | –0.56* | –3.45 | .001 | –0.88 | –0.24 |
| | 3 compl. | –0.67* | –3.87 | < .001 | –1.00 | –0.33 | –0.53* | –3.01 | .003 | –0.87 | –0.18 |
| **Negative Polarity** | | | | | | | | | | | |
| True | | | | | | | | | | | |
| | 1 compl. | –1.29* | –3.95 | < .001 | –1.93 | –0.65 | –1.33* | –4.22 | < .001 | –1.95 | –0.71 |
| | 2 compl. | –0.87* | –3.33 | .001 | –1.38 | –0.36 | –1.35* | –4.31 | < .001 | –1.96 | –0.73 |
| | 3 compl. | –0.89* | –3.75 | < .001 | –1.35 | –0.42 | –1.69* | –5.35 | < .001 | –2.31 | –1.07 |
| False | | | | | | | | | | | |
| | 1 compl. | –0.08 | –0.29 | .771 | –0.59 | 0.44 | –1.24* | –3.97 | < .001 | –1.85 | –0.63 |
| | 2 compl. | –0.65* | –2.61 | .009 | –1.13 | –0.16 | –1.55* | –4.95 | < .001 | –2.17 | –0.94 |
| | 3 compl. | –0.86* | –4.44 | < .001 | –1.24 | –0.48 | –1.38* | –4.32 | < .001 | –2.01 | –0.76 |
| **2 Possible Complet.** | | | | | | | | | | | |
| True | | | | | | | | | | | |
| | Aff. | –0.70* | –2.08 | .038 | –1.36 | –0.04 | –0.41* | –2.39 | .017 | –0.75 | –0.07 |
| | Neg. | –0.28 | –1.30 | .194 | –0.70 | 0.14 | –0.43* | –2.80 | .005 | –0.73 | –0.13 |
| False | | | | | | | | | | | |
| | Aff. | 0.07 | 0.26 | .795 | –0.43 | 0.57 | –0.42* | –2.52 | .012 | –0.74 | –0.09 |
| | Neg. | –0.50* | –2.23 | .026 | –0.95 | –0.06 | –0.73* | –4.57 | < .001 | –1.04 | –0.42 |
| **3 Possible Complet.** | | | | | | | | | | | |
| True | | | | | | | | | | | |
| | Aff. | –0.29 | –1.08 | .282 | –0.81 | 0.24 | 0.15 | 0.90 | .367 | –0.18 | 0.49 |
| | Neg. | –0.31 | –1.57 | .117 | –0.69 | 0.08 | –0.19 | –1.20 | .230 | –0.50 | 0.12 |
| False | | | | | | | | | | | |
| | Aff. | –0.75* | –3.29 | .001 | –1.20 | –0.30 | –0.33* | –2.08 | .037 | –0.64 | –0.02 |
| | Neg. | –0.97* | –5.32 | < .001 | –1.33 | –0.61 | –0.16 | –0.90 | .371 | –0.51 | –0.19 |

*Note.* Reference values for each factor are True (for effects of falsity), Affirmative (for effects of negative polarity), Single Possible True Completion (for effects of the presence of two possible true completions) and Two Possible True Completions (for effects of the presence of three possible true completions). For example, the first row of the table provides the simple effect of falsity, relative to truth, for affirmative sentences with a single possible true completion. The statistically significant beta value of –0.99 indicates that in the affirmative condition, when there was a single possible true completion, the odds of answering correctly were 63% lower (= $[e^{-0.99} – 1] \times 100$, since the coefficient represents the difference in log odds) for false sentences compared to true sentences. *$p < .05$

## Supplementary Table 2

Simple effects of each independent variable on response time (Experiment 1).

| Variable | | Response time (ms) | | | | | |
|---|---|---|---|---|---|---|---|
| | | ß | *t* | d.f. | *p* | 95% CI | |
| | | | | | | Lower | Upper |
| **Falsity** | | | | | | | |
| Aff. | | | | | | | |
| | 1 compl. | 219.6* | 5.98 | 4591.6 | < .001 | 146.6 | 291.6 |
| | 2 compl. | 180.7* | 4.87 | 4574.0 | < .001 | 108.0 | 253.3 |
| | 3 compl. | 184.9* | 4.87 | 4373.8 | < .001 | 110.5 | 259.4 |
| Neg. | | | | | | | |
| | 1 compl. | 15.2 | 0.41 | 4486.0 | .684 | –58.3 | 88.7 |
| | 2 compl. | 51.5 | 1.34 | 4476.1 | .180 | –99.2 | 114.2 |
| | 3 compl. | 46.8 | 1.15 | 4474.6 | .250 | –32.9 | 126.5 |
| **Negative Polarity** | | | | | | | |
| True | | | | | | | |
| | 1 compl. | 582.5* | 13.49 | 66.5 | < .001 | 497.9 | 667.1 |
| | 2 compl. | 589.8* | 13.78 | 166.2 | < .001 | 505.9 | 673.7 |
| | 3 compl. | 634.1* | 14.63 | 174.5 | < .001 | 549.2 | 719.1 |
| False | | | | | | | |
| | 1 compl. | 378.2* | 8.89 | 162.0 | < .001 | 294.8 | 461.6 |
| | 2 compl. | 460.6* | 10.73 | 168.1 | < .001 | 476.4 | 544.8 |
| | 3 compl. | 495.9* | 10.96 | 205.4 | < .001 | 407.3 | 584.6 |
| **2 Possible Complet.** | | | | | | | |
| True | | | | | | | |
| | Aff. | 126.0* | 3.36 | 83.6 | .001 | 52.5 | 199.5 |
| | Neg. | 133.5* | 3.29 | 85.5 | .001 | 54.1 | 213.0 |
| False | | | | | | | |
| | Aff. | 86.9* | 2.33 | 4372.9 | .020 | 13.9 | 160.0 |
| | Neg. | 169.3* | 4.47 | 4474.8 | < .001 | 95.0 | 243.6 |
| **3 Possible Complet.** | | | | | | | |
| True | | | | | | | |
| | Aff. | 50.6 | 1.36 | 4574.0 | .173 | –22.1 | 123.4 |
| | Neg. | 95.0* | 2.45 | 4574.7 | .014 | 18.9 | 171.2 |
| False | | | | | | | |
| | Aff. | 54.9 | 1.45 | 4775.1 | .148 | –19.5 | 129.3 |
| | Neg. | 90.3* | 2.24 | 4476.1 | .025 | 11.4 | 169.2 |

*Note.* Entries are interpreted in the same way as for Supplementary Table 1, with the same reference values, although coefficients represent differences in milliseconds, rather than changes in log odds. *$p < .05$

## Supplementary Table 3

Simple effects of each independent variable on response accuracy and proportion of trajectories falling into each cluster (Experiment 2).

| Variable | Response accuracy (proportion correct) | | | | | Proportion of direct trajectories | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ß | *z* | *p* | 95% CI | | ß | *z* | *p* | 95% CI | |
| | | | | Lower | Upper | | | | Lower | Upper |
| **Negative Polarity** | | | | | | | | | | |
| 1 poss. targ. | –0.68* | –3.15 | .002 | –1.11 | –0.26 | –1.37* | –6.02 | < .001 | –1.81 | –0.92 |
| 2 poss. targ. | –0.47* | –2.59 | .01 | –0.82 | –0.11 | –0.79* | –3.90 | < .001 | –1.19 | –0.39 |
| 3 poss. targ. | –0.16 | –0.89 | .38 | –0.50 | 0.19 | –0.50* | –2.55 | .01 | –0.89 | –0.12 |
| **2 Possible Targets** | | | | | | | | | | |
| Aff. | –0.74* | –3.43 | .001 | –1.16 | –0.32 | –1.07* | –5.22 | .001 | –1.47 | –0.67 |
| Neg. | –0.52* | –2.87 | .004 | –0.88 | –0.17 | –0.50* | –2.80 | .005 | –0.85 | –0.15 |
| **3 Possible Targets** | | | | | | | | | | |
| Aff. | –0.22 | –1.21 | .230 | –0.58 | –0.14 | –0.53* | –3.02 | .003 | –0.87 | –0.19 |
| Neg. | 0.09 | 0.52 | .605 | –0.25 | 0.42 | –0.24 | –1.34 | .180 | –0.59 | 0.11 |

*Note.* Entries are interpreted in the same way as for Table 1, with the same reference values for polarity. Reference values for the number of possible targets are the same as for the number of possible true completions in Experiment 1 (i.e., one possible target is the reference for two possible targets, and two possible targets is the reference for three). *$p < .05$

## Supplementary Table 4

Simple effects of each independent variable on response accuracy and proportion of trajectories falling into each cluster (Experiment 3).

| Variable | Response accuracy (proportion correct) | | | | | Proportion of direct trajectories | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ß | *z* | *p* | 95% CI | | ß | *z* | *p* | 95% CI | |
| | | | | Lower | Upper | | | | Lower | Upper |
| **Negative Polarity** | | | | | | | | | | |
| 1 poss. targ. | –1.00* | –4.79 | < .001 | –1.41 | –0.59 | –1.39* | –6.88 | < .001 | –1.79 | –1.00 |
| 2 poss. targ. | –0.82* | –4.65 | .001 | –1.17 | –0.47 | –0.99* | –5.53 | < .001 | –1.35 | –0.64 |
| 3 poss. targ. | –0.19 | –1.24 | .216 | –0.50 | 0.11 | –0.06 | –0.35 | .725 | –0.40 | 0.28 |
| **2 Possible Targets** | | | | | | | | | | |
| Aff. | –0.65* | –3.21 | .001 | –1.04 | –0.25 | –0.90* | –5.16 | .001 | –1.25 | –0.56 |
| Neg. | –0.46* | –3.11 | .002 | –0.76 | –0.17 | –0.50* | –3.53 | < .001 | –0.78 | –0.22 |
| **3 Possible Targets** | | | | | | | | | | |
| Aff. | –0.74* | –4.64 | < .001 | –1.06 | –0.43 | –0.98* | –6.76 | < .001 | –1.27 | –0.70 |
| Neg. | –0.12 | –0.84 | .402 | –0.39 | 0.16 | –0.05 | –0.35 | .728 | –0.32 | 0.23 |

*Note.* Entries are interpreted in the same way as for Supplementary Table 2, with the same reference values as those used in Supplementary Table 3. *$p < .05$

## Supplementary Table 5

Simple effects of each independent variable on response time (Experiment 2).

| Variable | Response time (ms) | | | | | |
|---|---|---|---|---|---|---|
| | ß | *t* | d.f. | *p* | 95% CI | |
| | | | | | Lower | Upper |
| **Negative Polarity** | | | | | | |
| 1 poss. targ. | 359.0* | 7.38 | 74.3 | < .001 | 263.7 | 454.3 |
| 2 poss. targ. | 243.5* | 4.84 | 84.2 | < .001 | 144.9 | 342.0 |
| 3 poss. targ. | 253.5* | 5.02 | 85.7 | < .001 | 154.3 | 351.8 |
| **2 Possible Targets** | | | | | | |
| Aff. | 276.1* | 5.86 | 60.2 | < .001 | 183.8 | 368.3 |
| Neg. | 160.5* | 3.30 | 67.4 | .002 | 65.1 | 256.0 |
| **3 Possible Targets** | | | | | | |
| Aff. | 92.4 | 1.99 | 62.4 | .051 | 1.3 | 183.4 |
| Neg. | 101.9* | 2.14 | 67.7 | .036 | 8.4 | 195.4 |

*Note.* Entries are interpreted in the same way as for Supplementary Table 3, with the same reference values for polarity and number of possible targets. *$p < .05$

## Supplementary Table 6

Simple effects of each independent variable on response time (Experiment 3).

| Variable | Response time (ms) | | | | | |
|---|---|---|---|---|---|---|
| | ß | *t* | d.f. | *p* | 95% CI | |
| | | | | | Lower | Upper |
| **Negative Polarity** | | | | | | |
| High pred. | 213.4* | 10.04 | 31.8 | < .001 | 171.7 | 255.0 |
| Med pred. | 177.0* | 7.84 | 32.2 | < .001 | 132.8 | 221.3 |
| Low pred. | 57.8* | 2.14 | 28.1 | .04 | 4.9 | 110.7 |
| **Med. Predictability** | | | | | | |
| Aff. | 143.5* | 7.00 | 35.6 | < .001 | 103.3 | 183.7 |
| Neg. | 107.2* | 3.90 | 29.8 | .001 | 53.3 | 161.0 |
| **Low Predictability** | | | | | | |
| Aff. | 156.2* | 7.70 | 43.4 | < .001 | 116.4 | 195.9 |
| Neg. | 37.0 | 1.58 | 32.3 | .124 | –8.9 | 82.8 |

*Note.* Entries are interpreted in the same way as for Supplementary Table 3, with the same reference values for polarity and number of possible targets. *$p < .05$

## Supplementary Table 7

Results of conducting the model comparisons without data trimming (Experiment 1).

| Model comparison testing for | Dependent variable | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Response accuracy | | | Response time | | | Trajectory clustering | | |
| | $\chi^2$ | d.f. | *p* | $\chi^2$ | d.f. | *p* | $\chi^2$ | d.f. | *p* |
| **Main effect of or interaction involving:** | | | | | | | | | |
| Polarity | 32.0* | 6 | < .001 | 74.0* | 6 | < .001 | 26.2* | 6 | < .001 |
| Possible true completions | 98.9* | 8 | < .001 | 90.6* | 8 | < .001 | 54.0* | 8 | < .001 |
| Truth value | 32.8* | 6 | < .001 | 50.3* | 6 | < .001 | 33.4* | 6 | < .001 |
| **Specifically an interaction involving** | | | | | | | | | |
| Polarity | 7.7 | 5 | .100 | 22.4* | 5 | < .001 | 6.2 | 5 | .289 |
| Possible true completions | 17.6* | 6 | .007 | 9.2 | 6 | .161 | 6.4 | 6 | .381 |
| Truth value | 15.13* | 5 | .010 | 17.6* | 5 | .003 | 2.9 | 5 | .711 |
| **Three-way interaction** | - | - | - | - | - | - | - | - | - |

*Note.* Model comparisons testing for a three-way interaction were only conducted if the preceding tests indicated involvement in an interaction for all three factors. *$p < .05$

## Supplementary Table 8

Results of conducting the model comparisons without data trimming (Experiment 2).

| Model comparison testing for | Dependent variable | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Response accuracy | | | Response time | | | Trajectory clustering | | |
| | $\chi^2$ | d.f. | *p* | $\chi^2$ | d.f. | *p* | $\chi^2$ | d.f. | *p* |
| **Main effect of or interaction involving:** | | | | | | | | | |
| Polarity | 21.9* | 3 | < .001 | 27.5* | 3 | < .001 | 18.6* | 3 | < .001 |
| Possible targets | 9.7* | 4 | .021 | 33.3* | 4 | < .001 | 47.5* | 4 | < .001 |
| **Polarity × poss. targ. interaction** | 4.3 | 2 | .114 | 0.4 | 2 | .82 | 2.1 | 2 | .349 |

*Note.* *$p < .05$

## Supplementary Table 9

Results of conducting the model comparisons without data trimming (Experiment 3).

| Model comparison testing for | Dependent variable | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Response accuracy | | | Response time | | | Trajectory clustering | | |
| | $\chi^2$ | d.f. | *p* | $\chi^2$ | d.f. | *p* | $\chi^2$ | d.f. | *p* |
| **Main effect of or interaction involving:** | | | | | | | | | |
| Polarity | 33.6* | 3 | < .001 | 49.9* | 3 | < .001 | 61.6* | 3 | < .001 |
| Possible targets | 29.9* | 4 | < .001 | 53.5* | 4 | < .001 | 172.9* | 4 | < .001 |
| **Polarity × poss. targ. interaction** | 17.7* | 2 | < .001 | 9.9* | 2 | .007 | 42.5* | 2 | < .001 |

*Note.* *$p < .05$