Conditionals on crutches: Expanding the modal horizon[1]

Dario PAAPE — *Universität Potsdam*
Malte ZIMMERMANN — *Universität Potsdam*

Abstract. Using truth-value judgment tasks, we investigated the on-line processing of counterfactual conditionals such as *If kangaroos had no tails, they would topple over*. Face-value plausibility of the counterfactual as well as the complexity of the antecedent were manipulated. Results show that readers' judgments deviate from face-value plausibility more often when the antecedent is complex, and when the counterfactual is plausible rather than implausible. We interpret our results based on the modal horizon assumption of von Fintel (2001) and argue that they are compatible with a variably strict semantics for counterfactuals (Lewis, 1973). We make use of computational modeling techniques to account for reaction times and truth-value judgments simultaneously, showing that implementing detailed process models deepens our understanding of the cognitive mechanisms triggered by linguistic stimuli.

Keywords: counterfactuals, modal horizon, computational modeling.

1. Introduction

Consider the sentence pair in (1), which constitutes a Sobel sequence (Sobel, 1970).

(1) a. If kangaroos had no tails, they would topple over.
b. If kangaroos had no tails but used crutches, they would not topple over.

Lewis (1973) uses these examples to show that counterfactual conditionals do not allow antecedent strengthening, which would be predicted if they were interpreted as strict conditionals. As the set of worlds in which kangaroos have no tails should contain the set of worlds in which they also use crutches, (1a) and (1b) cannot both be true under a strict analysis. Lewis's solution to the problem is variable strictness: Counterfactual conditionals (henceforth: counterfactuals) are interpreted by taking into account only the set of worlds which are sufficiently similar to the interpretation world. The argument is that the set of worlds being quantified over in (1a) differs from that in (1b) because *crutches* worlds are not among the most similar worlds – compared to the actual world – evoked by the antecedent *If kangaroos had no tails*. In (1b), meanwhile, the explicit mention of the *crutches* scenario in the antecedent forces evaluation relative to a more remote set of worlds, which changes the truth conditions.
As noted by von Fintel (2001), the problem with Lewis' approach is that it does not account for the evolution of the surrounding discourse. If (1b) is uttered as a (possibly pedantic) reply to (1a), the truth of (1a) is called into question. Based on this observation, von Fintel introduces the notion of the *modal horizon*: He proposes that when (1b) is encountered, *crutches* worlds are added to the set of worlds that are accessible to the interlocutors in the discourse, and that the updated set of worlds, when used to interpret (1a), yields a false proposition.
In his analysis, von Fintel (2001) endows every counterfactual sentence with a context change

[1]The authors would like to thank Joseph P. De Veaugh-Geiss, Mareike Philipp and Johannes Schneider for helpful comments and assistance with data collection. We also thank Michael Franke for an insightful discussion. The experiments presented in this paper were funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the collaborative research center "Limits of Variability in Language" (SFB 1287, Project number 317633480), sub-project C02 ("Limits of Variability in Interpretation").

Proceedings of Sinn und Bedeutung 24, vol. 2, pp.108–126. Osnabrück University.

potential, as formalized in (1). The formula states than when the counterfactual is interpreted, the modal horizon f – equivalent to the accessibility function – is updated with the set of worlds evoked by the antecedent ϕ, thus making these additional worlds accessible. The counterfactual is then interpreted using the updated modal horizon.[2]

$$f|\phi > \psi| = \lambda w.f(w) \cup \{w' : \forall w'' \in [\![\phi]\!]^{f,\leq} : w' \leq_w w''\} \tag{1}$$

In the present paper, we consider the possibility that a proper discourse is not needed in order for the modal horizon to be expanded. Strictly speaking, someone who utters (1b) in response to (1a) must *first* expand their internal modal horizon by considering some relatively remote worlds, realize that (1a) is false in at least one of those worlds, and then inform the interlocutor of this fact. We can thus distinguish between an *internal* modal horizon corresponding to the set of worlds that a reader of (1a) or (1b) is able to access (or chooses to access) during interpretation, and an *external* modal horizon that is shared by the interlocutors in a discourse. Our present work is concerned only with the internal mental horizon that readers use when evaluating the truth of a counterfactual statement in isolation.

Below, we present the results of two truth-value judgment experiments that examine the circumstances under which readers become more or less likely to spontaneously expand their internal modal horizon in order to make ostensibly true counterfactuals false, or vice versa.[3] In addition, we present a computational model of counterfactual evaluation that builds on an existing cognitive process model that explicitly links responses and their latencies, and that is largely able to reproduce the qualitative patterns in the empirical data. This type of computational modeling has been previously used in work on long-distance dependencies in sentence processing (Nicenboim and Vasishth, 2018), but has, to our knowledge, not been applied to truth-value judgment data.

1.1. Research questions and predictions

Our research questions can be summarized under three main points:

I. For counterfactuals that are ostensibly true or ostensibly false,[4] how likely are speakers to expand their modal horizon in order to arrive at the opposite truth value? Do they prefer to go from true to false or from false to true? How much do speakers vary in this regard?

II. When the counterfactual in question already has a relatively broad modal horizon, that is, when the antecedent makes reference to relatively remote worlds, does further expansion become more or less likely?

III. What is the relationship between working memory capacity and spontaneous expansion of the modal horizon? If expansion involves cognitive effort, increased capacity should allow for more expansion.

[2]See Gillies (2007) for a similar approach.

[3]For a comprehensive review of previous experimental work on counterfactual processing, see Kulakova and Nieuwland (2016). A review of important research questions concerning conditionals more generally, and of the main theoretical approaches, can be found in Byrne and Johnson-Laird (2009).

[4]We refer to a counterfactual as "ostensibly" true if it rings true "at face value", that is, without expansion of the modal horizon to accommodate further assumptions. Taking the counterfactual "at face value" corresponds to what Kratzer (1979) calls the "stick-close-to-the-relevant-facts" strategy of interpretation. See our explanation of "plausibility" below.

With regard to the questions summarized under (I), it has been observed that humans apparently operate on a truth bias (Zuckerman et al., 1981; Levine et al., 1999), possibly due to the fact that most of the statements an average person is exposed to during their daily life are (considered) true (O'Sullivan et al., 1988). Truth bias is assumed by Truth-Default Theory (Levine, 2014), a general theory of deception detection. While most counterfactual statements are not made in order to actively deceive the interlocutor, an experience-based truth bias should plausibly extend to contexts in which statements may be false in the absence of malicious intent. Truth bias thus predicts that readers should be more likely to expand their modal horizon in order to make an ostensibly false counterfactual true (*If kangaroos had no tails, they would not topple over*) than to make an ostensibly true one false (*If kangaroos had no tails, they would topple over*). Despite this general predicted tendency, depending on their proneness to pedantry (Klecha, 2015) and/or tolerance of "pragmatic slack" and "loose talk" (Lasersohn, 1999; Lauer, 2012), readers may still vary in their willingness to accept a given counterfactual as true.

The intuition behind question (II) is that imagining more distant possible worlds should be more difficult. Assuming that the evaluation of counterfactual statements involves mental simulation (Van Hoeck et al., 2015), it is plausible that simulating more unfamiliar worlds – that is, worlds containing elements that do not correspond to everyday experience, such as kangaroos with crutches – should be more effortful than simulating worlds that are close to the actual world. It is thus predicted that readers should be more likely to assign the ostensible truth value to counterfactual statements with more elaborate antecedents, given that expanding the modal horizon should be dispreferred. A tendency to choose the ostensible truth value in sentences with complex antecedents may also interact with face-value plausibility, so that a truth bias effect may be reduced in complex sentences, in which deviation is costly.

Question (III) is based on the finding that higher working memory capacity is a predictor of more accurate language comprehension (Caplan and Waters, 2005), supports the resolution of long-distance dependencies (Nicenboim et al., 2016), and has an effect on language processing strategies, with high-capacity readers showing more commitment to their chosen interpretations (von der Malsburg and Vasishth, 2013). Readers with higher working memory capacity may engage with counterfactual statements more deeply, considering more possibilities, than readers with lower working memory capacity. Furthermore, given that mental simulation taxes working memory (Ferguson and Cane, 2015; Van Hoeck et al., 2015), high-capacity readers may be able to represent a higher number of possible worlds within their modal horizon compared to low-capacity readers. Assuming that deviation from the ostensible truth value of a given counterfactual is a signal that the modal horizon has been expanded, high-capacity readers should thus show more deviations than low-capacity readers.

2. Truth-value judgment studies

2.1. Experimental design, subjects and materials

We pre-registered our study with the Open Science Framework (Foster and Deardorff, 2017; `https://osf.io/5xbjk`). Our experiments employed a 2×2 design with the factors antecedent complexity (simple vs complex) and plausibility (plausible vs implausible), as shown below. Plausible sentences are intended to be ostensibly true while implausible counterfactuals are intended to be ostensibly false. Plausibility was assessed introspectively by the authors and two additional referees. Implausible sentences were derived from their plausible counterparts by either adding or removing a negation in the consequent. Antecedent complexity was manipulated

by adding a conjunct in the complex version intended to invert the ostensible plausibility of the simple version. The presence of negation was counterbalanced across conditions, so that the same amount of negated consequents was encountered for each combination of the factor levels. Sentences were presented in German.

Simple, Plausible

a. If it was raining burning coals, there would be more forest fires.

Simple, Implausible

b. If it was raining burning coals, there would not be more forest fires.

Complex, Plausible

c. If it was raining burning coals and trees only grew underground, there would not be more forest fires.

Complex, Implausible

d. If it was raining burning coals and trees only grew underground, there would be more forest fires.

The experiments were run using the Linger software (Rohde, 2003). A set of German native speakers read the sentences, which were presented at once in their entirety, pressed the space bar, and then indicated with another key press whether they thought the sentence was TRUE or FALSE. Participants received either €7 or course credit as compensation.[5] Reading times for the entire sentence and response times for the judgments were recorded. There was no time limit. In Experiment 1, a total of 42 participants read 32 counterfactual statements each, plus 64 filler sentences. Experimental sentences were presented according to a Latin-squares procedure, and were randomly intermixed with fillers at runtime. Fillers consisted mainly of philosophical quotes ("Everybody has stupid thoughts, but a wise person keeps them to themselves") and predictions about the future ("In 2030, a manned mission to Mars' moon Phobos will be launched"). Experiment 2 was an attempt to replicate the findings of Experiment 1 with a new set of 41 participants, using the same experimental materials and the same setup. Most participants completed an operation span test before the main experiment (Conway et al., 2005) (see also von der Malsburg and Vasishth, 2013; Nicenboim et al., 2016) to obtain a measure of their working memory capacity. We obtained this measure for 30 of the subjects in Experiment 1 and for all subjects in Experiment 2.

2.2. Data analysis

Even though we collected two latency measures during the experiment – reading time and judgment response time – we simplified our analyses by computing an aggregated measure we call evaluation time, which is the sum of reading time and judgment response time. The reasoning behind this simplification is that participants likely start reasoning about which truth value they want to assign well before the prompt is given, that is, while they are still reading the sentence. To accurately gauge processing difficulty, both latency measures should thus be taken into account. For our initial analysis, as opposed to analyzing the proportion of TRUE and

[5] A subset of seven participants in Experiment 1 was recruited from among personal acquaintances and received no compensation.

FALSE answers across conditions, we analyzed the proportion of cases where the chosen truth value matches the ostensible truth value. This dependent measure is easier to interpret in terms of expansion of the modal horizon (match = no expansion, mismatch = expansion).
We analyzed evaluation times and the likelihood of choosing the ostensible truth value in R (R Core Team, 2019), using the brms package for Bayesian inference (Bürkner, 2017, 2018). All models were fitted using full variance-covariance matrices for the random effects (Barr et al., 2013). Lognormal distributions were fitted for evaluation times, and Bernoulli distributions with a logit link function were fitted for truth value choices. For the factor antecedent complexity, complex was coded as 1 and simple was coded as -1. For plausibility, plausible was coded as 1 and implausible was coded as -1. The interaction was coded as the product of the main effects. Centered, scaled sentence length in characters was entered into the evaluation time models to control for the length confound between simple and complex conditions. Normal(0,5) priors were used across all parameters for all models. LKJ priors (Lewandowski et al., 2009) with ν set to 2 were set for the variance-correlation matrices. Four MCMC chains with 2000 iterations each were run for each model. The first 1000 samples were discarded as warmup. $\hat{R}$ values close to 1 were used to monitor for any cases of non-convergence (Gelman and Rubin, 1992). All data and our analysis code will be released with the publication of this paper at `https://osf.io/y42ve/`.

2.3. Results

Tables 1 and 2 show evaluation times and proportions of chosen truth values by condition for the two experiments. In Experiment 1, participants were less likely to choose the ostensible truth value when the sentence was plausible rather than implausible ($\hat{\Delta} = -0.08$, CrI: $[-0.16, -0.01]$), and when the antecedent was complex rather than simple ($\hat{\Delta} = -0.19$, CrI: $[-0.3, -0.08]$). There was also an interaction ($\hat{\Delta} = -0.07$, CrI: $[-0.13, 0]$), due to complexity having a stronger negative effect on the likelihood of choosing the ostensible truth value in plausible sentences ($\hat{\Delta} = -0.26$, CrI: $[-0.39, -0.13]$) compared to implausible sentences ($\hat{\Delta} = -0.12$, CrI: $[-0.25, 0]$). Evaluation times also showed an interaction between complexity and plausibility ($\hat{\Delta} = 0.53$ s, CrI: [0.17 s, 0.9 s]), such that antecedent complexity only increased the evaluation time for plausible sentences ($\hat{\Delta} = 0.92$ s, CrI: [0.35 s, 1.45 s]).
In Experiment 2, participants were again less likely to choose the ostensible truth value for plausible compared to implausible sentences ($\hat{\Delta} = -0.07$, CrI: $[-0.14, -0.01]$), as well as for sentences with complex compared to simple antecedents ($\hat{\Delta} = -0.18$, CrI: $[-0.3, -0.06]$). As in Experiment 1, there was also an interaction ($\hat{\Delta} = -0.08$, CrI: $[-0.14, -0.02]$), due to a stronger effect of complexity in plausible sentences ($\hat{\Delta} = -0.26$, CrI: $[-0.39, -0.13]$) compared to implausible sentences ($\hat{\Delta} = -0.09$, CrI: $[-0.23, 0.04]$). Unlike in Experiment 1, there was a main effect of working memory capacity, such that higher capacity led to more deviations from the ostensible truth value ($\hat{\Delta} = 0.04$, CrI: $[0.01, 0.06]$).[6] Evaluation times again showed an interaction between complexity and plausibility ($\hat{\Delta} = 0.41$ s, CrI: [0.06 s, 0.77 s]), again due to antecedent complexity increasing evaluation time only for plausible sentences ($\hat{\Delta} = 0.83$ s, CrI: [0.24 s, 1.43 s]).
For both experiments, there was no evidence of any interactions between the experimental manipulations and working memory capacity in any of the measures. Nevertheless, we are

[6]Note that as the working memory predictor was scaled, the estimate represents the effect of increasing working memory capacity by one standard deviation.

interested in how much variability there is in the data with regards to how the manipulations affect each subject. Figure 1 plots the intercepts and slopes – population-level effect plus by-participant adjustments – for each participant on the probability scale. One interesting question to ask is whether all participants show an effect in the same direction, or whether there is evidence that some participants show no or even a reverse effect (Haaf and Rouder, 2017). As Figure 1 shows, we do not have enough data to answer this question for either the effect of plausibility or the plausibility×complexity interaction. For the main effect of complexity, however, we do have good evidence, especially in Experiment 1, that subjects consistently show reduced willingness to assign the ostensible truth value when the antecedent of the counterfactual is complex: As Figure 1 shows, the credible intervals for the complexity effect contain only negative values for almost all of the subjects.

		LOW-CAPACITY PARTICIPANTS			HIGH-CAPACITY PARTICIPANTS		
Plausibility	Complexity	$\overline{ET}$ (s)	p(TRUE)	p(Ost)	$\overline{ET}$ (s)	p(TRUE)	p(Ost)
plausible	simple	8.62	0.78	0.78	8.47	0.80	0.80
plausible	complex	9.91	0.57	0.57	10.44	0.54	0.54
implausible	simple	9.30	0.21	0.79	8.81	0.18	0.82
implausible	complex	9.32	0.27	0.73	8.92	0.30	0.70
filler	filler	8.47	0.45	—	8.55	0.48	—

Table 1: Mean evaluation time, proportion of TRUE answers, and proportion of answers matching the ostensible truth value by condition and WMC group (Experiment 1). Evaluation time has been residualized against sentence length in characters.

		LOW-CAPACITYPARTICIPANTS			HIGH-CAPACITY PARTICIPANTS		
Plausibility	Complexity	$\overline{ET}$ (s)	p(TRUE)	p(Ost)	$\overline{ET}$ (s)	p(TRUE)	p(Ost)
plausible	simple	8.19	0.84	0.84	7.83	0.81	0.81
plausible	complex	9.43	0.53	0.53	9.44	0.65	0.65
implausible	simple	9.07	0.22	0.78	8.13	0.13	0.87
implausible	complex	9.30	0.35	0.65	8.67	0.17	0.83
filler	filler	8.08	0.49	—	8.31	0.47	—

Table 2: Mean evaluation time, proportion of TRUE answers, and proportion of answers matching the ostensible truth value by condition and WMC group (Experiment 2). Evaluation time has been residualized against sentence length in characters.

2.4. Discussion

Across experiments and conditions, participants chose the opposite of the ostensible truth value of the counterfactual about 27% of the time. There is thus evidence that readers routinely extend the set of evaluation worlds beyond the set of worlds evoked by the antecedent.[7]

Most of our predictions are not supported by the results. Participants did not show a truth bias, contrary to the prediction of Truth-Default Theory (Levine, 2014), but instead showed what

[7]The interpretation of the result crucially depends on one's faith in the experimental materials, however. If our intuitive evaluation of the sentences' plausibility does not match that of the average reader, the "ostensible" truth value is merely subjectively ostensible.

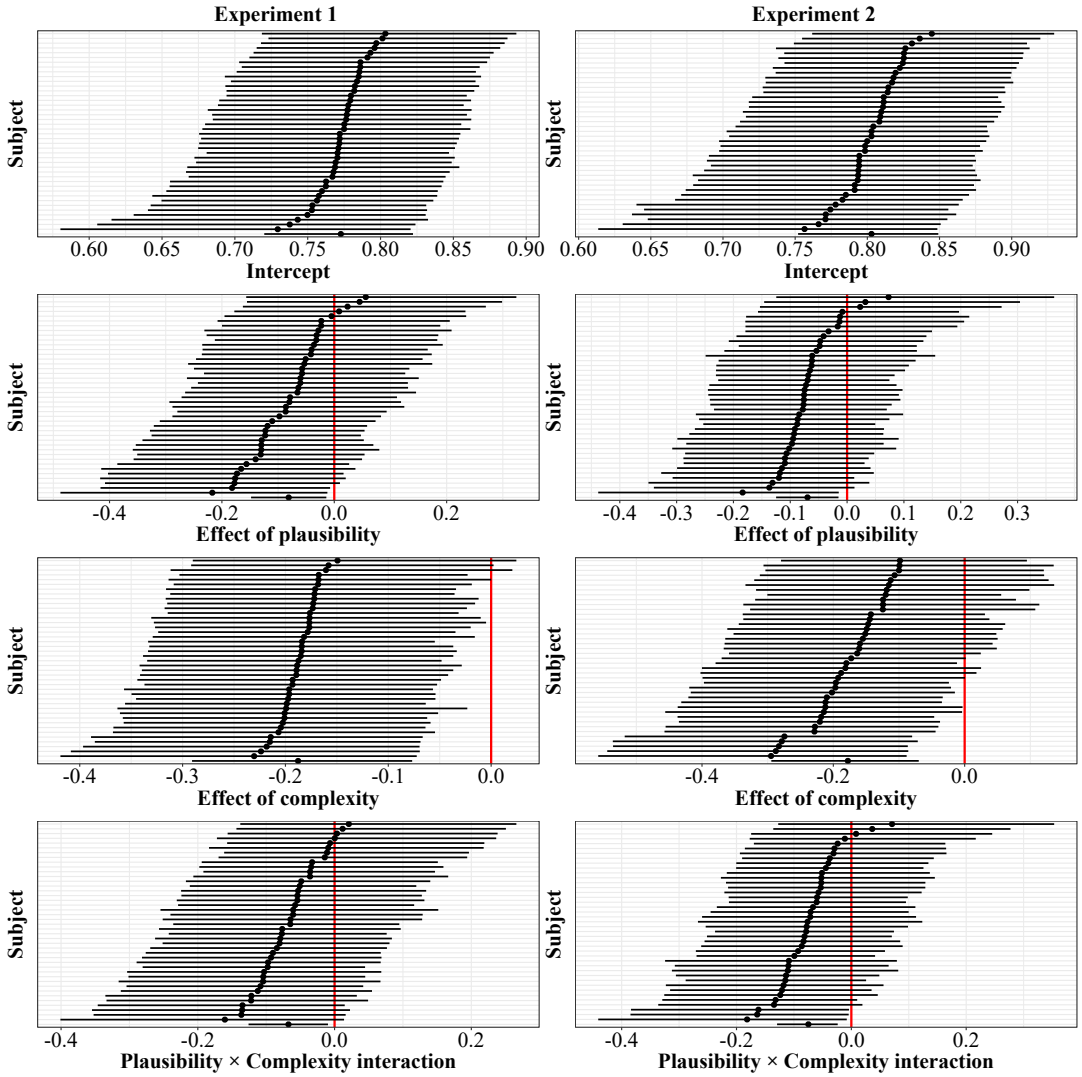

Figure 1: Estimated intercepts and effects of the experimental manipulations on the likelihood of choosing the ostensible truth value by subject and experiment. The average effect is shown at the bottom. Note that estimates are separately ordered by magnitude, so that rows do not map onto each other across plots.

can be thought of as a falsity bias, so that ostensibly true (plausible) sentences were judged to be false more often than ostensibly false (implausible) sentences were judged to be true. Furthermore, increased complexity of the antecedent caused more rather than fewer deviations from the ostensible truth value, contrary to the prediction that expansion of the modal horizon should become more difficult, and therefore less likely, with complex antecedents. While there was an interaction between plausibility and complexity, the direction is unexpected and contrary to our predictions: Participants show an additional increase in deviations for complex, plausible sentences compared to complex, implausible sentences, casting further doubt on both truth bias and any connection between expansion effort and remoteness of the antecedent-evoked set of worlds. The only piece of evidence in favor of expansion being effortful is the observed increase in evaluation times for complex, plausible sentences, combined with the high proportion of

inversions in this condition.
While there is evidence from Experiment 2 that high-capacity participants deviate more often from the ostensible truth value, matching the prediction that expansion should be easier for readers with higher capacity, Experiment 1 yielded no evidence for such a pattern. With regard to the predicted interaction between working memory capacity and antecedent complexity, the data from both experiments are inconclusive: The credible intervals for the interaction both cross zero (Experiment 1: $\hat{\Delta} = -0.01$, CrI: $[-0.06, 0.05]$, Experiment 2: ($\hat{\Delta} = -0.05$, CrI: $[-0.12, 0.01]$), so that the data are compatible with an effect in the predicted direction, an effect in the opposite direction, and with there being no effect at all (Alderson, 2004).

3. Computational modeling of the response process

The analysis presented in the previous section uses statistical models to draw inferences about the cognitive processes involved in interpreting counterfactual sentences. We now make an attempt to derive a plausible process model instead. A process model goes beyond the simple comparison of means across conditions and instead aims to directly account for the cognitive mechanisms that are recruited during interpretation.
We have suggested that the expansion of the modal horizon, which involves the mental simulation of possible worlds, is the process that creates measurable processing effort in our paradigm. This view, however, neglects the fact that the worlds in question also need to be evaluated in terms of whether the consequent holds in them. The variably strict view of counterfactual processing adopted by Lewis (1973) and von Fintel (2001) would dictate that whenever a world in which the consequent is false is encountered within the modal horizon, the entire counterfactual should be judged as false. Other models are possible. For instance, readers may take a counterfactual to be true if the consequent is true in the majority of worlds in the modal horizon. Such an approach can be formalized via the implicit addition of a default operator to the consequent (Ben-David and Ben-Eliyahu, 1994). Irrespective of how the final decision is made, the process involved in reaching a truth-value judgment can be viewed as one of evidence or information accumulation: The processor seeks information (in the form of possible worlds) in favor of answering either TRUE or FALSE, and once one of the options has accumulated enough evidence, the process is terminated and a response is produced.
One of the most well-known evidence-accumulation models for a two-choice task is the diffusion model of Ratcliff (1978). In the original task, a memory item has to be classified as being either recognized or not recognized. Evidence accumulation is controlled by the overlap in features between the recognition probe and potential memory targets from the study phase. Accumulation ends when either of two boundaries, "match" or "mismatch", is reached. When the overlap between the probe and a target is high, the decision is directed towards the "match" boundary, yielding a positive recognition response. Conversely, when overlap is small, the decision is directed towards the "mismatch" boundary. Crucially, reaching the mismatch boundary for one particular memory item is not enough to trigger a negative recognition response to the probe: Only when *all* items have been classified as mismatches is the negative response triggered. The recognition process is thus self-terminating for matches (one match is enough) but exhaustive for mismatches (no recognition only if no items are matched).
Adapting the diffusion model to our truth-value judgment task, the variably strict view of counterfactual interpretation would dictate that FALSE responses are the result of self-termination – because one world in which the consequent does not hold is enough to disprove the counterfactual – while TRUE responses are the result of exhaustive processing, in which all antecedent

worlds under consideration have yielded TRUE for the consequent. Note that such an implementation naturally predicts a falsity rather than a truth bias, as seen in our data, given that, other things being equal, self-termination should occur more often than exhaustive processing. Furthermore, self-termination should also result in faster responses on average compared to exhaustive processing. However, looking at the evaluation time measure, the experimental conditions with a large proportion of FALSE answers do not show reduced evaluation times compared to those with a large proportion of TRUE answers. Indeed, a model fit to evaluation times with the given response as a predictor yields no evidence that FALSE answers were given any faster than TRUE answers ($\hat{\Delta} = 0.01$ s, CrI: $[-0.2\,\text{s}, 0.24\,\text{s}]$). We thus expect that the adaptation of Ratcliff's (1978) model sketched here is not likely to yield an adequate fit to our data.

Another model that assumes gradual accumulation of evidence and directly links response preference to response speed is the lognormal race model of Rouder et al. (2015). In the lognormal race model, the possible responses themselves, rather than individual items in memory, accumulate evidence. The response for which the accumulation process finishes first is produced on a given experimental trial. The finishing times for each accumulator are log-normally distributed with mean μ and standard deviation σ. Effects of the experimental manipulations on finishing times can be estimated relatively straightforwardly by putting a slope on μ and fitting a linear regression, which yields an estimate of how much the corresponding accumulator speeds up or slows down in a given condition. The model assumes that the accumulators are independent and do not compete for mental resources. In terms of simulating mental worlds, the model would thus assume that the TRUE accumulator gathers evidence in favor of the consequent being true, while the FALSE accumulator simultaneously gathers evidence in favor of the consequent being false. This assumption does not rule out the possibility that the same worlds are being accessed during the process: For a given world *w*, the evidence in favor of TRUE is incremented if the consequent is true in *w*, otherwise the evidence in favor of FALSE is incremented.

While a variety of cognitive process models exist that could potentially be used to account for the processing of counterfactual sentences, the lognormal race model strikes us as intuitively plausible and relatively easy to implement. Compared to the more classical approach of analyzing responses and response latencies separately, it offers the advantage of taking into account both sources of information within one and the same trial. Moreover, the lognormal race model also offers a potentially more insightful view of between-participant variability: Rather than preferring to deviate or not deviate from the ostensible truth value of a given sentence, participants may vary in their underlying propensity to answer TRUE or FALSE, represented by faster or slower mean finishing times of the respective accumulators for a given participant. The model thus allows for a more straightforward evaluation of the claim that TRUE judgments should be preferred over FALSE judgments by most readers.

3.1. Implementation of the lognormal race model

We implemented the lognormal race model in Stan (Stan Development Team, 2018). Data from Experiment 1 and 2 were pooled for this analysis. As before, we included antecedent complexity (simple = -1, complex = 1) and plausibility (implausible = -1, plausible = 1) as predictors. Four MCMC chains with 4000 iterations each (with 2000 warmup iterations) were run. The model code with prior settings is given in the on-line supplementary materials at `https://osf.io/y42ve/`. Separate coefficients were estimated for all effects of interest for the TRUE and FALSE accumulators. Separate standard deviations were also assumed for the

accumulators. If FALSE answers are the result of self-termination (see above), one would expect FALSE answers to show more variable response times than TRUE answers, as self-termination may occur at any moment during processing when a FALSE world is found.
Besides main effects of the experimental manipulations and their interaction, coefficients were also estimated for the presence versus absence of negation (no negation = -1, negation = 1), which was manipulated as a cross-balanced between-items factor, as well as for working memory capacity and its interactions with antecedent complexity, plausibility, their two-way-interaction, and negation. For antecedent complexity and plausibility, the model also contains interactions with trial position in the experiment (first half = -1, second half = 1) to see if the experimental manipulations have different effects at the beginning of the experiment compared to the end. By-participants and by-item random effects were added to intercepts and slopes where appropriate.[8]
Instead of including a slope for sentence length in characters on the log scale, the lognormal race model contains a shift estimate on the original millisecond scale (Rouder, 2005). The shift parameter is intended to account for "more peripheral aspects of processing such as encoding stimuli or motor execution of responses" (Rouder, 2005: p. 377). Our goal was to arrive at an estimate of the time it takes to evaluate the truth of the counterfactual by factoring out as much as possible the more low-level aspects of word identification, structure assignment, and so forth. The shift is composed of an intercept and a slope for the number of characters in the sentence, both with estimated by-participant random effects. Fillers contribute to the estimate of each accumulator's intercept, as well as to the estimates for each subject's shift intercept and slope, thus yielding better estimates for these parameters. Compared to fitting separate regression models for TRUE and FALSE answers, the lognormal race model has an additional advantage: When a FALSE answer is given on a trial with some latency x+$shift$, not only do we learn that x must be the FALSE accumulator's finishing time on that particular trial, but also that the latency of the TRUE accumulator must have been larger than x, as the accumulator with the shortest latency always wins.[9]

3.2. Results

Figures 2 and 3 show the distributions of finishing times for the TRUE and FALSE accumulators by working memory group (high versus low capacity, median cut). The plots show the distribution of the means for each observation in the data set, calculated across 8000 post-warmup draws from the posterior predictive distribution of the model. Table 3 compares the data generated by the model with the original data. As the numbers show, the model is mostly able to recover the qualitative aspects of the data in both response proportions and evaluation times, even though it strongly underestimates the proportion of TRUE responses to filler items and simple, plausible sentences.
The estimated mean finishing times for experimental sentences are 13.6 s (CrI: [12 s, 15.48 s]) for the TRUE accumulator and 12.23 s (CrI: [11.08 s, 13.5 s]) for the FALSE accumulator. Plausibility speeds up the TRUE accumulator ($\hat{\Delta} = -7.25$ s, CrI: [-8.9 s, -5.63 s]) and simultaneously slows down the FALSE accumulator ($\hat{\Delta} = 5.38$ s, CrI: [4.14 s, 6.66 s]), resulting in a higher proportion

[8]For instance, negation was not freely manipulated within items and thus has no by-item adjustment, while working memory capacity has no by-participant adjustment.
[9]Note that this also results in the estimated mean finishing time for each accumulator being longer than the respective observed mean finishing time, because finishing times are only observable when the accumulator finishes more quickly than its competitor and otherwise remain latent.

of TRUE answers for plausible compared to implausible sentences. Antecedent complexity slows down TRUE ($\hat{\Delta} = 3.31$ s, CrI: [2.41 s, 4.19 s]), but there is not much evidence that it slows down FALSE ($\hat{\Delta} = 0.47$ s, CrI: [−0.4 s, 1.3 s]), thus more FALSE answers occur in complex sentences. A plausibility×complexity interaction results in a super-additive slowdown on TRUE ($\hat{\Delta} = 3.08$ s, CrI: [1.59 s, 4.6 s]) and a corresponding speedup on FALSE ($\hat{\Delta} = -1.96$ s, CrI: [−3.58 s, −0.4 s]) in complex, plausible sentences. The result is an asymmetrical pattern with more FALSE answers in complex, plausible sentences – meaning more deviations – but fewer FALSE answers in complex, implausible sentences – also meaning more deviations. The presence of negation in the consequent slows down FALSE ($\hat{\Delta} = 1.6$ s, CrI: [0.18 s, 3.03 s]), resulting in more TRUE answers across all conditions for negated sentences.
Working memory speeds up FALSE ($\hat{\Delta} = -0.42$ s, CrI: [−0.74 s, −0.07 s]), so that high-capacity participants give more FALSE answers across all conditions. There is also evidence of an interaction between working memory and plausibility, such that high-capacity participants' speedup for the FALSE accumulator is attenuated in plausible sentences ($\hat{\Delta} = 0.6$ s, CrI: [−0.08 s, 1.28 s]); this means that high-capacity participants are especially likely to respond FALSE to implausible sentences. Both accumulators show a speedup in the second half of the experiment (TRUE: $\hat{\Delta} = -1.4$ s, CrI: [−1.86 s, −0.94 s], FALSE: $\hat{\Delta} = -1.14$ s, CrI: [−1.52 s, −0.76 s]), indicating that responses were given faster overall.

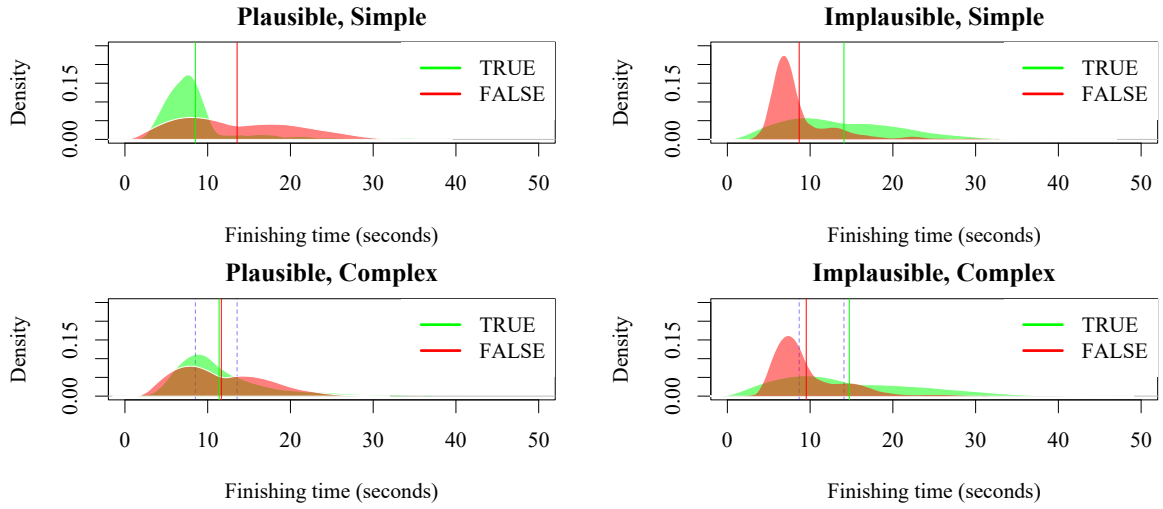

Figure 2: Predicted finishing times of TRUE and FALSE accumulators by condition for low-capacity participants. Vertical lines mark predicted mean finishing times.

3.3. Between-participant variability

Between-participant variability occurs within the model at the levels of the accumulator intercepts as well as the slopes. Variability at the level of the intercepts is shown in Figure 4 for a subset of 11 participants. As the figure shows, there are some participants for which the overall mean finishing time of the TRUE accumulator is faster than that of the FALSE accumulator, but for the majority of subjects the FALSE accumulator is faster.
Figure 5 shows between-participant variability in the slope estimates for plausibility, complexity, and their interaction for the TRUE accumulator. Participants mostly show consistent effects of the manipulations that are in line with the estimated population-level effects, with only

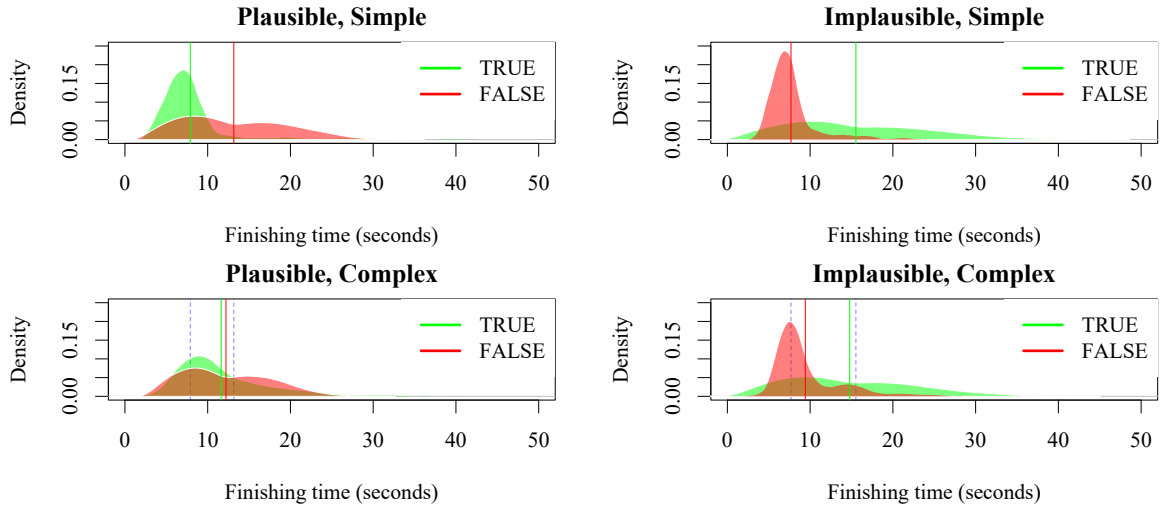

Figure 3: Predicted finishing times of TRUE and FALSE accumulators by condition for high-capacity participants. Vertical lines mark predicted mean finishing times.

		p(TRUE)		$\overline{ET}$ (s)	
Plausibility	Complexity	DATA	MODEL	DATA	MODEL
plausible	simple	0.81	0.66	8.28	8.55
plausible	complex	0.57	0.51	9.82	9.65
implausible	simple	0.19	0.21	8.87	8.80
implausible	complex	0.28	0.25	9.07	9.32
filler	filler	0.47	0.32	8.34	8.36

Table 3: Comparison of empirical and model-generated mean evaluation time and proportion of TRUE answers by condition.

some posterior distributions crossing over the zero line. There is more variability in the effect estimates for plausibility compared to complexity and the interaction, with some extreme effects, indicating participants with a strong tendency to choose the ostensible truth value (topmost four subjects, who show extreme speedups on TRUE for plausible sentences.).
Figure 6 shows between-participant variability in the slope estimates for plausibility, complexity, and the interaction for the FALSE accumulator. Again, the direction of effects is largely consistent across participants and matches the population-level estimates, with more of the posterior distributions for the complexity effect crossing over the zero line compared to the TRUE accumulator. Estimates for the plausibility effect on FALSE are numerically smaller than for the plausibility effect on TRUE, and show less variability. Between-participant variability in choosing the ostensible truth value thus appears to be driven mainly by effects on the TRUE accumulator.

3.4. Comparison with TRUE-default and FALSE-default models

Instead of introducing parameters representing the effects of the experimental manipulations on both accumulators of the lognormal race model, it is, in principle, possible to let the manipulations affect only one of the accumulators while the other acts as a noisy timer (Nicenboim

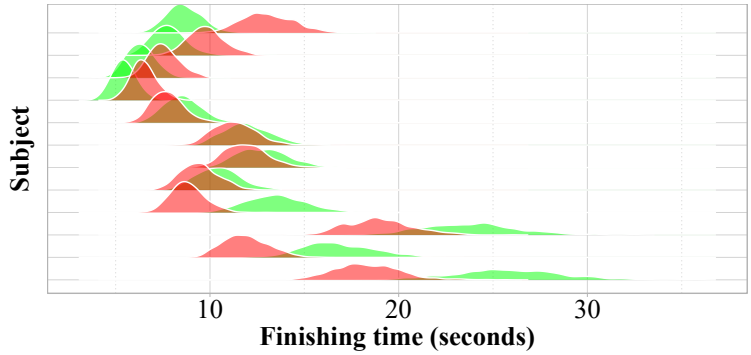

Figure 4: Estimated finishing time intercepts of TRUE and FALSE accumulators for a subset of 11 subjects. Green = TRUE, red = FALSE. Note that estimates are separately ordered by magnitude, so that rows do not map onto each other across figures.

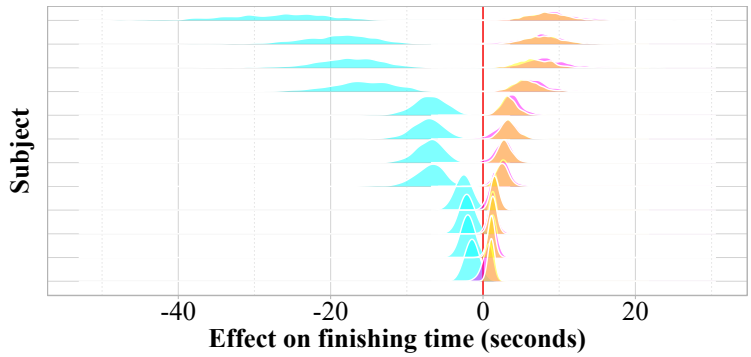

Figure 5: Estimated effects on finishing time of TRUE accumulator for a subset of 11 subjects. Turquoise = plausibility, purple = complexity, yellow = plausibility×complexity interaction. Note that estimates are separately ordered by magnitude, so that rows do not map onto each other across figures.

and Vasishth, 2018). For instance, if plausibility is assumed to affect only the TRUE accumulator while the FALSE accumulator has a fixed mean finishing time across conditions, the race assumption should still allow us to recover the qualitative pattern in the data: When the TRUE accumulator is slowed down in implausible sentences, the FALSE accumulator automatically becomes more likely to win due to the underlying race assumption. Analogous predictions can be made for all other effects in the model. The question is whether a model in which one of the accumulators acts as a "default" fits both the response and the latency profile of the data as well as the more complex model does.

We implemented both a TRUE-default and a FALSE-default model by removing all parameters on the respective accumulator except the intercept, adjustments by-participant and by-item random effects, and coefficients for negation and trial position in the experiment. The latter two parameters were kept because negation could plausibly delay the assignment of a default truth value while hastened responses in later parts of the experiment could also occur for default judgments. Model comparison was carried out using the loo package (Vehtari et al., 2019, 2017), which performs approximate leave-one-out cross-validation using Pareto-smoothed importance

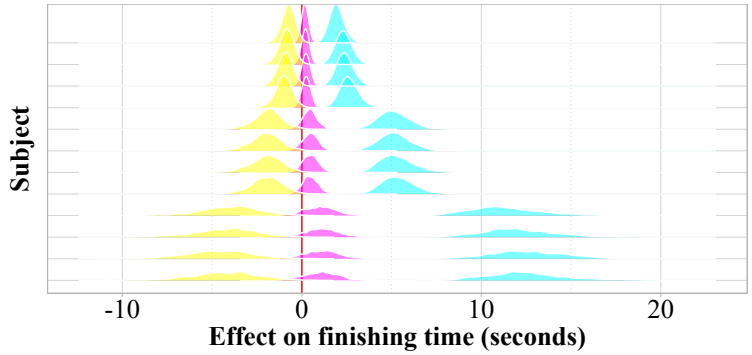

Figure 6: Estimated effects on finishing time of FALSE accumulator for a subset of 11 subjects. Turquoise = plausibility, purple = complexity, yellow = plausibility×complexity interaction. Note that estimates are separately ordered by magnitude, so that rows do not map onto each other across figures.

sampling (PSIS-LOO). This method estimates the expected log predictive density (elpd) of a model, which quantifies how well the model is able to account for unseen data (in this case, single data points from the original data set that are being withheld for validation).
Results of the model comparison show that the FALSE-default model outperforms the TRUE-default model in its predictive accuracy ($\hat{\Delta}$elpd = 102.9, SE = 32.5). The original model featuring the entire range of parameters outperforms both the TRUE-default model ($\hat{\Delta}$elpd = 354.7, SE = 25.9) and the FALSE-default model ($\hat{\Delta}$elpd = 251.8, SE = 20.3). The assumption of additional parameters in the original model is thus justifiable based on the increase in predictive fit. However, should one have strong theoretical reasons to assume a default response, the default is more likely to be FALSE than TRUE based on our model and data.

3.5. Discussion of the modeling results

Besides making the assumptions about cognitive processes involved in the interpretation of counterfactual sentences explicit, the implementation of the lognormal race model yields several insights that go beyond the conclusions drawn from the simple linear modeling approach. The first insight is that the FALSE accumulator gathers evidence more quickly than the TRUE accumulator overall, which results in an overall falsity bias. The second insight is that antecedent complexity mainly affects the rate of accumulation of the TRUE accumulator while plausibility has a nearly symmetrical effect on both accumulators. This pattern can be interpreted as showing that the antecedents of plausible sentences tend to evoke worlds in which the consequent is true while those of implausible sentences tend to evoke worlds in which the consequent is false, as intended by the manipulation. Meanwhile, adding complexity to the antecedent in the form of additional restrictions on the evoked worlds appears to result in fewer TRUE worlds being added to the modal horizon. This may signal that as the modal horizon – that is, the sphere of accessible worlds – expands, opportunities for falsification keep occurring at the same rate and require the same amount of effort, while more and more effort is required for verification. Such a conclusion fits well with the assumption that the interpretation of counterfactuals is, at its core, strict (Lewis, 1973) and that most counterfactuals are, in truth, false if one reasons deeply about them (Hájek, 2014).

Further conclusions from the computational model concern effects of working memory capacity, negation, and the question of whether there is a "default" answer in truth-value judgments of counterfactuals. With regard to working memory, it appears that high-capacity participants have easier access to FALSE worlds, especially in implausible sentences, in which they are naturally evoked by the antecedent. One interpretation of the finding is that participants with high working memory capacity allocate their mental resources more efficiently, allowing them to focus on falsifying the counterfactual as a time-saving strategy. Meanwhile, negation slows down the generation of FALSE responses, for which there are two plausible reasons: One is that responding FALSE to a negated sentence results in an implicit double negation, which may cause readers to doubt their judgment (cf. "The sentence *Cats are not animals* is false"). The other possible reason is that responding FALSE is, in some sense, the default, and that negation introduces uncertainty as to whether the default judgment is correct. That responding FALSE is more likely as to be the default than responding TRUE is also supported by the model comparison results, where the FALSE-default model outperformed the TRUE-default model in terms of predictive accuracy. However, as the "full" model yields even better predictive performance, it appears that a possible default preference for FALSE in the judgment of counterfactuals can be affected by manipulations of plausibility and antecedent complexity.

4. General discussion

Through experimentation and computational modeling, we have been able to shed new light on the semantic processing of counterfactual statements such as *If kangaroos had no tails but used crutches, they would not topple over*. The first, possibly trivial and possibly most important, insight is that there is no absolute consensus between readers as to what the truth value of a given counterfactual should be. The existence of such variability is often overlooked or at least relegated to footnote status in formal accounts of counterfactual interpretation. We have argued that between-participant variability is naturally accounted for by assuming that individual readers may be more or less likely to change their internal modal horizon to contain worlds that result in a flip of the ostensible truth value.

Despite disagreements between language users, there is also a striking amount of consistency in the sense that manipulations of plausibility and antecedent complexity tend to have, in the mean, comparable effects: In our experiments, sentences with ostensibly plausible antecedent-consequent combinations were judged to be TRUE more often than those with ostensibly implausible combinations. Furthermore, participants largely pattern alike in their asymmetrical response to antecedent complexity conditioned on plausibility: Increased antecedent complexity tends to lead to more deviations from the ostensible truth value in plausible compared to implausible sentences. This pattern can be seen as supporting the notion that the interpretation of counterfactuals is strict, as assumed by Lewis (1973), and that most counterfactuals are false Hájek (2014): It is difficult to prove and easy to disprove them, especially when they are based on outlandish premises.

We have also found evidence that implicates working memory in counterfactual processing, which is expected if mental simulation of possible worlds is involved, and matches previous evidence from the processing literature (Ferguson and Cane, 2015). Based on the finding that high-capacity participants are more likely to give FALSE judgments for counterfactuals, and especially in implausible cases, our preliminary conclusion is that these individuals may strategically allocate mental resources to falsification, which increases their efficiency at performing the task.

Finally, we have demonstrated that fitting cognitive process models to empirical data potentially results in a deeper understanding of the theoretical implications by bridging the gap between formal accounts of a phenomenon and the way actual human beings behave when confronted with language. To our minds, the lognormal race model of Rouder et al. (2015) is a natural candidate in this regard: Accumulation of evidence is implied by possible-world semantics and can be plausibly mapped onto evaluating the consequent in a set of antecedent-evoked worlds. Furthermore, the model establishes a direct link between a response and the speed with which it is given, thus linking effort and response preference in a transparent way. The model is mostly able to reproduce the patterns seen in our data, thus validating the approach, though the fit is by no means perfect. This, however, is only to be expected, given that

> [...] it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. [...] The only question of interest is "Is the model illuminating and useful?" (Box, 1979: p. 202–203)

We would argue that the lognormal race model as applied to truth-value judgments fulfills both conditions and should be applied to other types of sentences in future work.

4.1. A non-exhaustive list of caveats

Several caveats are in order. We have assembled them here in the form of a non-exhaustive list of bullet points for ease of reference.

- When looking at the entire experiment, we see a falsity bias even for filler sentences. It is possible that because participants encountered a large number of implausible fillers, they may have *developed* the falsity bias observed for experimental sentences during the experiment. While our analysis revealed no interactions between early versus late trials and our manipulations, it is possible that these interactions are more complex than our model can account for (Baayen et al., 2017). Furthermore, when participants are explicitly asked to judge whether a given sentence is true or false, any truth bias present in everyday language use may be suspended, and thus our findings may not generalize to more naturalistic settings.
- Intuitively, the discourses A-B and A'-B' below are not entirely parallel, though up until now we have been implicitly treating them as a minimal pair:

  A. If kangaroos had no tails, they would topple over.
  B. Not true! If kangaroos had no tails but used crutches, they would not topple over.

  A'. If kangaroos had no tails, they would not topple over.
  B'. True! If kangaroos had no tails but used crutches, they would not topple over.

  In A-B, the interlocutor expands the modal horizon to retroactively render the speaker's utterance false. By contrast, in A'-B', the interlocutor "saves" the speaker's ostensibly false utterance by changing the modal horizon. "Expansion" is something of a misnomer in the A'-B' case: Here, the modal horizon must *not* contain any not-crutches worlds after the update, so that worlds are actively being removed from the initial scope of the accessibility function. In A-B, not-crutches worlds can safely remain within the scope

of the accessibility function: As long as there is an accessible subset of crutches worlds, strictness guarantees the falsity of the initial utterance. The asymmetry casts doubt on the underlying assumption that the same processes are responsible for true-to-false and false-to-true changes, as it may be that elimination of worlds from the modal horizon is a separate mechanism with a discernible cost.

- We have chosen to adopt a possible world semantics for counterfactuals (Lewis, 1973) in combination with von Fintel's (2001) notion of a dynamic modal horizon as the theoretical starting point for our investigation. Both of these accounts have been criticized in the literature. Ciardelli et al. (2018) present experimental evidence from counterfactuals with disjunctive antecedents that they argue to be incompatible with the notion of a similarity-based accessibility function as employed by both Lewis and von Fintel. Moss (2012) argues that the relevant properties of Sobel sequences that motivate the modal horizon assumption can be accounted for by pragmatic factors governing the felicity of utterances in a context. We remain agnostic as to how our findings can be accounted for under alternative approaches to counterfactual interpretation, but note that deviations from the ostensible truth value of a sentence are in need of explanation.

5. Conclusion

Our experimental studies were not concerned so much with the truth conditions of counterfactuals, but rather with what Stalnaker (1986) calls their "belief conditions": We did not ask when counterfactuals are true,[10] but under which conditions language users *accept* them as true, and with how their judgments are reached. In essence, we assume that subjects conduct a Ramsey test (Ramsey, 1931; Stalnaker, 1986) by temporarily assuming that the antecedent is true, entering into the most accessible possible worlds evoked by this assumption, and evaluating the truth of the consequent in those worlds. Our results suggest that for counterfactuals with complex antecedents, subjects follow what Kratzer (1979) calls the "skeptical" strategy, which yields strict interpretations: If they can find evaluation worlds which render the counterfactual false, they appear to do so. When subjects choose what we call the ostensible truth value of the counterfactual, they are instead following the "keep-close-to-the-relevant-facts" strategy: They do not expand their internal modal horizon beyond what the antecedent necessitates.

References

Alderson, P. (2004). Absence of evidence is not evidence of absence. *BMJ 328*(7438), 476–477.

Baayen, H., S. Vasishth, R. Kliegl, and D. Bates (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language 94*, 206–234.

Barr, D. J., R. Levy, C. Scheepers, and H. J. Tily (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language 68*(3), 255–278.

Ben-David, S. and R. Ben-Eliyahu (1994). A modal logic for subjective default reasoning. In *Proceedings Ninth Annual IEEE Symposium on Logic in Computer Science*, pp. 477–486. IEEE.

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pp. 201–236. Elsevier.

[10] For a recent discussion of counterfactual truth values, see von Prince (2019).

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software 80*(1), 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal 10*(1), 395–411.

Byrne, R. M. and P. N. Johnson-Laird (2009). 'If' and the problems of conditional reasoning. *Trends in Cognitive Sciences 13*(7), 282–287.

Caplan, D. and G. Waters (2005). The relationship between age, processing speed, working memory capacity, and language comprehension. *Memory 13*(3–4), 403–413.

Ciardelli, I., L. Zhang, and L. Champollion (2018). Two switches in the theory of counterfactuals. *Linguistics and Philosophy 41*(6), 577–621.

Conway, A. R., M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review 12*(5), 769–786.

Ferguson, H. J. and J. E. Cane (2015). Examining the cognitive costs of counterfactual language comprehension: Evidence from ERPs. *Brain Research 1622*, 252–269.

Foster, E. D. and A. Deardorff (2017). Open Science Framework (OSF). *Journal of the Medical Library Association: JMLA 105*(2), 203–206.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Gillies, A. S. (2007). Counterfactual scorekeeping. *Linguistics and Philosophy 30*(3), 329–360.

Haaf, J. M. and J. N. Rouder (2017). Developing constraint in Bayesian mixed models. *Psychological Methods 22*(4), 779–798.

Hájek, A. (2014). Most counterfactuals are false. Ms, Australian National University.

Klecha, P. (2015). Two kinds of Sobel sequences: Precision in conditionals. In *West Coast Conference on Formal Linguistics (WCCFL)*, Volume 32, pp. 131–140.

Kratzer, A. (1979). Conditional necessity and possibility. In *Semantics from different points of view*, pp. 117–147. Springer.

Kulakova, E. and M. S. Nieuwland (2016). Understanding counterfactuality: A review of experimental evidence for the dual meaning of counterfactuals. *Language and Linguistics Compass 10*(2), 49–65.

Lasersohn, P. (1999). Pragmatic halos. *Language*, 522–551.

Lauer, S. (2012). On the pragmatics of pragmatic slack. In *Proceedings of Sinn und Bedeutung*, Volume 16, pp. 389–401.

Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology 33*(4), 378–392.

Levine, T. R., H. S. Park, and S. A. McCornack (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs 66*(2), 125–144.

Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis 100*(9), 1989–2001.

Lewis, D. K. (1973). *Counterfactuals*. Blackwell.

Moss, S. (2012). On the pragmatics of counterfactuals. *Noûs 46*(3), 561–586.

Nicenboim, B., P. Logačev, C. Gattei, and S. Vasishth (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology 7*, 280.

Nicenboim, B. and S. Vasishth (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language 99*, 1–34.

O'Sullivan, M., P. Ekman, and W. V. Friesen (1988). The effect of comparisons on detecting deceit. *Journal of Nonverbal Behavior 12*(3), 203–215.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsey, F. P. (1931). General propositions and causality. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and other Logical Essays*, pp. 237–255. London: Kegan Paul, Trench & Trubner.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review 85*(2), 59.

Rohde, D. (2003). Linger. Version 2.94. Available at: http://tedlab.mit.edu/∼dr/Linger/.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika 70*(2), 377–381.

Rouder, J. N., J. M. Province, R. D. Morey, P. Gomez, and A. Heathcote (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika 80*(2), 491–513.

Sobel, J. H. (1970). Utilitarianisms: Simple and general. *Inquiry 13*(1–4), 394–449.

Stalnaker, R. (1986). A theory of conditionals. In P. G. Harper W.L., Stalnaker R. (Ed.), *IFS. The University of Western Ontario Series in Philosophy of Science (A Series of Books in Philosophy of Science, Methodology, Epistemology, Logic, History of Science, and Related Fields), vol. 15*, pp. 41–55. Dordrecht: Kluwer.

Stan Development Team (2018). The Stan Core Library. Version 2.18.0.

Van Hoeck, N., P. D. Watson, and A. K. Barbey (2015). Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience 9*, 420.

Vehtari, A., J. Gabry, Y. Yao, and A. Gelman (2019). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.1.0.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing 27*, 1413–1432.

von der Malsburg, T. and S. Vasishth (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes 28*(10), 1545–1578.

von Fintel, K. (2001). Counterfactuals in a dynamic context. *Current Studies in Linguistics Series 36*, 123–152.

von Prince, K. (2019). Counterfactuality and past. *Linguistics and Philosophy 42*(6), 577–615.

Zuckerman, M., B. M. DePaulo, and R. Rosenthal (1981). Verbal and nonverbal communication of deception. In *Advances in Experimental Social Psychology*, Volume 14, pp. 1–59. Elsevier.