Journal of Experimental Psychology: Learning, Memory, and Cognition

Memory Modeling of Counterfactual Generation

Feiyi Wang, Ada Aka, Lisheng He, and Sudeep Bhatia

Online First Publication, April 4, 2024. https://dx.doi.org/10.1037/xlm0001335

CITATION

Wang, F., Aka, A., He, L., & Bhatia, S. (2024). Memory modeling of counterfactual generation.. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://dx.doi.org/10.1037/xlm0001335

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Learning, Memory, and Cognition

© 2024 American Psychological Association
ISSN: 0278-7393

https://doi.org/10.1037/xlm0001335

Memory Modeling of Counterfactual Generation

Feiyi Wang[1], Ada Aka[2], Lisheng He[3], and Sudeep Bhatia[1]
[1] Department of Psychology, University of Pennsylvania
[2] Department of Marketing, Stanford Graduate School of Business
[3] SILC Business School, Shanghai University

We use a computational model of memory search to study how people generate counterfactual outcomes in response to an established target outcome. Hierarchical Bayesian model fitting to data from six experiments reveals that counterfactual outcomes that are perceived as more desirable and more likely to occur are also more likely to come to mind and are generated earlier than other outcomes. Additionally, core memory mechanisms such as semantic clustering and word frequency biases have a strong influence on retrieval dynamics in counterfactual thinking. Finally, we find that the set of counterfactuals that come to mind can be manipulated by modifying the total number of counterfactuals that participants are prompted to generate, and our model can predict these effects. Overall, our findings demonstrate how computational memory search models can be integrated with current theories of counterfactual thinking to provide novel insights into the process of generating counterfactual thoughts.

Keywords: counterfactual thinking, memory, computational modeling, decision making

Counterfactual thinking, or the ability to imagine alternative possibilities to an event or outcome, is ubiquitous (Byrne, 2016; De Brigard & Parikh, 2019; Phillips et al., 2019). Once they come to mind, counterfactuals have a wide array of effects on cognition and behavior. For example, judgments of causality depend on salient counterfactuals, and counterfactual assessment is a key component in cognitive models of causal judgment (Gerstenberg & Tenenbaum, 2017; Sloman & Lagnado, 2015; Wells & Gavanski, 1989). In social settings, counterfactuals that come to mind determine judgments of responsibility and the moral evaluations of acts (Greene et al., 2004; Zultan et al., 2012). Counterfactuals also have important implications for mental health, as they are the basis of emotions like sadness, anxiety, and regret (Roese & Epstude, 2017). Finally, judgments, decisions, and evaluations rely critically on counterfactuals, with desirable but unattained counterfactuals reducing the judged desirability of attained choice outcomes (Loomes & Sugden, 1982; Mellers et al., 1997; Stewart et al., 2006).

Due to cognitive limitations, counterfactual thoughts that spontaneously come to mind at a given time are only a sample of the vast possibilities that one can consider. Thus, unsurprisingly, many researchers have attempted to study the determinants of counterfactual generation. In norm theory, Kahneman and Miller (1986) have argued that people generate counterfactual alternatives that are similar to the outcomes being evaluated. In a recent review, Phillips et al. (2019) have found that people tend to sample possible actions that possess two properties: (a) high desirability and (b) high likelihood of occurrence. Furthermore, Bear et al. (2020) have argued that this sampling strategy has an adaptive origin as it helps people efficiently make good decisions. Other related work has found that the exceptionality (Kahneman & Miller, 1986), causal ordering (Wells et al., 1987), controllability (Girotto et al., 1991), moral permissibility (Phillips & Cushman, 2017), and perceived similarity (De Brigard et al., 2021) of counterfactuals also influence the probability that they come to mind and influence cognition and behavior.

Although the above work has provided many important insights about counterfactual thinking, we still do not possess a computational model that formally describes how different variables and cognitive mechanisms guide and constrain the generation of counterfactual outcomes in response to an experienced outcome. Such models are common in memory research and are used to study list recall (Polyn et al., 2009; Raaijmakers & Shiffrin, 1981), semantic memory search (Abbott et al., 2015; Hills et al., 2012), free association (De Deyne et al., 2013, 2019), and decision making (Aka & Bhatia, 2021; Bhatia, 2019). However, extending them to counterfactual generation has been difficult. This is partially due to the types of tasks in which counterfactual thought is studied. These tasks typically involve high-level causal, social, and evaluative judgments, which can evoke (and can be influenced by) an almost infinite set of complex counterfactuals. Formally specifying the complete set of relevant counterfactuals, developing quantitative models that predict the sequence in which these counterfactuals are likely to come to mind, and fitting these models to sequences of counterfactual items generated by human participants are nearly impossible in such tasks.

Feiyi Wang https://orcid.org/0009-0003-2497-7545

Funding was received from the National Science Foundation (SES-1847794) and National Natural Science Foundation of China (72101156). Data, code, and preregistration are available at https://osf.io/497ct/.

Feiyi Wang served as lead for data curation, formal analysis, visualization, and writing–original draft and contributed equally to methodology. Ada Aka served in a supporting role for formal analysis, methodology, visualization, and writing–original draft. Lisheng He served as lead for hierarchical Bayesian modeling and served in a supporting role for writing–review and editing. Sudeep Bhatia served as lead for supervision and served in a supporting role for writing–review and editing. Feiyi Wang and Ada Aka contributed equally to conceptualization and writing–review and editing.

Correspondence concerning this article should be addressed to Feiyi Wang, Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, United States. Email: feiyiw@sas.upenn.edu

Yet developing computational memory models of counterfactual generation is of vital importance. Such an exercise would place established empirical findings on counterfactual generation within a formal theoretical framework. This framework could, in turn, be used to characterize the structure of variability in counterfactual thoughts across individuals, explain the influences of different tasks and contexts, and predict with high accuracy the counterfactuals that are generated in response to novel stimuli. A computational model could also be used to parameterize the effects of multiple distinct mechanisms and cues on counterfactual generation and quantitatively test which mechanisms play the largest role. Finally, a formal model of counterfactual generation would be able to test the existence of several established memory regularities in counterfactual thought. For instance, researchers have found that items that are retrieved from memory are semantically related to items that have previously been retrieved (Bousfield & Sedgewick, 1944; Cofer et al., 1966; Gruenewald & Lockhead, 1980; Howard & Kahana, 2002; Romney et al., 1993), a phenomenon known as semantic clustering. Word frequency, or how commonly a word appears in natural language, is also an important memory cue (Gorman, 1961; Hall, 1954; Lohnas & Kahana, 2013; Nelson et al., 2000; Sumby, 1963). Both semantic clustering and word frequency effects could be involved in counterfactual generation, with important implications for cognitions and behaviors that rely on counterfactual thought.

Our goal in this article is to build a computational model of counterfactual generation that can describe and predict the sequences of counterfactual outcomes that come to mind in response to a particular target outcome. For this purpose, we utilize a variant of a free association task. In our experiments, participants are told to consider a target outcome and, while they do so, are asked to list the set of counterfactual outcomes that come to their minds (in the order in which these counterfactuals come to their minds). Our model takes the form of a Markov random walk over items in memory, which treats counterfactual generation as a stochastic process in which the probability that an item comes to mind depends only on the most recently generated item (i.e., walking from one item to the next over a network of connected items). The Markov random walk is a basic model of memory search that emerges as a special case from more complex theories (e.g., Hills et al., 2012; Polyn et al., 2009; Raaijmakers & Shiffrin, 1981; Richie et al., 2023; Zhao et al., 2022; see Kahana, 2020 for a discussion). It is frequently used to study free association and semantic memory search (Abbott et al., 2015; De Deyne et al., 2019) and has also been applied to study memory processes in decision making (Aka & Bhatia, 2021).

To extend this model to counterfactual generation, we assume that the probability of generating a counterfactual outcome at a given point in time depends on several variables, including the variables studied in prior work (such as the desirability of the counterfactual outcome, the likelihood of occurrence of the counterfactual outcome, and the similarity between the counterfactual outcome and the target outcome). Critically, we also allow for the effect of new variables (such as the semantic similarity between the counterfactual outcome and the previously generated counterfactual, and the word frequency of the counterfactual outcome) that are implicated in memory search but have not been studied in the context of counterfactual generation. By jointly modeling several variables implicated in counterfactual research as well as new variables implicated in memory research, we can examine the effect of each variable while controlling for the others to offer comprehensive insights into the dynamics of counterfactual generation.

Experiment 1

Experiment 1 tested a computational model of counterfactual generation and examined the role of five key variables in its dynamics. To test our model, we designed a task involving fruits and vegetables. In this task, participants were asked to recall the last piece of fruit or vegetable that they ate. Then, they were asked to generate counterfactual fruits or vegetables that they could have eaten instead. Participants also provided baseline evaluations of fruits and vegetables in a prior session. We chose this scenario as it involves a finite and tractable set of items that could serve as counterfactual outcomes (i.e., fruits and vegetables). These items can be specified a priori and thus form the network of items over which our memory model operates. We fit our model to the data using hierarchical Bayesian modeling.

Method

Participants

Participants in Experiment 1 ($N = 59$; $M_{\text{age}} = 51$; 51% female, 46% male, 3% nonbinary) were recruited from Prolific Academic and performed the experiment online using their own computer interface. Participation was limited to native English speakers in the United States.

Procedures

Experiment 1 had two sessions (Figure 1A). In the first session, participants evaluated a comprehensive list of items. A week later, in the second session, participants were asked to generate a target item and list 10 counterfactual items that came to mind as they considered the target item. We used the item evaluations from the first session (along with other data sources) to model the counterfactual generation processes at play in the second session.

More specifically, in Session 1, participants were first asked to rate a list of fruits and vegetables in terms of their desirability (e.g., "how much would you like to eat apples?") and likelihood of occurrence (e.g., "how probable is it that you would eat apples?"). All ratings were made on a scale from 0 to 100 (with 0 corresponding to *extremely undesirable/unlikely* and 100 corresponding to *extremely desirable/likely*). To avoid systematic sequence effects in the ratings, we presented the list of items to each participant at a random order.

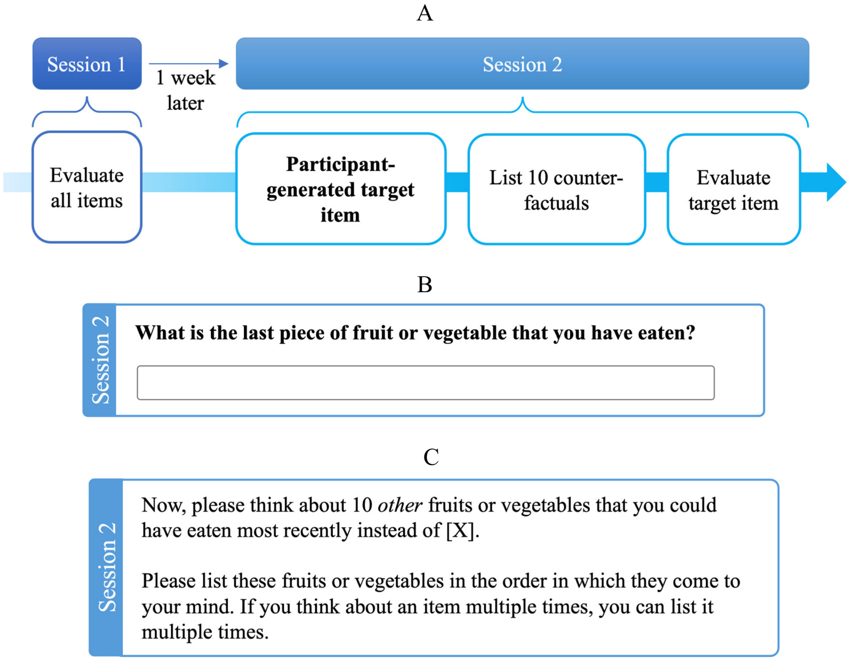
In Session 2, participants were first asked to indicate the last piece of fruit or vegetable that they ate (Figure 1B). To accommodate the possibility that the order of the words, "fruit" and "vegetable," in the phrase "fruit or vegetable" would bias generation toward one type of item, we counterbalanced the order of these two words in the survey and collapsed them for data analysis. After eliciting the target item, participants were then asked to think about 10 other fruits or vegetables that they could have eaten instead of the target item (Figure 1C). Participants were asked to list these counterfactuals on 10 successive screens and were allowed to list the same item multiple times. Finally, participants provided desirability and likelihood ratings of the target item as well as the counterfactual items that they generated.

Stimuli

We created a comprehensive list of 188 fruits and vegetables that exist in the Google News Word2Vec semantic space (Mikolov et al.,

Figure 1

Experiment 1 Design and Prompt Examples



What is the last piece of fruit or vegetable that you have eaten?

Now, please think about 10 *other* fruits or vegetables that you could have eaten most recently instead of [X].

Please list these fruits or vegetables in the order in which they come to your mind. If you think about an item multiple times, you can list it multiple times.

Note. (A) Schematic of the task design for Experiment 1. (B) Example prompt for target elicitation. (C) Example prompt for counterfactual generation. See the online article for the color version of this figure.

2013). This list can be found at https://osf.io/497ct/ (Wang et al., 2023). In Session 1, participants provided baseline ratings for each of these 188 items. Occasionally in Session 2, participants listed items that are not part of this list. We excluded one participant who listed a target item that was not one of those 188 items, and four additional participants who listed as counterfactuals more than 50% of such items. Among the remaining participants, 3.73% of the counterfactual items they listed were not one of those 188 items and these items were thus excluded from further analyses. Occasionally in Session 2, participants listed multiple items on the first screen when they were asked to list only the first item that comes to mind. We asked participants to list one item per screen in order to avoid direct cuing of the previous items on the subsequent items, so we excluded additional items that were listed on the same screen beyond the very first item. This is not an ideal procedure, but we chose it to keep as much data as possible (21 participants made such responses).

Overall, our participants generated 28 different target items and 90 different counterfactual items. The most frequently elicited target items were "apple" (7 times) and "broccoli" (5 times). The most common counterfactual items were "orange" (32 times) and "banana" (32 times).

Results

Empirical Patterns

In this section, we examined the effect of each variable on counterfactual generation in isolation (i.e., without controlling for other variables). Table 1 provides a summary of the descriptive results. These empirical patterns provide support for findings in prior literature. In the Modeling Results section, we will examine the effects of these variables jointly using our modeling framework.

Desirability and Likelihood. We first attempted to test the effect of desirability and likelihood of occurrence on counterfactual generation. Prior work has found that outcomes that are perceived as highly desirable and highly likely to occur also have higher probability of being generated as counterfactuals (see, e.g., Phillips et al., 2019). To test these effects in our experiments, we first computed the probability that each item gets listed as a counterfactual in Session 2, and then correlated it with each item's aggregate desirability and likelihood ratings elicited in Session 1. Consistent with previous literature, our tests revealed a positive correlation both for item desirability, $r(186) = .675$, $p < .001$, 95% confidence interval (CI) = [0.589, 0.746], and item likelihood, $r(186) = .665$, $p < .001$, 95% CI = [0.577, 0.738]. These relationships are shown in Figure 2A and 2B, which segment items into 10 equal-width bins based on their desirability or likelihood ratings, respectively, and then plot each bin's aggregate generation probability.

We also examined the relationship between the order in which the counterfactual items are generated in Session 2 and these item's desirability and likelihood evaluations in Session 1. The order is specified as a number from 1 to 10, with 1 being the first item generated, and 10 being the last. We observed a significant negative relationship between average item desirability and order (Spearman's rank correlation was computed because the latter is on an ordinal scale), $r_s(8) = -.636$, $p = .048$, 95% CI = [−0.904, −0.011], and

Table 1
Summary of Statistical Tests in the Observed (O) and Simulated (S) Data in Experiments 1–4

Value	Pearson's *r* between value and generation probability (O; S)	Spearman's rho between value and order of generation (O; S)	One sample *t* test between value and chance (O; S)	Spearman's rho between value and bin number (O; S)
Experiment 1				
Des.	.675***; .759***	−.636*; −.842**	—	—
Lik.	.665***; .766***	−.927***; −.863**	—	—
Freq.	.576***; .701***	−.515; −.721*	—	—
Tar. Sim.	—	−.939***; −.915***	14.45***; 122.22***	—
CRP	—	—	—	.745*; .891***
Experiment 2 List 5				
Des.	.635***; .690***	−.700; −.999***	—	—
Lik.	.626***; .683***	−.900*; −.999***	—	—
Freq.	.551***; .713***	−1.000***; −.900*	—	—
Tar. Sim.	—	−.500; −.999***	10.67***; 77.61***	—
CRP	—	—	—	.796**; .524
Experiment 2 List 20				
Des.	.745***; .820***	−.764***; −.893***	—	—
Lik.	.710***; .813***	−.817***; −.710***	—	—
Freq.	.595***; .799***	−.756***; −.853***	—	—
Tar. Sim.	—	−.281; −.702***	15.94***; 141.19***	—
CRP	—	—	—	.794**; .939***
Experiment 3A				
Des.	.677***; .706***	−.479; −.770**	—	—
Lik.	.805***; .865***	−.430; −.842**	—	—
Freq.	.541***; .608***	−.139; −.818**	—	—
Tar. Sim.	—	−.697*; −733*	17.04***; 134.00***	—
CRP	—	—	—	.952***; .999***
Experiment 3B				
Des.	.652***; .702***	−.733*; −.915***	—	—
Lik.	.772***; .858***	−.770**; −.685*	—	—
Freq.	.546***; .631***	−.430; −.770**	—	—
Tar. Sim.	—	−.891***; −.855**	22.29***; 154.13***	—
CRP	—	—	—	.999***; .988***
Experiment 3C				
Des.	.492***; .718***	−.576; −.879***	—	—
Lik.	.538***; .756***	−.236; −.952***	—	—
Freq.	.490***; .764***	−.564; −.976***	—	—
Tar. Sim.	—	−.879***; −.879***	17.82***; 175.93***	—
CRP	—	—	—	.745*; .999***
Experiment 4 List 5				
Des.	.615***; .739***	−.999***; −.900*	—	—
Lik.	.692***; .848***	−.999***; −.900*	—	—
Freq.	.438***; .588***	−.600; −.300	—	—
Tar. Sim.	—	−.800; −.700	25.67***; 156.15***	—
CRP	—	—	—	.976***; .976***
Experiment 4 List 20				
Des.	.706***; .745***	−.732***; −.729***	—	—
Lik.	.790***; .861***	−.814***; −.928***	—	—
Freq.	.626***; .697***	−.580**; −.695***	—	—
Tar. Sim.	—	−.947***; −.926***	32.46***; 249.41***	—
CRP	—	—	—	.999***; .988***

Note. Des. = desirability; Lik. = likelihood of occurrence; Word Freq. = log-transformed word frequency; Tar. Sim. = similarity with the target; CRP = conditional response probability.
* $p < .05$. ** $p < .01$. *** $p < .001$.

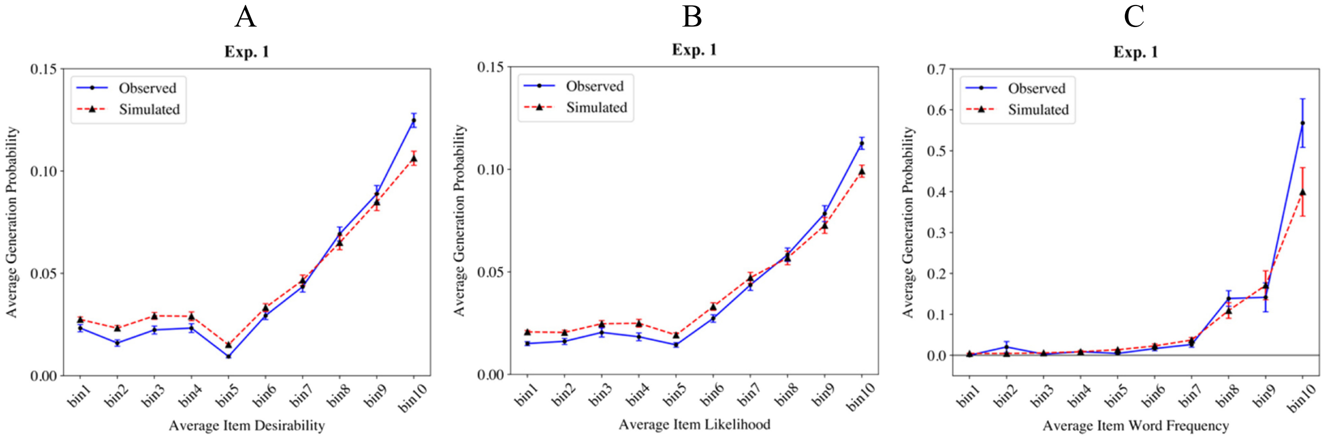
between average item likelihood and order, $r_s(8) = -.927, p < .001$, 95% CI = [−0.983, −0.714]. These relationships are shown in Figure 3A and 3B, respectively.

Finally, we would like to point out that desirability and likelihood ratings were highly correlated, $r(186) = .959, p < .001$, 95% CI = [0.946, 0.969]. This is quite reasonable because people are more likely to eat fruits and vegetables which they find desirable rather than undesirable. However, it indicates that positive effects of desirability on counterfactual generation could have been due to the likelihood variable, or vice versa. We return to this issue in the Memory Model section.

Similarity With the Target. Prior literature has also suggested that people are more likely to think about counterfactual outcomes that are similar with the target outcome than ones that are dissimilar (Kahneman & Miller, 1986). To objectively measure the degree of similarity between the target and the counterfactuals and to quantitatively test this relationship, we used the 300-dimensional distributed vector representations from the Google News Word2Vec model (Mikolov et al., 2013). As in prior work, we measured the similarity between two items by their vectors' cosine similarity (Aka & Bhatia, 2021; Bhatia, 2019; see Bhatia et al., 2019 for a

Figure 2

Experiment 1 Observed Versus Predicted Generation Probabilities



Note. Average observed and model-predicted probabilities of counterfactual generation as a function of item (A) desirability decile, (B) likelihood decile, and (C) word frequency (log-transformed) decile. Error bars display ±1 *SE*. See the online article for the color version of this figure.

review). If similarity with the target cues counterfactual generation, then the cosine similarity between a counterfactual item and the target item should be higher than the expected cosine similarity between two randomly selected items in the list of 188 fruits and vegetables (which is .426). We tested this using a one sample *t* test and found that the average similarity between participant-generated counterfactuals and their corresponding targets is significantly higher than chance, $t(567) = 14.452$, $p < .001$, 95% CI = [0.496, 0.518].

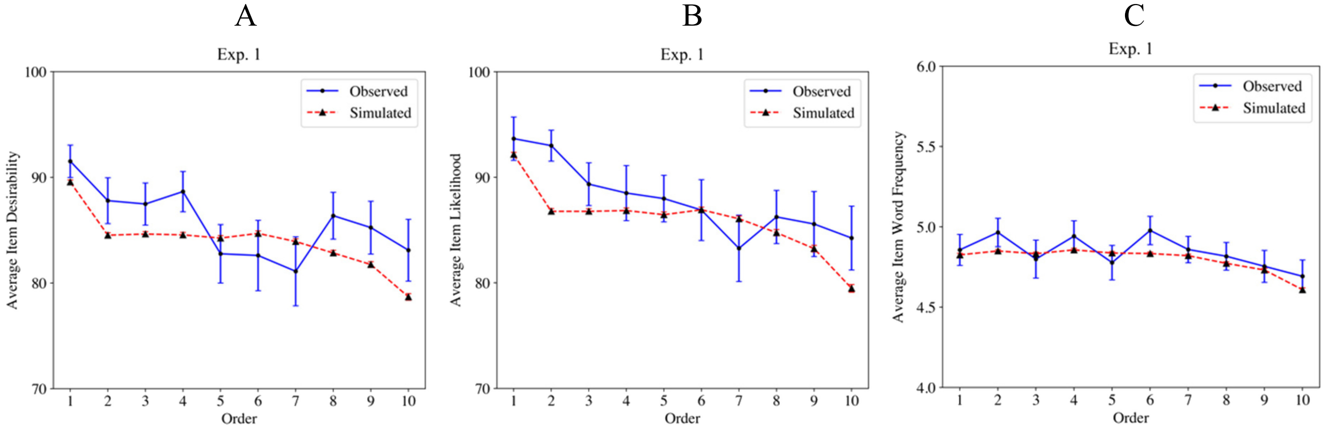
We also examined whether this similarity effect varied as a function of order. We did this by correlating the position at which a counterfactual item was listed with the average cosine similarity between the counterfactual and the target item. This revealed a significant negative relationship, $r_s(8) = -.939$, $p < .001$, 95% CI= [-0.985, -0.757], as shown in Figure 4A. Moreover, each of the 10 generated counterfactuals were more similar with the target item than expected by chance, which is plotted as a dashed line in Figure 4A. Together, these results suggest that counterfactuals that are more similar to the target not only come to mind more frequently but were also generated earlier than counterfactuals that are less similar to the target.

It is important to note that this set of analysis that considered one variable at a time may be susceptible to other confounding variables. In a more rigorous analysis using a memory model that jointly considered different predictors in a single framework, the effect of similarity to target on generation probability became very small. We will return to this issue in the Memory Model section.

Word Frequency. The past two sections have shown that previously documented determinants of counterfactual generation (i.e., item desirability, likelihood, and similarity with the target) also played a role in our task. Now we wish to examine the effects of variables that have been implicated in memory tasks but have not been tested in counterfactual generation. The first of these

Figure 3

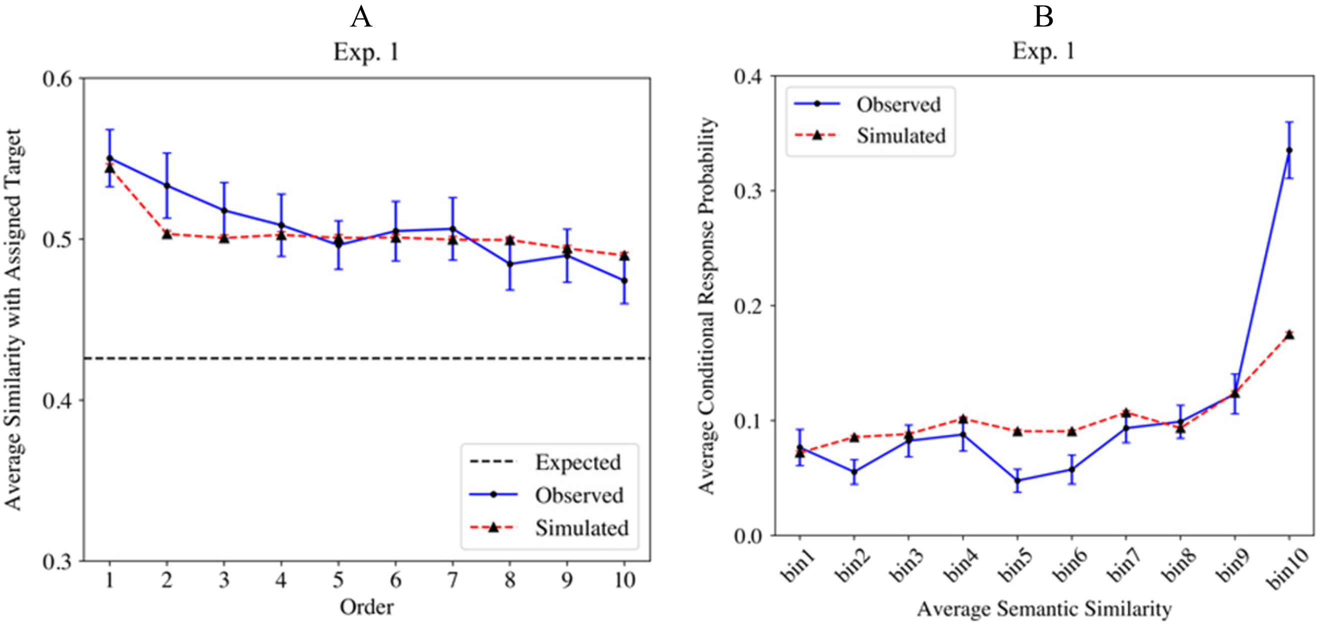
Experiment 1 Observed Versus Predicted Desirability, Likelihood, and Word Frequency



Note. Average observed and model-predicted (A) desirability, (B) likelihood, and (C) word frequency (log-transformed), plotted over the order in which counterfactual items were generated. Error bars display ±1 *SE*. See the online article for the color version of this figure.

Figure 4

Experiment 1 Observed Versus Predicted Target Similarity and CRP



Note. (A) Observed and model-predicted average cosine similarities of counterfactual items with the target item as a function of order. The dashed black line displays the average cosine similarity expected by chance. (B) Observed and model-predicted average conditional response probabilities for 10 semantic similarity bins. Error bars display ±1 *SE*. CRP = conditional recall probabilities. See the online article for the color version of this figure.

variables is word frequency, that is, the frequency with which a word or a concept occurs in a language. Prior work has found that high-frequency words are more likely to be recalled in free recall from lists (Hall, 1954; Lohnas & Kahana, 2013; Sumby, 1963) and in free association tasks (Nelson et al., 2000). This is likely due to the relationship between word frequency and familiarity. People are more likely to be exposed to high-frequency words, and these words, in turn, are likely to have higher baseline activations.

We tested whether high-frequency items were more likely to come to mind in our counterfactual task. We obtained frequency information of each item in our task from the iWeb corpus (https://www.english-corpora.org/iweb/), and we log-transformed the word frequencies for all analyses reported here. By correlating logarithmic word frequency of each item with the probability that the item comes to mind as a counterfactual item in Session 2, we found that higher frequency items also had a higher generation probability, $r(186) = .576$, $p < .001$, 95% CI = [0.472, 0.664], as illustrated in Figure 2C. Note that items belonging to the bin of the highest word frequency are much more likely to come to mind than those from the other bins, resulting in a steep slope in the plot. In contrast, the differences between the last two bins were not as pronounced for desirability (Figure 2A) and likelihood (Figure 2B). As shown in Figure 3C, higher frequency items were also generated somewhat earlier than lower frequency items, although this relationship did not reach significance, $r_s(8) = -.515$, $p = .128$, 95% CI = [−0.864, 0.169].

In addition, we found that word frequency is positively correlated with both desirability, $r(186) = .638$, $p < .001$, 95% CI = [0.545, 0.716], and likelihood, $r(186) = .660$, $p < .001$, 95% CI = [0.571, 0.734]. This is reasonable because the mere exposure to higher frequency words could influence subjective evaluation of desirability and likelihood in the context of this task, and it is also plausible that more desirable and likely fruits and vegetables are used more often in our common language. Overall, our results show that, as in memory tasks like free recall and free association, word frequency is implicated in counterfactual generation.

A related measure of word occurrence in natural language that has been implicated in memory research is contextual diversity, which is the variability of contexts in which a word appears. Previous research has documented the effect of contextual (or semantic) diversity on word learning (Johns et al., 2016) as well as free recall tasks (Lohnas et al., 2011), and studies on word recognition task found that the effect of contextual diversity could override that of word frequency (Adelman et al., 2006). One of the benefits of our modeling approach is that we can simultaneously include word frequency and contextual diversity in the same model and examine whether the effect of one is eliminated in the presence of the other (Chapman & Martin, 2022). We will show that the effect of word frequency is pronounced while controlling for contextual diversity in our counterfactual generation task. To test this, we obtained contextual diversity measures from Brysbaert and New (2009) and were able to find a measure for 129 out of the 188 items in our task. We will return to this issue in the Modeling Results section.

Semantic Clustering. Another variable that has been implicated in memory studies of free recall from lists and free association is semantic similarity with the previously retrieved item (Bousfield & Sedgewick, 1944; Cofer et al., 1966; Gruenewald & Lockhead, 1980; Howard & Kahana, 2002; Romney et al., 1993). As previously retrieved items cue the retrieval of items that are similar to them, this can lead to semantic clustering in the data. To examine whether this effect also emerges in counterfactual generation, we computed conditional response probabilities (CRPs) using the method proposed by Howard and Kahana (2002). Given a previously listed item, CRP specifies the probability that another item comes to mind as a function of the similarities between these two successive items. As in previous analyses, we specified similarity using cosine similarities in the Google News Word2Vec model. We calculated CRP for 10 equally sized bins, with the first bin corresponding to the smallest similarity between two successively generated items (i.e., Bin 1) and the last bin corresponding to the largest similarity between two successively generated items (i.e., Bin 10). If thinking about one counterfactual item leads to thinking about a semantically similar item, then one would observe greater CRPs for higher semantic similarity bins.

As illustrated in Figure 4B, we found that average CRPs increased with bin number when aggregating across participants, $r_s(8) = .745$, $p = .013$, 95% CI = [0.218, 0.935]. The CRPs for the last similarity bins were substantially higher than the average CRPs for the remaining bins, indicating that two successive items are most likely to be highly similar with each other. These results showed that, after one counterfactual item came to mind, the next item that comes to mind was most likely to be semantically related to the previous item.

Memory Model

Model Structure. The above sections have shown the effects of several variables on counterfactual generation. However, the tests for each of these variables have been performed in isolation. To better understand the effect of each variable in the context of other implicated variables, we developed and tested a model that attempts to capture the joint effect of an item's desirability, likelihood, similarity with the target, word frequency, and similarity with the previously generated items, on its probability of being listed as a counterfactual during the generation task. We did this using a Markov random walk, in which the state (i.e., generation) of a participant at time t is a random variable $\chi(t)$. $\chi(t)$ can take on a limited number of distinct values based on the number of possible counterfactuals in the task. Here in Experiment 1, the list of 188 fruits and vegetables serve as the possible distinct states in the model.

We attempted to understand the dynamics at play in counterfactual generation by fitting the Markov random walk to the generated sequences of counterfactual items in our task. For this, we can write out the probability of moving from state i to state j as P_{ij}. We assumed that

$$P_{ij} = \text{softmax}(\beta_D \times \text{DES}_j + \beta_L \times \text{LIK}_j + \beta_F \times \text{FREQ}_j + \beta_T \times \text{SIM}_{jT} + \beta_P \times \text{SIM}_{ij}).$$

That is, P_{ij} is a linear function of item j's desirability (DES_j), likelihood of occurrence (LIK_j), log-transformed word frequency (FREQ_j), similarity with the target (SIM_{jT}), and similarity with the previous item i (SIM_{ij}), passed through a softmax transformation. Here, the β estimates specify the individual effects of these variables and are fit to the data. To specify desirability and likelihood in the model, we used each participant's idiosyncratic Session 1 ratings, and standardized these ratings for each participant. We also standardized the log-transformed word frequencies obtained from the iWeb corpus. Finally, we obtained cosine similarities using the Word2Vec model, and for each participant, we standardized the similarities between each of the 188 items and the target, as well as the similarities between each of the 188 items and the previously listed item for all except the first listed counterfactual item.

By definition, the very first counterfactual that come to mind does not have a previously generated counterfactual item. To model this starting state, we used the above equation but set the similarity between the current and the previous item to 0. We also fit separate parameters to model the effect of the remaining variables on the starting state. By allowing for different parameters for the starting probabilities and the transition probabilities, we make a parsimonious attempt to capture differences in counterfactual generation as a function of order.

We also allowed for revisiting in the model, which means that we did not restrict subsequent states to states that have not been generated previously. We believe that this follows the natural tendency of counterfactual generation and other decision-making processes, as people often mentally revisit previously recalled items or options to make comparisons and construct evaluations. In our task, we explicitly told participants that they are allowed to list the same item multiple times, and we observed that 32.2% of participants listed at least one item twice.

Modeling Fitting. We used hierarchical Bayesian model fitting to estimate the parameters in our model. The hierarchical Bayesian model fitting was carried out in the R interface to Stan (Stan Development Team, 2021). The group-level grand means, μ_k, were set at the standard normal distribution, where k denotes parameters in the model. The individual-level degree of deviation from the grand means, σ_k, was set to follow a half-Cauchy distribution (with location = 0 and scale = 5). The model also allowed each participant's parameters to deviate from the grand mean with different sizes (drawn from a prior standard normal distribution), $\delta_{k,l}$, where l indexes each participant. The individual-level parameter could thus be written as $\beta_{k,l} = \mu_k + \sigma_k\delta_{k,l}$. All group- and individual-level parameters were estimated simultaneously via fitting the individual-level counterfactual generation data. To ensure that Markov chain Monte Carlo samples converge, we ran five independent chains for each fit and estimated the potential scale reduction statistic, $\hat{R}$ (Gelman & Rubin, 1992). Each of the five chains contained 2,000 iterations after 1,000 warmup samples, totaling 10,000 formal samples for each fit. All $\hat{R}$ values were below 1.02, indicating good convergence.

Finally, we performed posterior predictive checks to examine whether our model was able to generate the behavioral patterns originally observed with human participants. Data was simulated for each of the 10,000 samples, but to make it more manageable, we randomly selected 20 distinct samples from each of the five chains, resulting in 100 model-predicted sequences of counterfactual items per participant. The descriptive results from the simulation are summarized in Table 1.

Modeling Results. Hierarchical Bayesian modeling provides both group- and individual-level estimation of each variable in our memory model. Group-level estimation and 95% confidence intervals (95% CIs), the proportion of individual-level 95% CIs above 0, as well as Bayes factors (BFs) are reported in Table 2. To test whether our data favors the inclusion of each variable in our memory model, we individually dropped each of the variables from our

Table 2

Group-Level Estimation and 95% CIs, Proportion of Individual-Level 95% CIs Above 0, and Bayes Factor Between the Full Model and the Nested Model

Variables	β_k	95% CI	Proportion (%)	*BF*
Transition probabilities				
Des.	0.424	[0.207, 0.673]	22	1.76×10^{4}
Lik.	0.719	[0.473, 0.996]	68	5.66×10^{7}
Word Freq.	0.918	[0.785, 1.053]	100	2.13×10^{44}
Tar. Sim.	0.061	[−0.011, 0.127]	0	0.032
Prev. Sim.	0.115	[0.050, 0.176]	49	6.43×10^{1}
Starting probabilities				
Des.	0.828	[0.160, 1.630]	0	2.13×10^{1}
Lik.	1.655	[0.819, 2.687]	73	2.99×10^{3}
Word Freq.	0.716	[0.327, 1.131]	12	8.55×10^{2}
Tar. Sim.	0.219	[0.052, 0.373]	0	0.033

Note. CI = confidence interval; BF = Bayes factor; Des. = desirability; Lik. = likelihood of occurrence; Word Freq. = log-transformed word frequency; Tar. Sim. = similarity with the target; Prev. Sim. = similarity with the previous item.

model and computed BFs between the full model and each of the nested models. We interpret BF > 1 as showing support for the full model over the nested model, and BF < 1 indicating that the data support the nested model over the full model (Lee & Wagenmakers, 2014). In our data, the desirability and the likelihood of eating were highly correlated. Despite so, the hierarchical Bayesian model fitting should allow us to produce precise parameter estimates for both variables (Jaya et al., 2019).

On the group level, an item's subjective desirability and likelihood of occurrence both positively contribute to the probability that it comes to mind as a counterfactual item (all BF > 10). Our model was able to accurately capture the desirability effect observed in the empirical data by closely mimicking the relationship between item desirability and probability of generation (Figure 1A), as well as the relationship between desirability and order of generation (Figure 2A). Repeating prior statistical tests with our model's predictions, rather than participant data, reveals that the participants' observed desirability of items was positively correlated with the model's predicted probability of counterfactual generation, $r(186) = .759$, $p < .001$, 95% CI = [0.690, 0.813], and negatively correlated with the order of generation, $r_s(8) = -.842$, $p = .002$, 95% CI = [−0.961, −0.452]. Our model was also able to accurately capture the effect of likelihood on generation probability (Figure 1B) and order (Figure 2B). Once again, repeating the above statistical tests with our model's predictions rather than participant data reveals that the observed likelihood of items was positively correlated with the simulated generation probability, $r(186) = .766$, $p < .001$, 95% CI = [0.699, 0.819], and negatively correlated with the order of the simulated counterfactuals, $r_s(8) = -.863$, $p = .001$, 95% CI = [−0.961, −0.452]. On the individual level, most participants reliably displayed similar tendencies for likelihood but not desirability, as indicated by the proportion of participants whose individual estimates above the random level. Nevertheless, the BFs indicate a strong support for the inclusion of both desirability and likelihood in the model.

Table 2 shows that, with other variables statistically controlled in the model, there was only a weak effect of target similarity on counterfactual generation and the 95% credible interval (i.e., the Bayesian confidence interval) contained the null effect (i.e., zero). Furthermore, we did not find support for the inclusion of target similarity in the model (both BF < 1). Despite the null effect, our model was able to mimic the empirical patterns as shown in Figure 4A, and the cosine similarities between the model's generated items and the target items were higher than the expected similarity between two randomly selected items, $t(58999) = 122.22$, $p < .001$, 95% CI = [0.502, 0.505]) and it dropped as a function of order, $r_s(8) = -.915$, $p < .001$, 95% CI = [−0.980, −0.674]. This indicates that the effect of target similarity documented in Figure 4A can be explained by other variables, such as desirability. This would not be surprising, since other fruits and vegetables that are similar to the actual fruit or vegetable consumed by our participant, would also be desirable fruits and vegetables for the participant. To disentangle the effects of desirability and target similarity we randomly assigned targets to participants in Experiments 3 and 4.

Finally, word frequency and semantic clustering positively contribute to counterfactual generation. As expected from the observed empirical patterns, word frequency has a robust effect in our model. Moreover, our model replicated the word frequency effects observed earlier in this article. As shown in Figure 1C, an item's word frequency was positively correlated with its model-predicted probability of becoming a counterfactual, $r(186) = .701$, $p < .001$, 95% CI = [0.680, 0.807], and as shown in Figure 2C, word frequency is negatively correlated with the order in which an item comes to mind, $r_s(8) = -.721$, $p = .019$, 95% CI = [−0.928, −0.168].

To examine whether the effect of word frequency is confounded by contextual diversity, we ran our model with two additional variables: (a) contextual diversity for starting probabilities, and (b) contextual diversity for transitional probabilities. As noted previously, we only found contextual diversity measures for 129 out of the 188 items, so we excluded any counterfactual items that participants listed which were not one of those 129 items (this excluded 6.34% of the observations we used to fit our reported model, which is reasonable because items without a contextual diversity measure were also less commonly seen in the United States). For model fitting, we also standardized the contextual diversity measures for these 129 items. Then, we individually dropped each of the word frequency and contextual diversity variables from this new model and computed BFs between the full model and each of the nested models. We found no evidence for the effect of contextual diversity on transitional probabilities (BF < 0.001) nor on starting probabilities (BF < 0.001). In contrast, we found significant evidence for the effect of word frequency on transitional probabilities (BF = 7.789×10^{32}) but not on starting probabilities (BF < 0.001). Overall, these results suggest that the effect of word frequency, rather than contextual diversity, influences the generation of counterfactual items in our task.

Additionally, the model revealed that counterfactual items that are more similar with previously listed items are more likely to come to mind (BF = 6.43×10^{1}). Model-simulated counterfactual items are clustered semantically such that the average conditional recall probabilities (CRP) are positively correlated with bin number, $r_s(8) = .891$, $p < .001$, 95% CI = [0.596, 0.974]. Note that the model underpredicted the last CRP bin in Figure 4B. This could be due to several reasons, including our modeling assumptions. For simplicity, our memory model assumed a linear effect of semantic clustering, and we suspect that better predictions would be obtained with more complex, nonlinear functions.

Effects of Counterfactuals on Target Evaluation

Although our task was designed to study counterfactual generation, we also tested the effect of generated counterfactuals on target evaluation. Prior studies have found that desirable counterfactuals reduce the evaluations of targets, whereas undesirable counterfactuals increase the evaluations of targets (e.g., Mellers et al., 1997). Similarly, when choosing between two options with unknown rewards, people infer that the observed value of their chosen option is inversely related to the unobserved value of the unchosen option (Biderman & Shohamy, 2021). In our experiment, people generated their own counterfactual outcomes rather than making forced choice between given options, but it is possible that the perceived desirability of the target outcome changes as an inverse function of the average desirability of the counterfactual outcomes.

To test this, we compared the target item's desirability ratings in Session 2 (which was made in the context of the counterfactual items that came to mind) with the target's desirability ratings in Session 1 (which was made in the context of the full set of items that could be listed as counterfactuals). We expected that the difference between an individual's Session 2 and Session 1 ratings for the target would be a negative function of the average Session 1 ratings of the counterfactuals generated by that individual in Session 2. That is, if we write individual *i*'s rating of the target in Session 1 as r_{i1}, their rating of the target in Session 2 as r_{i2}, and their average Session 1 rating of the counterfactuals generated in Session 2 as r_{ic}, then we can test our hypothesis with the regression $d_i \sim \beta_0 + \beta_1 \cdot r_{ic}$, where $d_i = r_{i2} – r_{i1}$. We found that the average desirability ratings of the counterfactual items did not have a significant effect on the target item's desirability rating for any of the three experiments ($\beta_1 = -0.024$, $p = .871$, 95% CI = [−0.318, 0.270]). We also attempted a variant of this test in which we *z*-scored participants' ratings (within each participant) to control for participant heterogeneity in how they use the rating scale; however, our null results persisted ($\beta_1 = -0.229$, $p = .178$, 95% CI = [−0.565, 0.108]).

Our last attempt was to compare the cross-session change in average desirability separately for the target item and the counterfactuals. On the one hand, participants' average desirability of the target item was 75.4 in Session 2 but 90.0 in Session 1, although this difference did not reach significant as revealed by a paired *t* test, $t(58) = -1.695$, $p = .096$, 95% CI = [−31.720, 2.636]. On the other hand, participants' average desirability rating of the counterfactuals was 79.9 in Session 2 but 85.6 in Session 1, and this difference was significant, $t(567) = -7.237$, $p < .001$, 95% CI = [−6.397, −3.666]. When we *z*-scored participants' ratings (within each participant), there was no difference between the average desirability of the target item (1.317 in Session 2 vs. 1.336 in Session 1), $t(58) = -0.275$, $p = .784$, 95% CI = [−0.154, 0.117], but the average desirability of the counterfactuals were significantly lower in Session 2 compared to Session 1 (0.885 in Session 2 vs. 1.116 in Session 1), $t(567) = -0.201$, $p < .001$, 95% CI = [−0.255, −0.146]. This indicates that there are session-level effects on how the scale is used, but that these do not influence the relative rating of the target in the contexts of the counterfactuals. We suspect that the Session 2 ratings may have changed in such a way as to make them consistent with the just-produced judgments.

We suspect that the null result could be a product of our task, which was designed to study dynamics of counterfactual generation and not the effect of generated items on target evaluation. In particular, Session 1 ratings are a good proxy of the desirability (and likelihood of occurrence) of the counterfactual items and are useful inputs into our memory model. However, these ratings may not provide a decontextualized measure of the target's baseline desirability. In fact, it could be the case that participants engaged in counterfactual thinking and generated a similar set of counterfactuals when rating the target in Session 1 as they did when rating the target in Session 2. Thus, comparing the target's Session 2 desirability with its Session 1 desirability can fail to measure the effect of the counterfactuals.

Discussion

In Experiment 1, we have examined the determinants of counterfactual generation using quantitative modeling. First, we have replicated prior research findings on the positive effects of an item's desirability, likelihood of occurrence, and similarity to the target, by studying each variable in isolation. Second, we have shown that established memory effects, notably the word frequency effect and the semantic clustering effect, also play a role in counterfactual generation. Third, we have quantitatively predicted the empirical patterns by fitting a computational memory model to observed data. One of the benefits of quantitative model fitting is that we can determine the relative strength of the effects of various mechanisms on counterfactual generation. Our model revealed that desirability and likelihood of occurrence are strong drivers of counterfactual generation. The more desirable an outcome is perceived and more likely that it could have occurred, the more probable that it comes to mind as a counterfactual outcome, even after controlling for other mechanisms. However, we did not find support for the inclusion of target similarity in the model, which we suspect could be due to the fact that target similarity is confounded with other variables like desirability in our task (in Experiments 3 and 4, we tackle this problem by randomly assigning targets across participants). Finally, our modeling results suggest that the memory effects of word frequency and semantic clustering are also implicated in counterfactual thinking. This implies that accounting for memory effects might enhance our comprehension of what drives counterfactual generation.

Experiment 2

Experiment 2 aims to modify the set of counterfactual outcomes generated by participants by varying the number of counterfactuals they are prompted to consider. We have found in Experiment 1 that the effects of desirability and likelihood were more pronounced at the start of the generation process, and their effects weaken as more counterfactual outcomes were generated. To test whether considering fewer versus more counterfactual outcomes alters the overall desirability and likelihood of the counterfactuals that come to mind, we adapted the task in Experiment 1 and instead asked participants to generate either five or 20 counterfactuals. As our model has shown promise in accurately representing the relationships between order and each of these variables, it should also be able to demonstrate the disparities between the two lengths of generation in this experiment.

Method

Experiment 2 was preregistered at https://osf.io/9ymhe. Participants ($N = 159$; $M_{\text{age}} = 40.0$; 55% female, 42% male, 3% nonbinary) were recruited from Prolific Academic and performed the experiment online using their own computer interface. Participation was limited to native English speakers in the United States. The design used in this experiment was identical to that in Experiment 1 except for the number of counterfactuals that participants were asked to generate. While participants in Experiment 1 were asked to list 10 counterfactuals that come to mind as they considered the target, half of the participants in Experiment 2 were asked to list five counterfactuals, and the other half were asked to list 20 counterfactuals, with random assignment of conditions.

We excluded one participant who gave the same likelihood rating for all items in Session 1. Occasionally in Session 2, participants listed items that are among the 188 fruits or vegetables in our list. As in Experiment 1, we excluded 11 participants who listed a target item that was not one of the 188 fruits or vegetables in our list, and eight additional participants who listed as counterfactuals more than 50% of such items (due to oversight, this criterion was not preregistered; we show in the Appendices A and B that the results hold in the presence of these participants). Among the remaining participants, some of the counterfactual items they listed were not among those 188 items and these items were thus excluded from further analyses (4.10% of all listed counterfactuals in the List 5 condition, and 5.13% in the List 20 condition). As in Experiment 1, we also excluded extra items that were listed on the same screen beyond the very first item, rather than excluding participants who have mistaken the instructions.

Overall, participants in the List 5 condition generated 29 different target items and 65 different counterfactual items. The most common target items were "apple" (10 times) and "broccoli" (6 times). The most common counterfactual items were "banana" (26 times) and "tomato" (22 times). Participants in the List 20 condition generated 31 different target items and 113 different counterfactual items. The most common target items were "broccoli" (8 times) and "corn" (5 times). The most common counterfactual items were "apple" (65 times) and "carrot" (57 times).

Results

Replicating Experiment 1 With Observed Data

Desirability and Likelihood. Experiment 2 replicated the effects of desirability and likelihood of occurrence on counterfactual generation. Across both conditions, an item's desirability ratings elicited in Session 1 was positively correlated with the probability that it gets listed as a counterfactual in Session 2, $r(186) = .635$, $p < .001$, 95% CI = [0.541, 0.713] in the List 5 condition; $r(186) = .745$, $p < .001$, 95% CI = [0.674, 0.803] in the List 20 condition (see Table 1 for a summary of the descriptive results). Similarly, item likelihood also positively correlated with generation probability, $r(186) = .626$, $p < .001$, 95% CI = [0.530, 0.706] in the List 5 condition; $r(186) = .710$, $p < .001$, 95% CI = [0.630, 0.774] in the List 20 condition. These observations are shown in Figure 5A and 5B, respectively.

We were also able to replicate the empirical patterns of desirability and likelihood on the order in which counterfactual items were generated. As shown in Figure 6A and 6B, counterfactuals that were generated earlier were also on average more desirable, $r_s(3) = -.700$, $p = .188$, 95% CI = [−0.978, 0.476] in the List 5 condition; $r_s(18) = -.764$, $p < .001$, 95% CI = [−0.901, −0.486] in the List 20 condition, and more likely to occur, $r_s(3) = -.900$, $p = .037$, 95% CI = [−0.993, −0.087] in the List 5 condition; $r_s(18) = -.817$, $p < .001$, 95% CI = [−0.925, −0.587] in the List 20 condition, relative to counterfactuals generated later.

Similarity With the Target. Experiment 2 also replicated the empirical pattern of target similarity. Overall, participants generated counterfactuals that were more similar to the target item than expected by chance using one sample *t* tests, $t(397) = 10.671$, $p < .001$, 95% CI = [0.478, 0.501] in the List 5 condition; $t(1441) = 15.937$, $p < .001$, 95% CI = [0.474, 0.488] in the List 20 condition. No significant correlation was found between target similarity and order as illustrated in Figure 6C, $r_s(3) = -.500$, $p = .391$, 95% CI = [−0.959, 0.684] in the List 5 condition; $r_s(18) = -.281$, $p = .230$, 95% CI = [−0.643, 0.184] in the List 20 condition.

Word Frequency. Consistent with Experiment 2, the probability that an item comes to mind was positively correlated with its word frequency, $r(186) = .551$, $p < .001$, 95% CI = [0.443, 0.643] in the List 5 condition; $r(186) = .595$, $p < .001$, 95% CI = [0.494, 0.680] in the List 20 condition. In contrast, word frequency is negatively correlated with order, $r_s(3) = -1.000$, $p < .001$, 95% CI = [−1.000, −1.000] in the List 5 condition; $r_s(18) = -.756$, $p < .001$, 95% CI = [−0.898, −0.472] in the List 20 condition. These results are shown in Figures 5C and 6D, respectively.

Semantic Clustering. The semantic clustering effect also replicated using the CRP analysis introduced in Experiment 1. In both conditions, CRPs were positively correlated with bin number when aggregated across participants, $r_s(3) = .796$, $p = .006$, 95% CI = [0.334, 0.949] in the List 5 condition; $r_s(18) = .794$, $p = .006$, 95% CI = [0.329, 0.949] in the List 20 condition. This indicates that participants were more likely to think about counterfactuals that are similar to what has previously come to mind.

Effect of Length on Counterfactual Generation

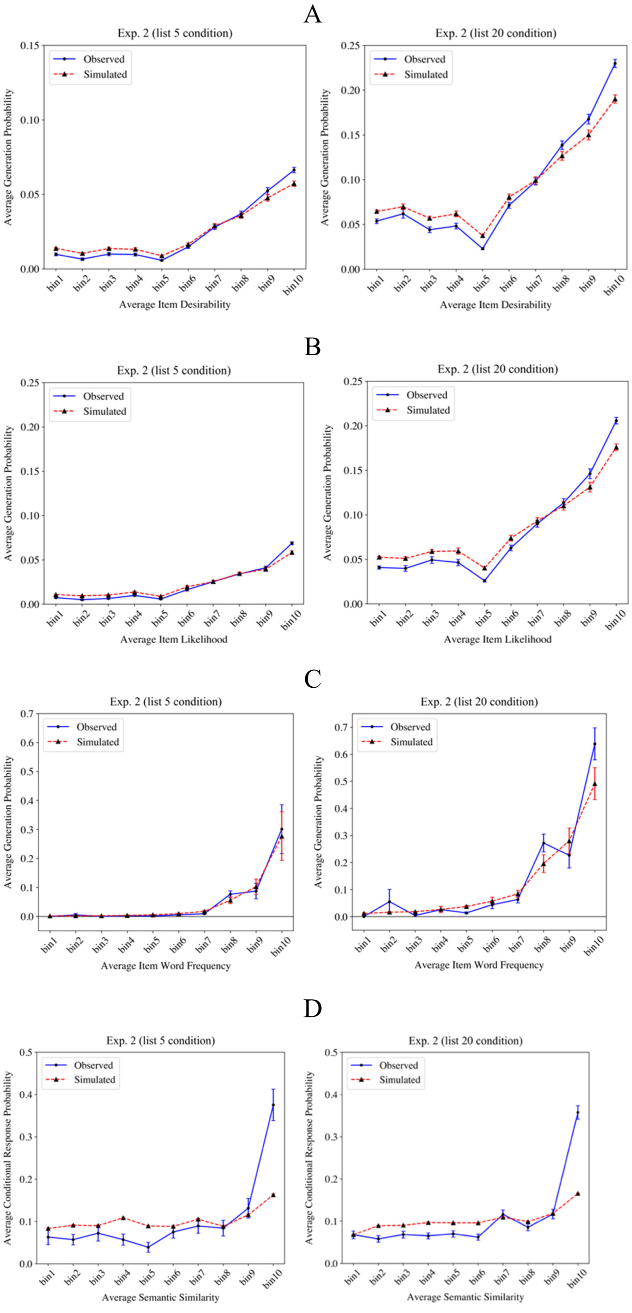
The main goal of our experiment was to manipulate counterfactual outcomes by varying the number of counterfactuals that participants were asked to generate. We predicted that participants in the List 5 condition would list counterfactuals that are perceived as more desirable and more likely to occur than those in the List 20 condition. We found that these predictions were supported. Overall, the average desirability of the counterfactuals in the List 5 condition was 84.7, compared to 80.7 in the List 20 condition. This difference is statistically significant with an unpaired *t* test, $t(1838) = 3.116$, $p = .002$, 95% CI = [1.479, 6.504]. Similarly, a simple linear regression of the Session 1 desirability ratings of counterfactuals on condition (with the List 5 condition coded as 1 and the List 20 condition coded as 0) revealed that participants generated significantly more desirable items in the List 5 condition than in the List 20 condition ($\beta_1 = 3.992$, $p = .002$, 95% CI = [1.481, 6.502]). This difference is illustrated in Figure 6A. Note that, on average, the generation probabilities were naturally high in the List 20 condition because participants had more opportunities to generate counterfactuals.

We performed similar tests for likelihood of occurrence and found the same pattern. As illustrated in Figure 6B, the average likelihood of occurrence of counterfactuals was 86.4 in the List 5 condition but 82.9 in the List 20 condition, and this difference is significant with an unpaired *t* test, $t(1838) = 2.480$, $p = .013$, 95% CI = [0.742, 6.359], and a regression of likelihood ratings on condition ($\beta_1 = 3.550$, $p = .013$, 95% CI = [0.744, 6.357]).

In addition, Figure 6C revealed that participants were more likely to generated counterfactuals that are more similar with the target when they were asked to list five instead of 20 items. The average target similarity of the generated counterfactuals was 0.489 in the List 5

Figure 5

Experiment 2 Observed Versus Predicted Generation Probabilities and CRP



Note. Average observed and model-predicted probabilities of counterfactual generation as a function of item (A) desirability decile, (B) likelihood decile, and (C) word frequency (log-transformed) decile. (D) Observed and model-predicted average conditional response probabilities for 10 semantic similarity bins. Error bars display ±1 *SE*. CRP = conditional recall probabilities. See the online article for the color version of this figure.

condition but 0.481 in the List 20 condition, although this difference was not significant, $t(1838) = 1.158$, $p = .247$, 95% CI = [−0.006, 0.023]; $\beta_1 = 0.008$, $p = .247$, 95% CI = [−0.006, 0.023].

Furthermore, although the correlation between word frequency and order of generation did not reach significance in Experiment 1, we found a significant difference between the average word frequency of the generated counterfactuals in the two conditions tested (4.97 in the List 5 condition but 4.69 in the List 20 condition), $t(1838) = 6.554$, $p < .001$, 95% CI = [0.196, 0.363]; $\beta_1 = 0.279$, $p < .001$, 95% CI = [0.196, 0.363]. This difference is illustrated in Figure 6D.

Overall, these results show that our manipulation was able to successfully alter the set of generated counterfactuals. On average, participants listed items that were more desirable, more likely, and more similar with the target when they were asked to generate fewer items.

Memory Model

We fit the same model from Experiment 1 separately for each condition here. As before, desirability ratings, likelihood ratings, similarities with the target, similarities with the previous item, and log-transformed word frequencies were standardized. In both conditions, we observed instances of revisiting (10.8% of participants in the List 5 condition list the same counterfactual item at least twice, compared to 63.2% in the List 20 condition). The lower revisiting rate in the List 5 condition is reasonable as participants were allowed fewer opportunities to generate (repeated) counterfactuals compared to the List 20 condition. Model fit and posterior predictive checks were performed using the same approach as in Experiment 2B. All $\hat{R}$ values were below 1.02. Hierarchical Bayesian modeling, with the same assumptions as described earlier, was used to fit the data, and posterior predictive checks were applied using the same approach as previously specified. The simulated patterns are shown in Table 1.

Replicating Experiment 1 With Model Predictions

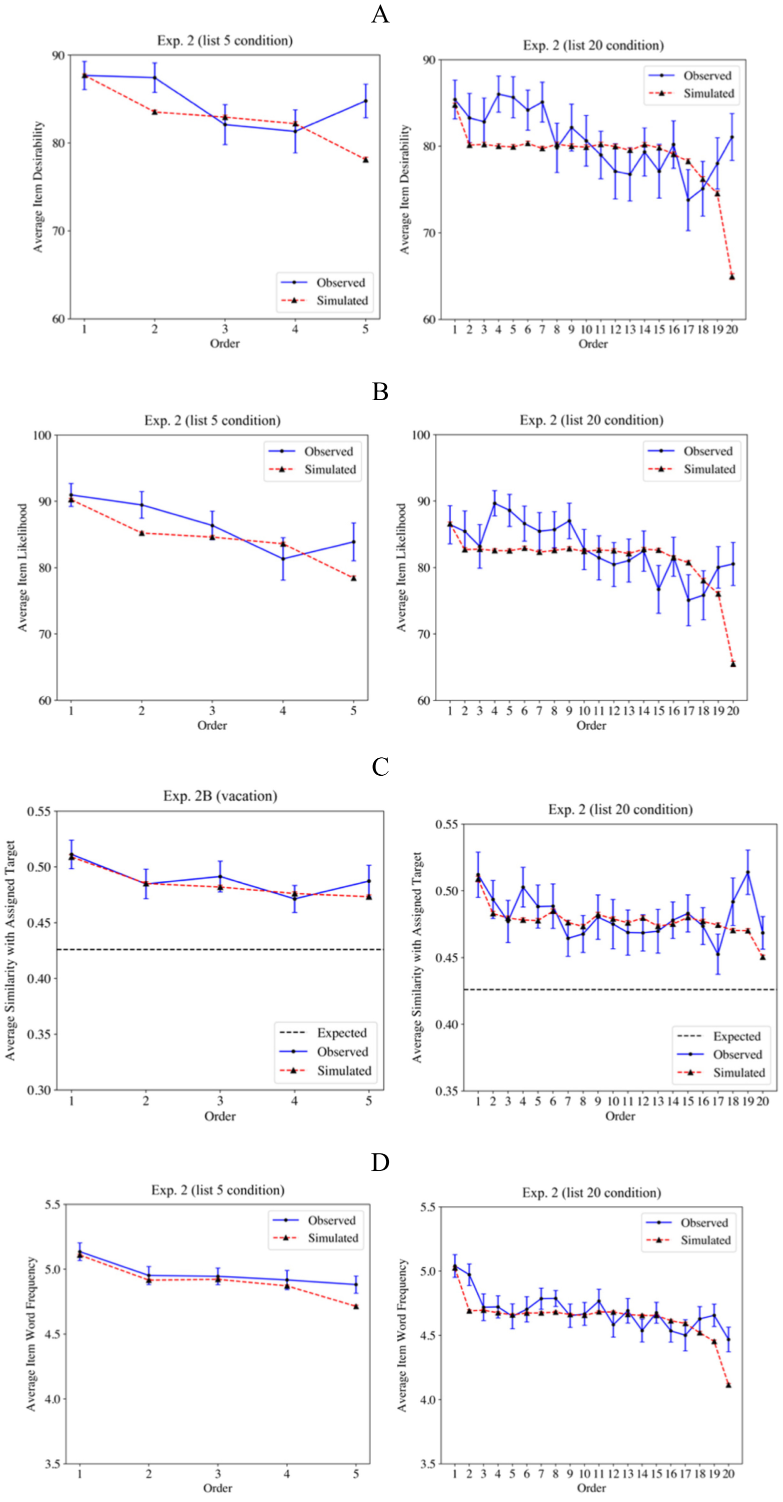
We began by replicating the main results of Experiment 1. We found that the effects of desirability and likelihood were robust in both conditions. Our model successfully predicted that each item's model-predicted generation probability is positively associated with its observed desirability, $r(186) = .690$, $p < .001$, 95% CI = [0.607, 0.758] in the List 5 condition; $r(186) = .820$, $p < .001$, 95% CI = [0.767, 0.862] in the List 20 condition, as well as its observed likelihood, $r(186) = .683$, $p < .001$, 95% CI = [0.599, 0.752] in the List 5 condition; $r(186) = .813$, $p < .001$, 95% CI = [0.758, 0.856] in the List 20 condition.

In addition, the model generated counterfactuals that were more similar with the target item than random chance, $t(41499) = 77.614$, $p < .001$, 95% CI = [0.484, 0.487] in the List 5 condition; $t(151999) = 141.190$, $p < .001$, 95% CI = [0.477, 0.578] in the List 20 condition.

Moreover, the model effectively accounted for the effect of word frequency on generation probability, $r(186) = .713$, $p < .001$, 95% CI = [0.635, 0.777] in the List 5 condition; $r(186) = .799$, $p < .001$, 95% CI = [0.741, 0.845] in the List 20 condition. The model underpredicted the influence of semantic clustering, $r_s(3) = .527$, $p = .117$, 95% CI = [−0.154, 0.868] in the List 5 condition; $r_s(18) = .939$, $p < .001$, 95% CI = [0.851, 0.976] in the List 20 condition. Again, we suspect that this could imply a nonlinear effect.

Figure 6

Experiment 2 Observed Versus Predicted Desirability, Likelihood, Target Similarity, and Word Frequency



Note. (A) Average observed and model-predicted item desirability as a function of order. (B) Average observed and model-predicted item likelihoods as a function of order. (C) Average observed and model-predicted cosine similarity with the assigned target as a function of order. (D) Average observed and model-predicted log-transformed word frequency as a function of order. Error bars display ±1 *SE*. See the online article for the color version of this figure.

Model Predictions of the Length Effect

The model was able to successfully predict differences across the conditions as shown in Figure 12A–12D. First, the average desirability of the counterfactuals in the five-length simulations was 82.9, compared to 78.9 in the 20-length simulations, and this difference is statistically significant, $t(193498) = 31.24$, $p < .001$, 95% CI = [3.748, 4.249]; $\beta_1 = 4.000$, $p < .001$, 95% CI = [3.748, 4.249]. The average likelihood of occurrence of our model generated counterfactuals was 84.4 in the List 5 condition but 81.2 in the List 20 condition, and this difference is significant, $t(193498) = 22.233$, $p < .001$, 95% CI = [2.909, 3.472]; $\beta_1 = 3.191$, $p < .001$, 95% CI = [2.909, 3,472]. These contrasts are illustrated in Figure 12A and 12B, respectively. In addition, the average similarity with model-predicted counterfactuals and the participant-generated target item was 0.485 in the List 5 condition but only 0.477 in the List 20 condition, and this difference is significant, $t(193498) = 9.442$, $p < .001$, 95% CI = [0.006, 0.009]; $\beta_1 = 0.008$, $p < .001$, 95% CI = [0.006, 0.009]. This relationship is shown in Figure 6C. Finally, the average word frequency of the model-predicted counterfactuals was 4.91 in the List 5 condition but 4.64 in the List 20 condition, and this difference is significant, $t(193498) = 59.745$, $p < .001$, 95% CI = [0.261, 0.279]; $\beta_1 = 0.270$, $p < .001$, 95% CI = [0.261, 0.279]. Overall, these results show that our model was able to successfully predict the effect of length.

Our model was less adequate at capturing the gradual reduction in the desirability, likelihood, and target similarity of the generated counterfactuals in the List 20 condition. This is because the model only allows for differential effects of these variables for the first generated counterfactual and assumes that the same memory parameters govern how subsequent counterfactuals are generated. In other words, the model predicts that the strengths of the effects do not change after the second counterfactual outcome. In addition, our model expected substantially more instances of revisiting in the 20-length simulations (92.0%) than the observed 20-length data (63.2%). However, this was not the case for the five-length simulations (15.2%) compared to the observed five-length data (10.8%). This is because the number of items that could get listed as counterfactuals is limited (i.e., 188 fruits and vegetables) and participants rated a narrow subset of items as highly desirable or highly likely to occur, which drive the model to yield more repetitions for the longer generation lengths. With more repetitions among the generated counterfactuals, the effects of the variables are more likely to remain the same over the order of generation.

Discussion

Experiment 2 aimed to manipulate the total number of counterfactuals that participants were asked to consider. Building on the results of Experiment 1, we hypothesized that participants prompted to produce only five outcomes would list counterfactuals that are perceived as more desirable and more likely to occur than those who were asked to produce 20 outcomes. Our analyses confirmed these predictions, with participants in the List 5 condition generating outcomes that were rated higher in desirability and likelihood than those in the List 20 condition. Importantly, our computational model was able to predict these differences. Furthermore, we successfully replicated the primary findings of Experiment 1 using a distinct cohort of participants.

It is worth noting that our computational model exhibited a tendency to underpredict the effects of later counterfactual items in comparison to earlier ones. This is likely due to a sizeable proportion of participants generating at least one invalid or implausible item as part of their list of counterfactual outcomes—14 participants in the List 5 condition and 44 in the List 20 condition. As a result, there was a reduced amount of data available to the model later in the sequence of generated counterfactuals, which may have limited its ability to accurately calibrate and predict the effects of later items in the series. We show in the Appendices A and B section that excluding participants who listed any invalid items further improves the model's ability to mimic the observed data. Moreover, these invalid items typically include food items other than fruits or vegetables, as well as nonfood items, which indicates a lack of attention. Despite so, the model performed fairly well in the presence of these invalid items, as shown earlier.

Experiments 3A–3C

Experiments 1 and 2 have demonstrated the utility of formal computational models for jointly studying the effects of various mechanisms on counterfactual generation, including the roles of established memory processes and the length of generation. However, the target items that initiated counterfactual thinking were actual items that were recalled by the participants. We did not manipulate these items through random assignment. This could be one reason why we did not observe the target similarity effect in our model fits for Study 1.

In Experiments 3A–3C, we attempted to address this shortcoming using three different hypothetical scenarios. All experiments followed the procedure in Experiment 1, except that we randomly assigned a target item to each participant. Experiment 3A asked participants to evaluate a job offer in a foreign country and list counterfactual countries that came to their minds as they thought about the target country. Experiment 3B asked participants to evaluate a vacation in a foreign country and list counterfactual countries that came to their minds as they thought about the target country. Finally, Experiment 3C asked participants to evaluate a food tasting of a fruit or a vegetable and list the counterfactual fruits or vegetables that came to their minds as they thought about the target fruit or vegetable.

Method

Participants

Experiment 3A was preregistered at OSF.[1] Participants in Experiment 3A ($N = 53$; $M_{\text{age}} = 20$; 55% female, 43% male, 2% nonbinary), Experiment 3B ($N = 53$; $M_{\text{age}} = 32$; 57% female, 37% male, 6% nonbinary), and Experiment 3C ($N = 40$; $M_{\text{age}} = 20$; 67% female, 33% male) performed the experiment online using their own computer interface. Participants in Experiment 3A and

[1] Note that due to the Covid outbreak, and the resulting March 2020 lockdown, we were unable to collect our prespecified number of samples for Experiment 3A. However, we found that the 53 participants were sufficient for modeling memory effects in counterfactual generation (as each participant recalled 10 items, we had a total of 530 observations which gave us sufficient power for our memory models). We decided to retain this sample size for subsequent experiments.

3C were undergraduate students at the University of Pennsylvania and participated in the study for course credit. Participants in Experiment 3B were recruited via Prolific Academic, and participation was limited to native English speakers who are citizens of the United States.

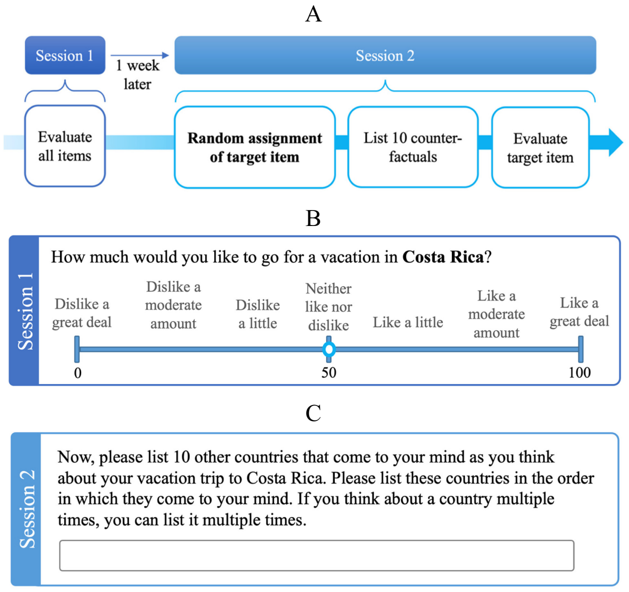
Procedures

As in Experiments 1 and 2, Experiments 3A–3C had two sessions (Figure 7A). In Session 1, participants evaluated a comprehensive list of items (countries in Experiments 3A and 3B, and fruits and vegetables in Experiment 3C). For example, participants in Experiment 3B were asked to rate how much they would like to go on a vacation in each of the 193 countries in the world (Figure 7B).

A week later, in Session 2, participants were first shown a description of a hypothetical event in which they were asked to imagine obtaining a target outcome (job offer in a country in Experiment 3A, vacation trip to a country in Experiment 3B, and food tasting of a piece of fruit or vegetable in Experiment 3C). Next, participants were asked to list 10 counterfactual outcomes that came to their minds as they considered their assigned target outcome. For example, some participants in Experiment 3B were told that they had won a vacation to Costa Rica and were then asked to list the other countries that came to their minds as they considered their Costa Rica vacation (Figure 7C). Participants were asked to list the counterfactual items on 10 successive screens. Finally, participants rated the target outcome in terms of desirability and likelihood on one screen, and they also rated their listed counterfactual outcomes in terms of desirability and likelihood on another screen.

Figure 7

Experiments 3A–3C Design and Prompt Examples



How much would you like to go for a vacation in **Costa Rica**?

Now, please list 10 other countries that come to your mind as you think about your vacation trip to Costa Rica. Please list these countries in the order in which they come to your mind. If you think about a country multiple times, you can list it multiple times.

Note. (A) Schematic of the task design for Experiments 3A–3C. (B) Example of the desirability rating question in Session 1 of Experiment 3B. (C) Example of the counterfactual generation task in Session 2 of Experiment 3B. See the online article for the color version of this figure.

Stimuli

For Experiments 3A (job offer) and 3B (vacation), we created a comprehensive list of all countries in the world using the 193 member states of the United Nations (as of February 7, 2020). For Experiment 3C (food tasting), we used the same list of 188 fruits and vegetables as in Experiment 1. Out of each list, we selected four items (i.e., country or fruits and vegetables) to use as target outcomes for each experiment. To ensure the robustness of our results, we attempted to select target items that were as dissimilar to each other as possible. For this purpose, we applied multidimensional scaling on the Word2Vec representations of each item to visualize all the items on a two-dimensional space. By inspecting this visualization, we selected four target outcomes from different clusters that emerged in this space. Our target outcomes for Experiment 3A (job offer) were Germany, Kenya, Guatemala, and Saudi Arabia. The target outcomes for Experiment 3B (vacation) were France, Costa Rica, Japan, and South Africa. For Experiment 3C (food tasting), we selected strawberry, passionfruit, collard, and zucchini as the target outcomes.

Occasionally, participants listed counterfactuals in Session 2 that had not been evaluated in Session 1 (e.g., some participants listed names of cities instead of countries). We excluded these counterfactuals from our data analyses (except when participants listed the capital cities of countries, in which case we counted them as their corresponding countries). Overall, 1.32% of all listed counterfactuals were dropped in Experiment 3A, 5.85% were dropped in Experiment 3B, and 8.25% were dropped in Experiment 3C. As in previous experiments, we excluded additional items that were listed on the same screen beyond the very first item, rather than excluding participants who have mistaken the instructions.

For Experiment 3A (job offer), our undergraduate participants generated 101 different counterfactual countries, and the most common ones were "United Kingdom" (32 times) and "France" (32 times). For Experiment 3B (vacation), our Prolific participants generated 93 different counterfactuals, and the most common ones were "France" (26 times), "Spain" (25 times), and "United Kingdom" (25 times). For Experiment 3C (food tasting), our undergraduate participants generated 81 different counterfactual fruits and vegetables, and the most frequent items were "cabbage" (19 times), "apple" (15 times), and "lettuce" (15 times).

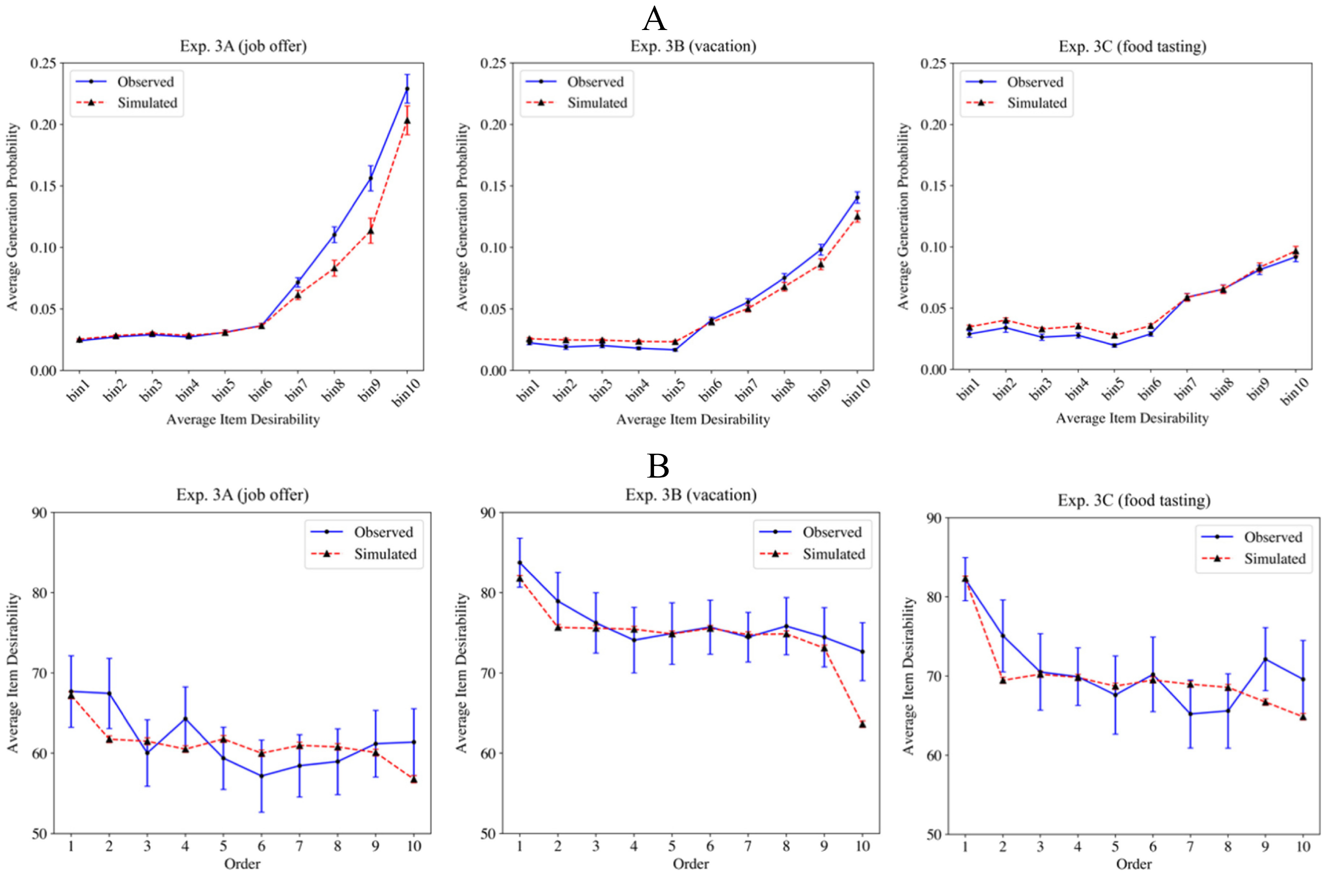
Results

Replicating Experiment 1 With Observed Data

Desirability and Likelihood. Experiments 3A–3C replicated the effects of desirability and likelihood of occurrence on counterfactual generation using hypothetical scenarios. Across the three experiments, the probability that an item gets listed as a counterfactual was positively correlated with its desirability, $r(191) = .677$, $p < .001$, 95% CI = [0.592, 0.747] in Experiment 3A; $r(191) = .652$, $p < .001$, 95% CI = [0.563, 0.727] in Experiment 3B; $r(186) = .492$, $p < .001$, 95% CI = [0.375, 0.593] in Experiment 3C, and likelihood, $r(191) = .805$, $p < .001$, 95% CI = [0.748, 0.849] in Experiment 3A; $r(191) = .772$, $p < .001$, 95% CI = [0.708, 0.824] in Experiment 3B; $r(186) = .538$, $p < .001$, 95% CI = [0.428, 0.633] in Experiment 3C (see Table 1). These observations are illustrated in Figure 8A.

Figure 8

Experiments 3A–3C Observed Versus Predicted Generation Probabilities and Desirability



Note. (A) Average observed and model-predicted counterfactual generation as a function of desirability decile. (B) Average observed and predicted desirability of items, plotted over the order in which counterfactual items were generated. Error bars display ±1 *SE*. See the online article for the color version of this figure.

In all three experiments, we also observed a negative relationship between average item desirability and order of generation, $r_s(8) = -.479$, $p = .162$, 95% CI = [−0.851, 0.215] in Experiment 3A; $r_s(8) = -.733$, $p = .016$, 95% CI = [−0.932, −0.192] in Experiment 3B; $r_s(8) = -.576$, $p = .082$, 95% CI = [−0.884, 0.084] in Experiment 3C, as well as between average item likelihood and order, $r_s(8) = -.430$, $p = .214$, 95% CI = [−0.833, 0.273] in Experiment 3A; $r_s(8) = -.770$, $p = .009$, 95% CI = [−0.942, −0.273] in Experiment 3B; $r_s(8) = -.236$, $p = .511$, 95% CI = [−0.753, 0.462] in Experiment 3C. These relationships are illustrated in Figure 9A and 9B, respectively.

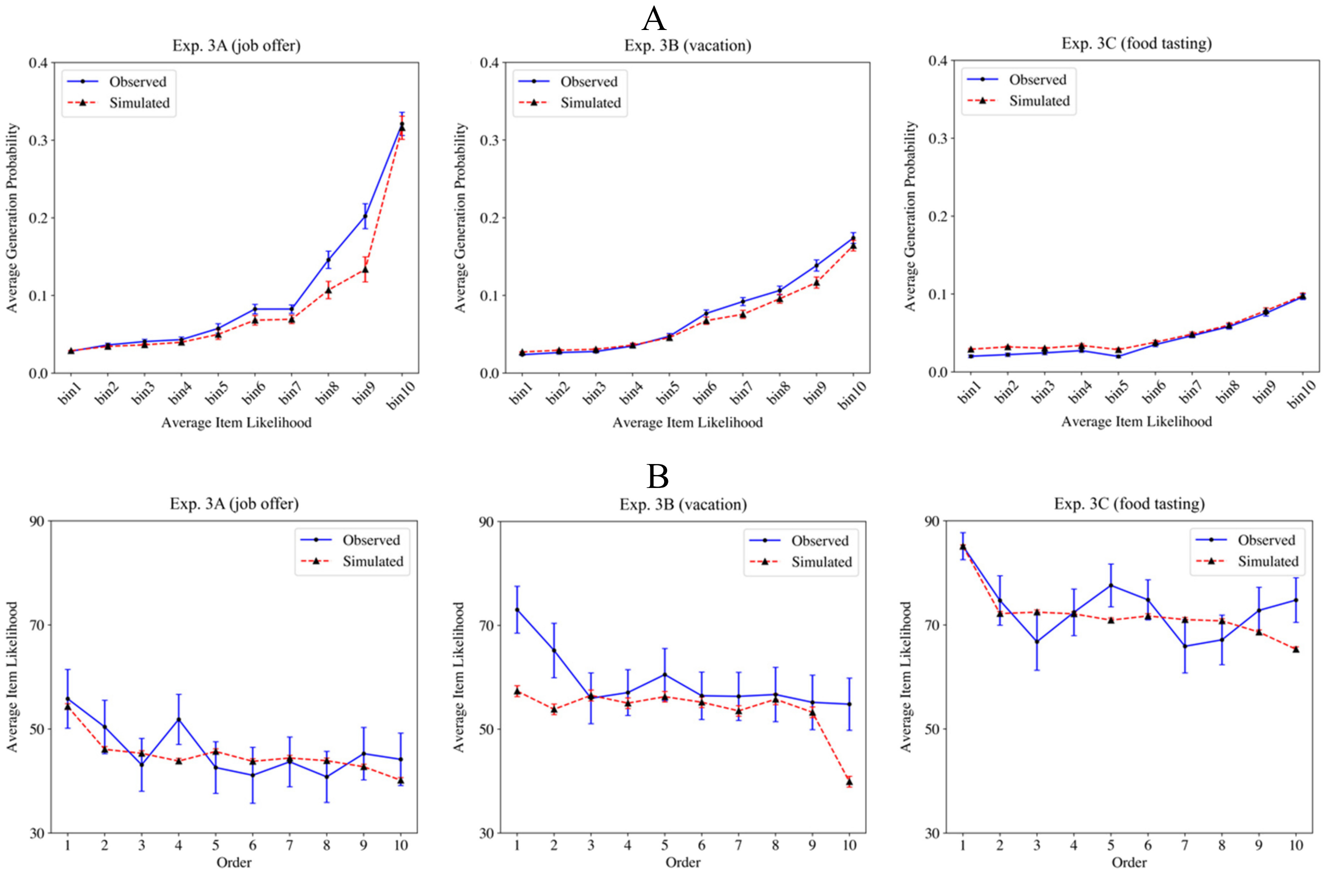
Additionally, we found that desirability and likelihood were highly correlated in all three experiments, $r(191) = .935$, $p < .001$, 95% CI = [0.914, 0.951] in Experiment 3A; $r(191) = .926$, $p < .001$, 95% CI = [0.903, 0.944] in Experiment 3B; $r(186) = .949$, $p < .001$, 95% CI = [0.933, 0.962] in Experiment 3C. This is quite reasonable for the evaluative scenarios examined in our experiments (e.g., people are likely to accept jobs only in countries that they find desirable).

Similarity With the Target. All three experiments replicated the relationship between target similarity and counterfactual generation. The average cosine similarity between counterfactuals and their corresponding targets is significantly higher than the expected cosine similarity between two randomly selected items (which is 0.322 in Experiments 3A and 3B, and 0.426 in Experiment 3C), shown by a one sample *t* test, $t(522) = 17.037$, $p < .001$, 95% CI = [0.425, 0.451] in Experiment 3A; $t(498) = 22.291$, $p < .001$, 95% CI = [0.445, 0.469] in Experiment 3B; $t(366) = 17.816$, $p < .001$, 95% CI = [0.513, 0.534] in Experiment 3C.

Furthermore, target items were randomly assigned to participants in these experiments. Thus, we should also expect the counterfactual items listed by a given participant to be more similar with the target item assigned to that participant, relative to the other three target items that were not assigned. We examined this this using paired *t* tests which compared the cosine similarity between each counterfactual and the assigned target, against the average of the cosine similarities between the generated counterfactual and the three unassigned target items. Overall, we found stronger and significant similarity effect of the assigned target relative to the unassigned targets, $t(522) = 10.829$, $p < .001$, 95% CI = [0.075, 0.108] in Experiment 3A; $t(498) = 7.896$, $p < .001$, 95% CI = [0.043, 0.072] in Experiment 3B; $t(366) = 8.756$, $p < .001$, 95% CI = [0.036, 0.057] in Experiment 3C. To examine whether the similarity effect varied

Figure 9

Experiments 3A–3C Observed Versus Predicted Generation Probabilities and Likelihood



Note. (A) Average observed and model-predicted counterfactual generation as a function of likelihood decile. (B) Average observed and predicted likelihoods of items, plotted over the order in which counterfactual items were generated. Error bars display ±1 *SE*. See the online article for the color version of this figure.

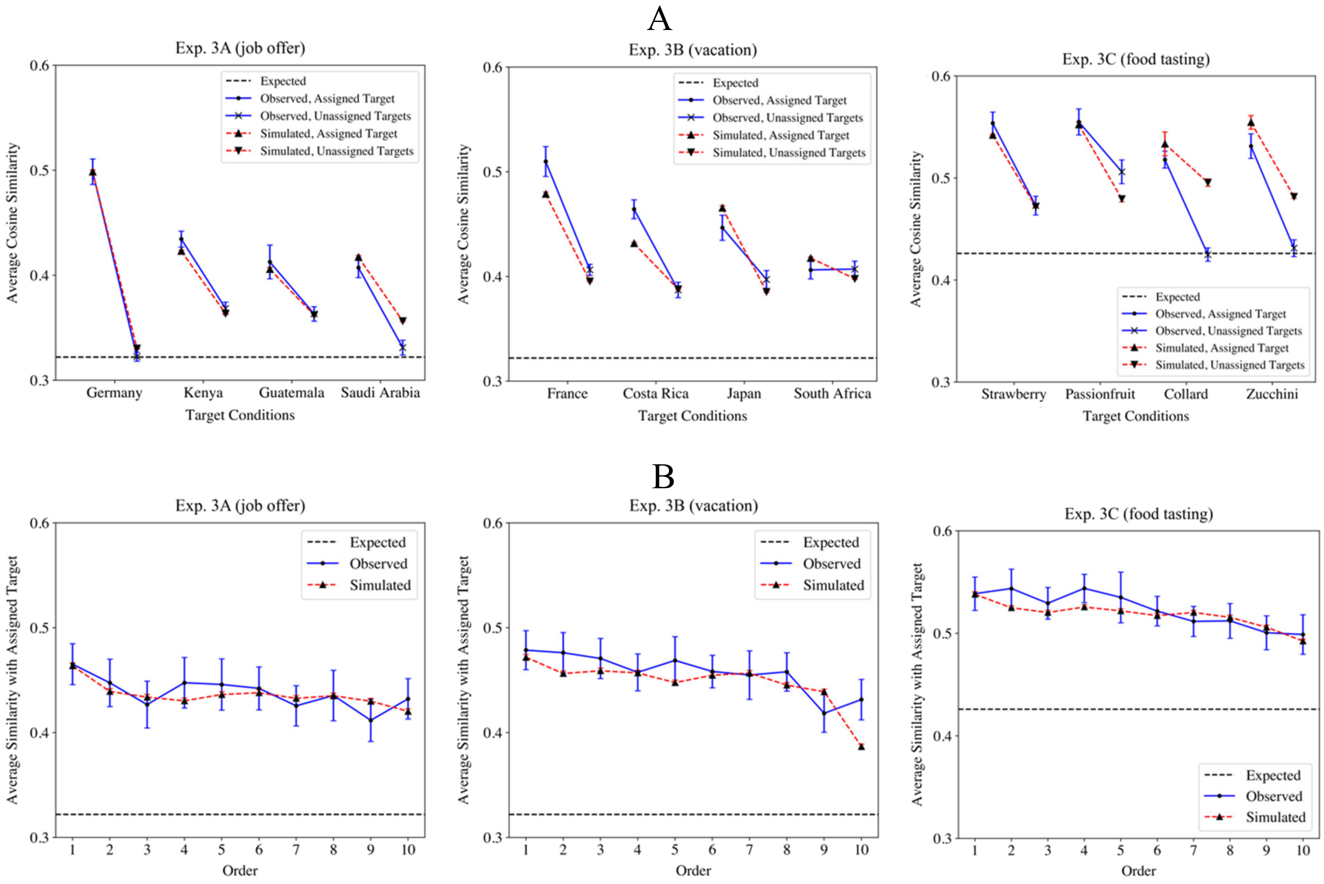
as a function of the different target items, we also conducted a similar test for each of the four targets in all three experiments. Again, we found a significant effect for eight of the twelve tests ($p < .001$ for the Germany, Saudi Arabia, and Guatemala conditions, and $p = .017$ for the Kenya condition in Experiment 3A; $p < .001$ for the Costa Rica and France conditions, and $p = .002$ for the Japan condition in Experiment 3B; $p < .001$ for the strawberry and zucchini conditions in Experiment 3C). For the South Africa condition in Experiment 3B and the passionfruit and collard conditions in Experiment 3C, we did not find a significant effect. These results are plotted in Figure 10A, which shows the average cosine similarity of counterfactuals in each condition with the assigned and with the unassigned targets. This figure also plots the expected cosine similarity between two randomly selected items as a dashed black line.

As in Experiment 1, we found that the similarity effect varied as a function of order. Specifically, the average cosine similarity between the counterfactuals and their corresponding target negatively correlated with the order in which counterfactuals are generated, $r_s(8) = -.697$, $p = .025$, 95% CI = [−0.922, −0.121] in Experiment 3A; $r_s(8) = -.891$, $p < .001$, 95% CI = [−0.974, −0.596] in Experiment 3B; $r_s(8) = -.879$, $p < .001$, 95% CI = [−0.971, −0.559] in Experiment 3C. This relationship is shown in Figure 10B, which also plots the expected cosine similarity between two randomly selected items.

Word Frequency. In all three experiments, we found a strong effect of word frequency. Items (i.e., countries or fruits and vegetables) with higher frequency in the English language were more likely to come to mind during counterfactual generation, $r(191) = .541$, $p < .001$, 95% CI = [0.433, 0.634] in Experiment 3A; $r(191) = .546$, $p < .001$, 95% CI = [0.438, 0.638] in Experiment 3B; $r(186) = .490$, $p < .001$, 95% CI = [0.373, 0.592] in Experiment 3C. Higher frequency items also appeared somewhat earlier during the generation process, although this relationship is not significant, $r_s(8) = -.139$, $p = .701$, 95% CI = [−0.706, 0.537] in Experiment 3A; $r_s(8) = -.430$, $p = .214$, 95% CI = [−0.833, 0.273] in Experiment 3B; $r_s(8) = -.564$, $p = .090$, 95% CI = [−0.880, 0.101] in Experiment 3C. These results are shown in Figure 11A and 11B, respectively. Word frequency is also positively correlated with both desirability, $r(191) = .445$, $p < .001$, 95% CI = [0.324, 0.552] in Experiment 3A; $r(191) = .395$, $p < .001$, 95% CI = [0.269, 0.508] in Experiment 3B; $r(186) = .568$, $p < .001$, 95% CI = [0.462, 0.658] in Experiment 3C, and likelihood, $r(191) = .543$, $p < .001$, 95% CI = [0.435, 0.635] in Experiment 3A; $r(191) = .578$, $p < .001$, 95% CI = [0.476, 0.665] in

Figure 10

Experiments 3A–3C Observed Versus Predicted Target Similarity



Note. (A) Observed and model-predicted average cosine similarities of counterfactuals items in each condition with the assigned target versus the unassigned targets. (B) Observed and model-predicted average cosine similarities of counterfactual items with the assigned target as a function of order. Error bars display ±1 *SE*, and the dashed black lines display average similarities expected by chance. See the online article for the color version of this figure.

Experiment 3B; $r(186) = .618$, $p < .001$, 95% CI = [0.521, 0.699] in Experiment 3C. Like fruits and vegetables, it is plausible that countries as destinations for work or travel may influence subjective evaluations through mere exposure or have greater appearance in common language.

Semantic Clustering. The memory effect of similarity with the previous item also appears in all three experiments. As illustrated in Figure 12, we found that average conditional recall probabilities (CRPs) increased with bin number, $r_s(8) = .952$, $p < .001$, 95% CI = [0.805, 0.988] in Experiment 3A; $r_s(8) = .999$, $p < .001$, 95% CI = [0.996, 1.000] in Experiment 3B; $r_s(8) = .745$, $p = .013$, 95% CI = [0.218, 0.935] in Experiment 3C. These results showed that thinking about a counterfactual item increases the chance that a semantically related counterfactual item will come to mind next.

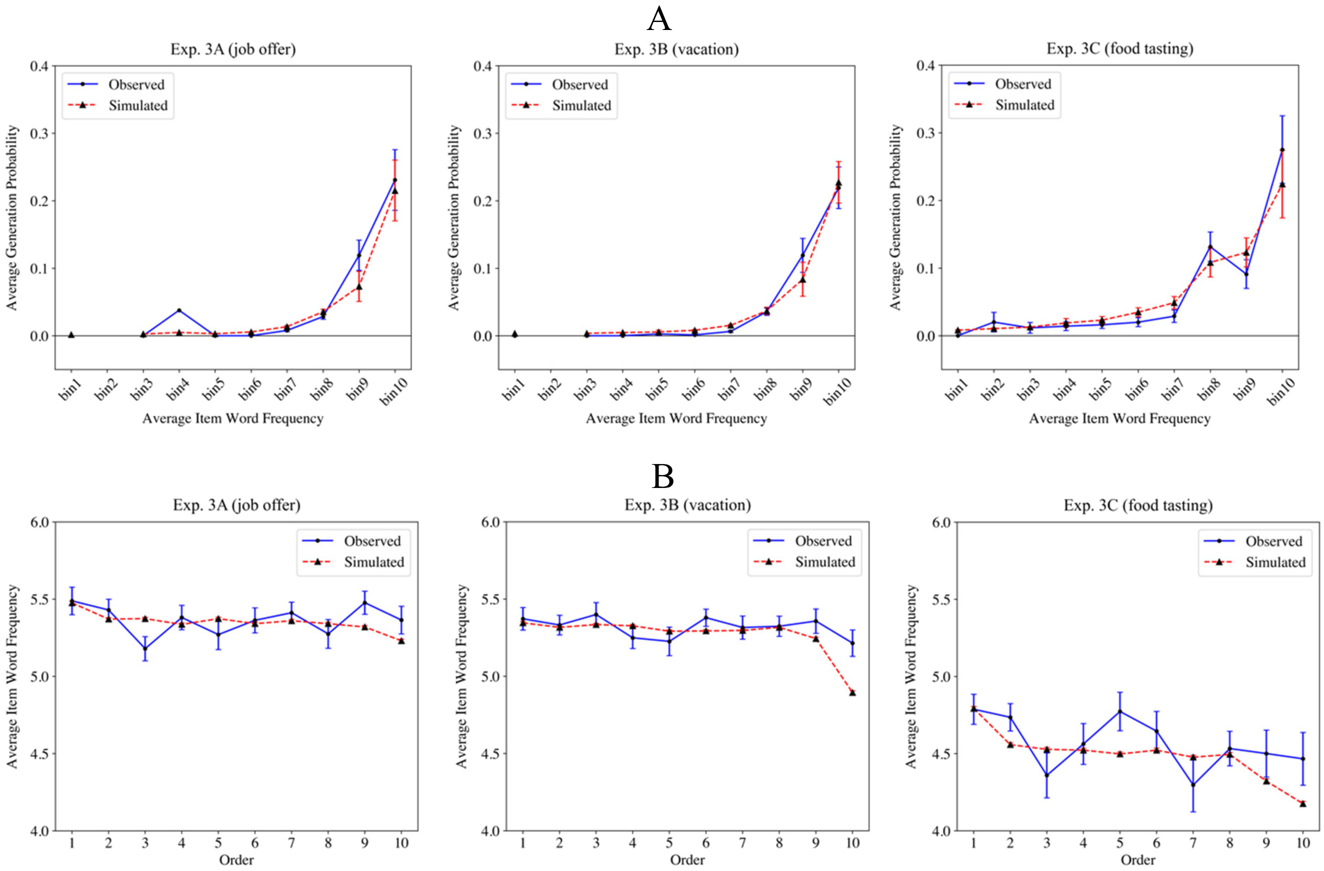
Memory Model

Model Structure and Fit. We fit the same model from Experiment 1 to each experiment here. In Experiments 3A and 3B, the states in the model are countries in the world, whereas in Experiment 3C, the states in the model are fruits and vegetables. As before, desirability ratings, likelihood ratings, similarities with the target, similarities with the previous item, and log-transformed word frequencies were standardized. We also observed instances of revisiting in the data (52.8% of participants listed the same counterfactual item twice or more in Experiment 3A, 30.2% in Experiment 3B, and 45.0% in Experiment 3C). The same approach was used to fit the model and simulate data. All $\hat{R}$ values were below 1.02. The simulated results are summarized in Table 1.

Modeling Results. Group-level means and 95% CIs, proportion of individual-level 95% CIs above 0, as well as BFs are shown in Table 3. As expected from the empirical patterns, we found robust group-level effects of desirability and likelihood across three experiments. The model predicted that more desirable items are not only more likely to come to mind as counterfactuals (Figure 8A), but they are also generated earlier than less desirable items (Figure 8B). Observed desirability of items correlated positively with their simulated probability of being listed as counterfactuals, $r(191) = .706$, $p < .001$, 95% CI = [0.627, 0.770] in Experiment 3A; $r(191) = .702$, $p < .001$, 95% CI = [0.623,

Figure 11

Experiments 3A–3C Observed Versus Predicted Generation Probabilities and Word Frequency



Note. (A) Average observed and model-predicted counterfactual generation as a function of word frequency (log-transformed) decile. (B) Average observed and predicted word frequencies (log-transformed) of items, plotted over the order in which counterfactual items were generated. Error bars display ±1 *SE*. See the online article for the color version of this figure.

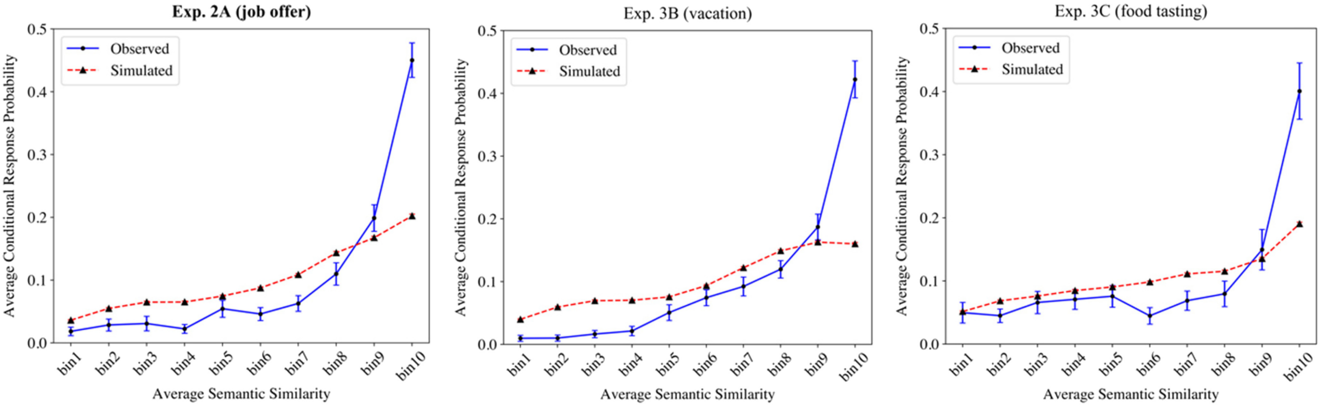
0.767] in Experiment 3B; $r(186) = .718$, $p < .001$, 95% CI = [0.641, 0.781] in Experiment 3C, and negatively with the order of the simulated counterfactuals, $r_s(8) = -.770$, $p = .009$, 95% CI = [−0.942, −0.273] in Experiment 3A; $r_s(8) = -.915$, $p < .001$, 95% CI = [−0.980, −0.674] in Experiment 3B; $r_s(8) = -.879$, $p < .001$, 95% CI = [−0.971,−0.559] in Experiment 3C. BFs indicated support for desirability in Experiment 3A and 3B, but not in Experiment 3C, which could be attributed to the degree to which the hypothetical scenarios cue desirable counterfactuals such as jobs or vacations, relative to foods at a tasting event.

Additionally, the model predicted the effect of likelihood on generation probability (Figure 9A) and order (Figure 9B). Observed item likelihood correlated positively with model-predicted generation probability, $r(191) = .865$, $p < .001$, 95% CI = [0.824, 0.897] in Experiment 3A; $r(191) = .858$, $p < .001$, 95% CI = [0.815, 0.891] in Experiment 3B; $r(186) = .756$, $p < .001$, 95% CI = [0.687, 0.811] in Experiment 3C, and negatively with the order of the simulated counterfactuals, $r_s(8) = -.842$, p = .002, 95% CI = [−0.961, −0.452] in Experiment 3A; $r_s(8) = -.685$, $p = .029$, 95% CI = [−0.918, −0.098] in Experiment 3B; $r_s(8) = -.952$, $p < .001$, 95% CI = [−0.988, −0.805] in Experiment 3C. As shown in Table 3, only data from Experiment 3B favored the inclusion of likelihood in the model as in Experiment 1 (i.e., BF > 1), which suggests that the vacation scenario may recruit similar mechanisms as the actual scenario compared to the other two scenarios employed.

Moreover, unlike the null effect in Experiment 1, we observed a significant effect of target similarity here (see Table 3). This could imply that the effect of desirability is disentangled from that of target similarity by randomly assigning targets to participants. The model also replicated the empirical patterns. As shown in Figure 10A and confirmed by paired *t* tests, simulated counterfactuals were more similar to the assigned target relative to the unassigned targets, $t(52999) = 84.342$, $p < .001$, 95% CI = [0.081, 0.085] in Experiment 3A; $t(52999) = 63.418$, $p < .001$, 95% CI = [0.054, 0.057] in Experiment 3B; $t(39999) = 113.13$, $p < .001$, 95% CI = [0.069, 0.071] in Experiment 3C. Paired *t* tests further revealed that the simulated counterfactuals were more similar to the assigned target relative to chance, $t(52999) = 134$, $p < .001$, 95% CI = [0.434, 0.438] in Experiment 3A; $t(52999) = 154.13$, $p < .001$, 95% CI = [0.446, 0.449] in Experiment 3B; $t(39999) = 175.93$, $p < .001$, 95% CI = [0.541, 0.544] in Experiment 3C. As

Figure 12

Experiments 3A–3C Observed Versus Predicted CRP



Note. Observed and model-predicted average conditional response probabilities for 10 semantic similarity bins. Error bars display ± 1 *SE*. CRP = conditional recall probabilities. See the online article for the color version of this figure.

shown in Figure 10B, the similarity of simulated counterfactuals with the assigned target also dropped as a function order, $r_s(8) = -.733$, $p = .016$, 95% CI = [−0.932, −0.192] in Experiment 3A; $r_s(8) = -.855$, $p = .002$, 95% CI = [−0.965, −0.489] in Experiment 3B; $r_s(8) = -.879$, $p < .001$, 95% CI = [−0.971, −0.559] in Experiment 3C.

Here again, our two new memory mechanisms, namely word frequency and semantic clustering, are highly robust. Figure 11A and 11B shows that the model replicated the results that word frequency was positively correlated with the simulated probability of generation, $r(191) = .608$, $p < .001$, 95% CI = [0.511, 0.690] in Experiment 3A; $r(191) = .631$, $p < .001$, 95% CI = [0.537, 0.709] in Experiment 3B; $r(186) = .764$, $p < .001$, 95% CI = [0.697, 0.818] in Experiment 3C, and negatively correlated with order, $r_s(8) = -.8188$, $p = .004$, 95% CI = [−0.955, −0.389] in Experiment 3A; $r_s(8) = -.770$, $p = .009$, 95% CI = [−0.942, −0.273] in Experiment 3B; $r_s(8) = -.976$, $p < .001$, 95% CI = [−0.994, −0.899] in Experiment 3C. Figure 12 shows that, as in Experiment 1, the model underpredicted CRP in the last bin, but it captured the overall pattern that semantically clustered counterfactuals are more likely to come to mind. Across all three experiments, CRP is positively correlated with bin numbers, $r_s(8) = .999$, $p < .001$, 95% CI = [0.996, 1.000] in Experiment 3A; $r_s(8) = .988$, $p < .001$, 95% CI = [0.949, 0.997] in Experiment 3B; $r_s(8) = .999$, $p < .001$, 95% CI = [0.996, 1.000] in Experiment 3C. Across all three experiments, BFs indicate that these memory effects are supported by data.

Finally, it is reasonable that the model underpredicted the effect of desirability, likelihood, and word frequency for later generated counterfactual items relative to earlier ones. This is because a substantial number of participants listed at least one invalid items as counterfactuals (six participants in Experiment 3A, 26 in Experiment 3B, and 18 in Experiment 3C), so there is less data later in the sequence of generated counterfactuals on which the model could calibrate. When we excluded these participants from the data rather than removing their invalid counterfactual items, the model was able to closely mimic the observed order effects for the last generated counterfactuals (see the Appendices A and B section).

Effects of Counterfactuals on Target Evaluation

We again tested whether the desirability of the counterfactual items influenced participants' evaluation of the target, and we did not find a significant effect for any of the three experiments ($\beta_1 = -0.254$, $p = .144$, 95% CI = [−0.596, 0.089] in Experiment 3A; $\beta_1 = -0.286$, $p = .161$, 95% CI = [−0.690, 0.118] in Experiment 3B; $\beta_1 = 0.047$, $p = .865$, 95% CI = [−0.512, 0.607] in Experiment 3C).

Standardizing the scale within each participant yielded mixed results ($\beta_1 = 0.213$, $p = .248$, 95% CI = [−0.153, 0.579] in Experiment 3A; $\beta_1 = -0.524$, $p = .036$, 95% CI = [−1.012, −0.036] in Experiment 3B; $\beta_1 = 1.207$, $p = .008$, 95% CI 0.342, 2.071] in Experiment 3C). These results could be due to reasons previously explained as well as the hypothetical nature of Experiments 3A–3C, and we note that the scenario in Experiment 3B (vacation) appears to align with prior findings on the effects of counterfactual thinking on target evaluation.

Discussion

In Experiments 3A–3C, we manipulated the target outcome to further examine the determinants of counterfactual generation. We used three different hypothetical scenarios (getting a job offer in a country, travelling for vacation in a country, and tasting a piece of fruits and vegetables), and randomly assigned to each participant one of four selected target outcomes (i.e., countries or fruits and vegetables). Across all three experiments, we replicated Experiment 1's findings on desirability, likelihood, similarity to the target, word frequency, and semantic clustering. Moreover, our memory model was able to capture these effects on the content as well as the sequence in which counterfactual outcomes come to mind, and the dynamics captured by our model persist across scenarios. The model also revealed that the established memory effects of word frequency and semantic clustering are important factors in counterfactual generation.

Experiment 4

Experiment 4 aims to test the effect of generation length by adapting the vacation scenario from Experiment 3B and asking participants

Table 3

Group-Level Estimation, Group-Level 95% CIs, Proportion of Individual-Level 95% CIs Above 0, and Bayes Factors

	Experiment 3A				Experiment 3B				Experiment 3C			
Variables	β_k	95% CI	Proportion (%)	*BF*	β_k	95% CI	Proportion (%)	*BF*	β_k	95% CI	Proportion (%)	*BF*
Transition probabilities												
Des.	0.538	[0.307, 0.772]	51	5.38×10^{16}	0.461	[0.286, 0.654]	40	5.20×10^{6}	0.196	[−0.013, 0.410]	5	0.070
Lik.	0.022	[−0.109, 0.154]	0	0.026	0.187	[0.035, 0.351]	8	4.09×10^{1}	0.199	[−0.007, 0.407]	0	0.006
Word Freq.	0.991	[0.820, 1.163]	100	3.87×10^{33}	0.851	[0.692, 1.009]	100	4.052	0.748	[0.593, 0.903]	100	8.05×10^{19}
Tar. Sim.	0.281	[0.119, 0.420]	32	4.25×10^{18}	0.241	[0.158, 0.316]	49	1.14×10^{7}	0.304	[0.205, 0.396]	100	4.45×10^{3}
Prev. Sim.	0.192	[0.125, 0.258]	77	6.26×10^{4}	0.314	[0.247, 0.377]	100	1.34×10^{16}	0.205	[0.100, 0.299]	22	2.11×10^{1}
Starting probabilities												
Des.	0.683	[0.117, 1.359]	0	2.10×10^{2}	0.813	[0.312, 1.431]	6	9.247×10^{1}	0.719	[0.147, 1.351]	8	0.046
Lik.	0.314	[−0.088, 0.832]	2	0.405	0.751	[0.310, 1.288]	15	1.02×10^{4}	0.706	[0.034, 1.430]	0	0.027
Word Freq.	0.964	[0.446, 1.515]	51	2.33×10^{1}	0.460	[0.017, 0.904]	0	1.546	0.940	[0.479, 1.421]	93	1.47×10^{1}
Tar. Sim.	0.685	[0.456, 0.907]	100	2.20×10^{5}	0.392	[0.191, 0.586]	62	7.154	0.439	[0.199, 0.652]	88	0.265

Note. CI = confidence interval; BF = Bayes factor; Des. = desirability; Lik. = likelihood of occurrence; Word Freq. = log-transformed word frequency; Tar. Sim. = similarity with the target; Prev. Sim. = similarity with the previous item.

to generate either five or 20 counterfactual outcomes. Since the effects of desirability, likelihood, and target similarity significantly decrease as a function of order in Experiment 3B, we expect that generating fewer (vs. more) counterfactuals would lead to an increase in the overall desirability, likelihood, and target similarity of the generated outcomes. Given that the success of the model in capturing the relationship between order and each of these variables, we also expect the model predictions in Experiment 4 to reflect the differences between the two generation lengths. In this way this experiment serves as a conceptual replication of Experiment 2.

Method

All participants in Experiment 4 ($N = 210$; $M_{\text{age}} = 36.1$; 49% female, 48% male, 3% nonbinary) were recruited from Prolific Academic and performed the experiment online using their own computer interface. Participation was limited to native English speakers in the United States. The design used in this experiment was identical to that in Experiment 3B (vacation) except for the number of counterfactuals that participants were asked to generate. While participants in Experiment 3B were asked to list 10 counterfactuals that come to mind as they considered the target, half of the participants in Experiment 4 were asked to list five counterfactuals, and the other half were asked to List 20 counterfactuals, with random assignment of conditions. As in Experiment 3B, the target conditions in Experiment 4 were France, Costa Rica, Japan, and South Africa.

As in previous experiments, we removed from items that were not evaluated in Session 1 from the counterfactuals that participants generated in Session 2 (4.30% of all listed counterfactuals in the List 5 condition, and 6.18% in the List 20 condition). Additionally, since the primary analysis of Experiment 3 involved the exact order in which counterfactuals were generated, we excluded 10 participants whose responses contained 30% or more invalid items. As in previous experiments, we excluded extra items that were listed on the same screen beyond the very first item. Overall, participants in the List 5 condition generated 94 different counterfactuals, and the most common ones were “Italy” (34 times) and “Ireland” (25 times). Participants in the List 20 condition generated 166 different counterfactuals, and the most common ones were “United Kingdom” (75 times) and “Italy” (71 times).

Results

Replicating Experiment 3B With Observed Data

To begin, we replicated the key findings of Experiment 3B (focused on vacations) using a fresh set of participants and modified the length of counterfactual generation. We found that counterfactuals generated at an earlier stage were more desirable, $r_s(3) = -.999$, $p < .001$, 95% CI = [−1.000, −1.000] in the List 5 condition; $r_s(18) = -.732$, $p < .001$, 95% CI = [−0.871, −0.429] in the List 20 condition, and more likely to occur, $r_s(3) = -.999$, $p < .001$, 95% CI = [−1.000, −1.000] in the List 5 condition; $r_s(18) = -.814$, $p < .001$, 95% CI = [−0.923, −0.581] in the List 20 condition, relative to counterfactuals generated later. We also found that the probability that an item gets listed as a counterfactual was positively correlated with its average desirability, $r(191) = .615$, $p < .001$, 95% CI = [0.518, 0.696] in the List 5 condition; $r(191) = .706$, $p < .001$, 95% CI = [0.627, 0.771] in the List 20 condition, and average likelihood, $r(191) = .692$, $p < .001$, 95%

CI = [0.610, 0.759] in the List 5 condition; $r(191) = .790$, $p < .001$, 95% CI = [0.730, 0.838] in the List 20 condition. Table 1 provides a summary of these results.

Second, similarity with target also dropped as a function of order, $r_s(3) = -.800$, $p = .104$, 95% CI = [−0.986, 0.279] in the List 5 condition; $r_s(18) = -.947$, $p < .001$, 95% CI = [−0.979, −0.969] in the List 20 condition. Overall, participants generated counterfactuals that were more similar to their assigned target than expected by chance using one sample t tests, $t(511) = 25.668$, $p < .001$, 95% CI = [0.445, 0.465] in the List 5 condition; $t(1744) = 32.458$, $p < .001$, 95% CI = [0.416, 0.429] in the List 20 condition. Moreover, paired t tests revealed that these counterfactuals were also more similar with their assigned target than the unassigned targets, $t(511) = 10.003$, $p < .001$, 95% CI = [0.055, 0.082] in the List 5 condition; $t(1744) = 9.558$, $p < .001$, 95% CI = [0.027, 0.041] in the List 20 condition.

Third, word frequency is negatively correlated with order only when participants listed 20 counterfactuals, $r_s(18) = -.580$, $p = .007$, 95% CI = [−0.813, −0.185] in the List 20 condition, but not when they generated five counterfactuals, $r_s(3) = .600$, $p = .285$, 95% CI = [−0.599, 0.969] in the List 5 condition. Nevertheless, the probability that a counterfactual comes to mind is positively correlated with its word frequency, $r(191) = .438$, $p < .001$, 95% CI = [0.316, 0.546] in the List 5 condition; $r(191) = .626$, $p < .001$, 95% CI = [0.532, 0.705] in the List 20 condition. We also we replicated the semantic clustering effect using the CRP analysis introduced in Experiment 1. Here, we found that CRPs were positively correlated with bin number when aggregated across participants, $r_s(3) = .976$, $p < .001$, 95% CI = [0.672, 0.998] in the List 5 condition; $r_s(18) = .999$, $p < .001$, 95% CI = [0.996, 1.000] in the List 20 condition. This indicates that participants are more likely to think about counterfactuals that are similar to what has previously come to mind.

Effect of Length on Counterfactual Generation

The main goal of our experiment was to manipulate counterfactual outcomes by varying the number of counterfactuals that participants were asked to generate. We predicted that participants in the List 5 condition would list counterfactuals that were on average more desirable, more likely, and more similar with the target than those in the List 20 condition. We found that these predictions were supported. Overall, the average desirability of the counterfactuals in the List 5 condition was 78.4, compared to 73.4 in the List 20 condition. This difference is statistically significant with an unpaired t test, $t(2255) = 3.434$, $p < .001$, 95% CI = [2.131, 7.804]. Additionally, a simple linear regression of the Session 1 desirability ratings of counterfactuals on condition (with the List 5 condition coded as 1 and the List 20 condition coded as 0) revealed that participants generated significantly more desirable items in the List 5 condition than in the List 20 condition ($\beta_1 = 4.968$, $p < .001$, 95% CI = [2.133, 7.803]). These results are illustrated in Figure 13A.

We performed similar tests for likelihood of occurrence and found the same pattern. As illustrated in Figure 13B, the average likelihood of occurrence of generated counterfactuals was 60.3 in the List 5 condition but 55.8 in the List 20 condition, and this difference is significant with an unpaired t test, $t(2255) = 3.434$, $p < .001$, 95% CI = [0.995, 8.016], and a regression of likelihood ratings on condition ($\beta_1 = 4.505$, $p = .012$, 95% CI = [0.996, 8.014]).

In addition, Figure 13C revealed that participants are more likely to generated counterfactuals that are more similar with the target when they were asked to list five instead of 20 items. The average target similarity of the simulated counterfactuals was 0.455 in the List 5 condition but 0.422 in the List 20 condition, and this difference is significant with an unpaired t test, $t(2255) = 5.063$, $p < .001$, 95% CI = [0.020, 0.045], and a regression of counterfactual similarity on condition ($\beta_1 = 0.032$, $p < .001$, 95% CI = [0.020, 0.045]).

The memory effect of word frequency was not expected to correlate with strong effect with length as we have observed an insignificant correlation in Experiment 3B. As illustrated in Figure 13D, we did not find a difference between average word frequency of the counterfactuals in the five-length versus 20-length simulations, $t(2255) = 1.168$, $p = .243$, 95% CI = [−0.025, 0.099]; $\beta_1 = 0.037$, $p = .243$, 95% CI = [−0.025, 0.099].

Overall, these results show that our manipulation was able to successfully alter the set of generated counterfactuals. On average, participants listed items that were more desirable, more likely, and more similar with the target when they were asked to generate fewer items.

Memory Model

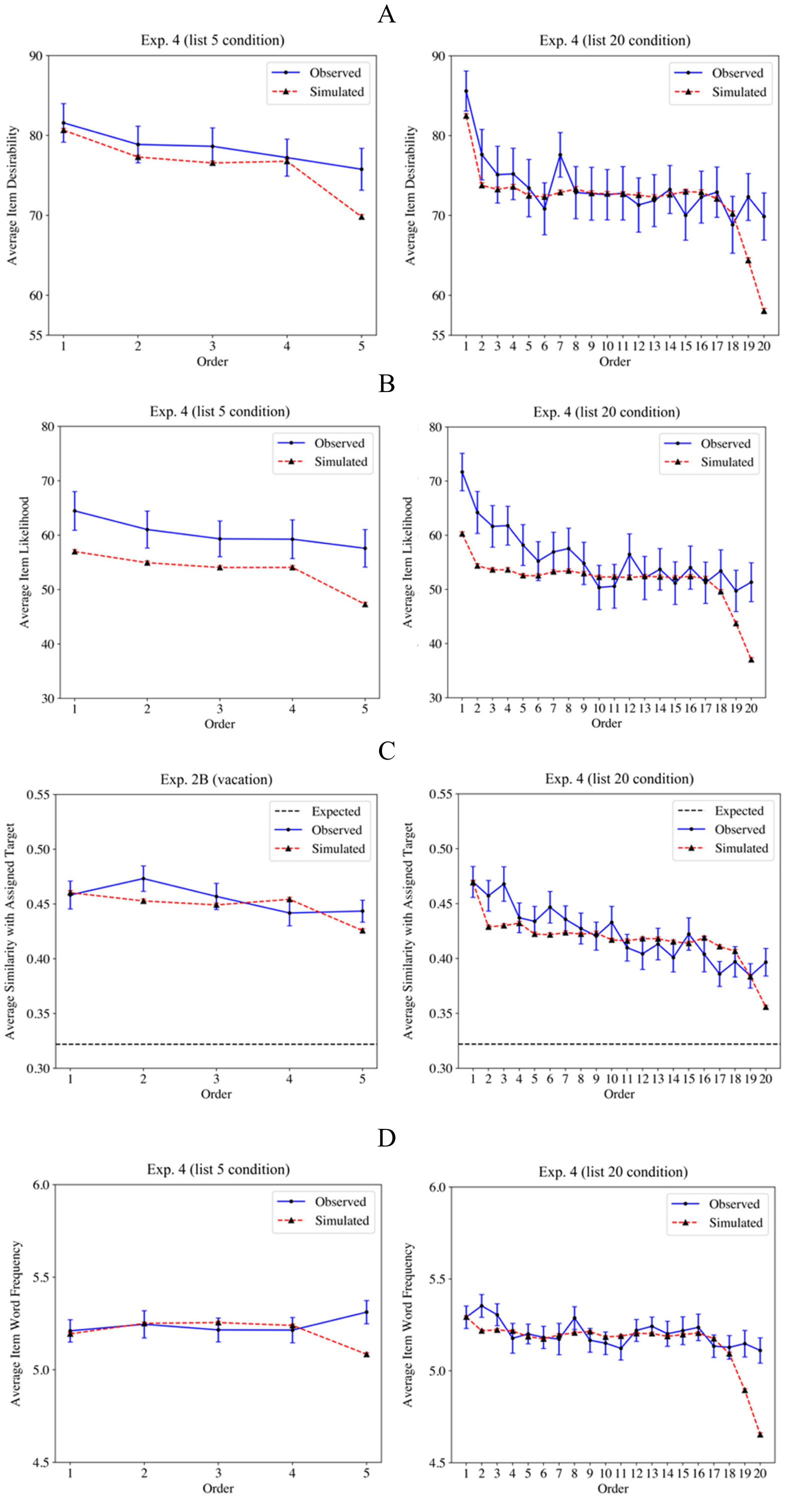
Model fit and posterior predictive checks were performed using the same approach as in Experiment 3B. All $\hat{R}$ values were below 1.02. We observed more instances of revisiting in the List 5 condition (15.0%) compared to the List 20 condition (59.1%), which is reasonable as participants had fewer chance to revisit in the List 5 condition. Simulations also showed substantially more revisiting in the 20-length simulations (92.1%) than in the five-length simulations (18.0%). The simulated results are summarized in Table 1.

Replicating Experiment 3B With Model Predictions

First, the model captures the relationship between generation probability and desirability, $r(191) = .739$, $p < .001$, 95% CI = [0.667, 0.797] in the List 5 condition; $r(191) = .745$, $p < .001$, 95% CI = [0.674, 0.802] in the List 20 condition, as well as the relationship between generation probability and likelihood, $r(191) = .848$, $p < .001$, 95% CI = [0.802, 0.883] in the List 5 condition; $r(191) = .801$, $p < .001$, 95% CI = [0.819, 0.893] in the List 20 condition. Second, the model generated counterfactuals that are more similar with the assigned target similarity than random chance, $t(53499) = 156.15$, $p < .001$, 95% CI = [0.447, 0.450] in the List 5 condition; $t(185999) = 249.41$, $p < .001$, 95% CI = [0.417, 0.418] in the List 20 condition, and more similar with the assigned target than the unassigned targets, $t(53499) = 68.708$, $p < .001$, 95% CI = [0.058, 0.061] in the List 5 condition; $t(185999) = 82.761$, $p < .001$, 95% CI = [0.033, 0.034] in the List 20 condition). Third, the model captures the memory effect of word frequency on generation probability, $r(191) = .588$, $p < .001$, 95% CI = [0.487, 0.673, in the List 5 condition; $r(191) = .697$, $p < .001$, 95% CI = [0.616, 0.763] in the List 20 condition, and semantic clustering, $r_s(3) = .976$, $p < .001$, 95% CI = [0.672, 0.998] in the List 5 condition; $r_s(18) = .988$, $p < .001$, 95% CI = [0.969, 0.995] in the List 20 condition.

Figure 13

Experiments 4 Observed Versus Predicted Desirability, Likelihood, Target Similarity, and Word Frequency



Note. (A) Average observed and model-predicted item desirability as a function of order. (B) Average observed and model-predicted item likelihoods as a function of order. (C) Average observed and model-predicted cosine similarity with the assigned target as a function of order. (D) Average observed and model-predicted log-transformed word frequency as a function of order. Error bars display ±1 *SE*. See the online article for the color version of this figure.

Model Predictions of the Length Effect

As shown in Figure 13A–13D, our model was able to successfully predict differences across the conditions. Overall, the average desirability of the counterfactuals in the five-length simulations was 76.2, compared to 72.0 in the 20-length simulations, and this difference is statistically significant with an unpaired *t* test, $t(239498) = 29.16$, $p < .001$, 95% CI = [3.915, 4.479]. This relationship is illustrated in Figure 13A. In addition, a regression of counterfactual desirability on condition (1 for the List 5 condition and 0 for the List 20 condition) showed that the model generated data contained significantly more desirable items in the List 5 condition than in the List 20 condition ($\beta_1 = 4.197$, $p < .001$, 95% CI = [3.915, 4.479]).

The average likelihood of occurrence of our model generated counterfactuals was 53.5 in the List 5 condition but 51.8 in the List 20 condition, and this difference is significant, $t(239498) = 9.703$, $p < .001$, 95% CI = [1.369, 2.062]. A regression of counterfactual likelihood on condition (1 for the List 5 condition and 0 for the List 20 condition) revealed that participants generated items with significantly higher likelihood in the List 5 condition than in the List 20 condition ($\beta_1 = 1.715$, $p < .001$, 95% CI = [1.369, 2.062]). This relationship is shown in Figure 13B.

In addition, the average target similarity was .449 in the List 5 condition but only .417 in the List 20 condition, and this difference is significant with an unpaired *t* test, $t(239498) = 37.29$, $p < .001$, 95% CI = [0.030, 0.033]. A regression of target similarity on condition (1 for the List 5 condition and 0 for the List 20 condition) revealed that the model generated items that were more similar to the target in the List 5 condition than in the List 20 condition ($\beta_1 = 0.031$, $p < .001$, 95% CI = [0.030, 0.033]). This relationship is shown in Figure 13C. Overall, these findings demonstrate the model's ability to accurately capture the impact of length on generation.

Discussion

Experiment 4 conceptually replicate the findings from Experiment 2, demonstrating that these effects extended to a different set of counterfactuals related to vacation destinations. We observed that counterfactuals generated at an earlier stage tended to be more desirable and likely, while also exhibiting greater similarity to the target. The consistency of these results underscores the importance of considering retrieval length as an important factor shaping the nature of counterfactuals in human cognition.

General Discussion

The present study attempted to build a formal parametric model of counterfactual generation and test its predicted dynamics. We used a Markov memory model that specifies generation as a random walk over counterfactual items in memory, and it allows for the desirability, likelihood, similarity with the target, word frequency, and semantic similarity to influence their generation probabilities. Our tests showed that both desirability and likelihood have a strong impact on counterfactual generation, even when controlling for other variables in the model. To begin with, people seem to have a general tendency to think about what was likely to have occurred when asked to consider counterfactual alternatives to a given outcome. This result is in line with prior work. For example, Phillips et al. (2019) found that, across diverse tasks, the alternative possibilities that people consider by default are biased toward what is probable and what is valuable.

Our findings also revealed an underexplored relationship between counterfactual thinking and memory search. Although neuroscientific evidence indicates an overlap between recalling experiences and imagining counterfactuals (e.g., Schacter et al., 2015), past work has not studied this relationship using computational memory models. Inspired by the free recall list learning paradigm in memory research, we devised a task in which participants listed counterfactual thoughts in response to a particular target outcome or event. Our tests found a positive effect of semantic similarity with the target item, in line with existing findings on the importance of target similarity in counterfactual thinking (De Brigard et al., 2021; Kahneman & Miller, 1986). Replicating research on semantic congruence in free recall and semantic memory search (Bhatia, 2019; Hills et al., 2012; Howard & Kahana, 2002), we also found a very strong effect semantic similarity with the previously generated item. Furthermore, we found that a robust effect of word frequency (rather than contextual diversity). These findings suggest that counterfactuals that were closely related to the target outcome in consideration and to previously generated counterfactuals, as well as highly frequent in language, were more likely to come to mind. It is worth noting that the semantic congruence and word frequency effects were stronger than the other effects, as indicated by model fits. This indicates that fundamental memory processes may have a substantial impact on counterfactual thinking.

We further showed that counterfactual items that come to mind earlier are more desirable, likely to occur, and similar to the target than the later ones. A growing body of literature in the domain of decision making provides insights into this tendency with the finding that people often start by considering highly valuable choice options when faced with a decision (Hauser, 2014). In fact, what comes to mind first is often considered as the best option, which people end up choosing (Johnson & Raab, 2003; Klein et al., 1995). It is also not surprising that counterfactuals that are generated earlier are more similar to the target, as the target probably cues relevant information in memory, but its effect diminishes as deliberation progresses and additional information becomes available.

Importantly, we showed that our model was able to capture the influence of order and, as shown in Experiments 2 and 4, predict how manipulating the number of generated items alters the overall desirability and likelihood. In this way, we demonstrated the value of our approach for modeling the effects of contextual and task-related factors on the dynamics of counterfactual generation. We believe that further exploring the role of these contextual factors is an important topic for future work, and that prior research on memory processes provides a useful set of empirical results with which to develop these tests. For example, we would expect that irrelevant primes would alter counterfactual generation by activating associated items, and that manipulations like cognitive load would make generation more stochastic and by doing so reduce the effect of desirability, likelihood, and target similarity, on generation. Testing whether our modeling approach can successfully account for these manipulations is good topic for future work.

Another avenue for future work is the further refinement of our model. We found that although our model was able to successfully capture all observed behavioral patterns, it underpredicted the semantic clustering effect in all experiments and the likelihood effect in Experiment 3B and Experiment 4, possibly due to nonlinearity. In future work we hope to address this limitation by developing a modified variant of our model with more complex generation functions. Another limitation of our model is its inability to capture the gradual reduction in the desirability, likelihood, and target similarity effects as deliberation progresses. In its current form, the model allows for two sets of parameters: one set for the starting generation probabilities, and another set for the remaining generation probabilities. This means that the model can capture a change in generation tendencies between the first and subsequent positions, but not between any other positions in the sequence of generated counterfactuals. Future work could thus attempt to develop a model that allows for a more graded change in retrieval as deliberation progresses.

Theories in the decision-making literature have treated counterfactuals as a reference class which influences the subsequent choice evaluations (Loomes & Sugden, 1982; Mellers et al., 1997; Stewart et al., 2006). Some recent work has also identified an inverse relationship between the actual value of a chosen option and the inferred value of an unchosen option (Biderman & Shohamy, 2021). Even though our experiment was designed (and powered) to test for generation, and not evaluation, we examined the possibility that when people generate counterfactual outcomes in response to a target outcome, a target outcome would appear less desirable in the context of desirable counterfactuals, whereas the target would appear more desirable in the context of undesirable counterfactuals.

These tests showed that the Session 2 evaluation of the target largely depended on its initial (Session 1) evaluation rather than on the desirability of the generated counterfactuals in Session 2. At least for the scenarios employed in our experiments, people seemed to have a stable desirability rating for the target after retrieving counterfactuals. This does not mean that counterfactuals do not play a role in evaluation. Rather it could be the case that our design, which asked subjects to generate their own counterfactuals, led to the generation of the same counterfactuals in Session 1 as in Session 2. Developing an empirical paradigm to measure the impact of self-generated counterfactuals on evaluation is an important topic for future work.

In addition, our counterfactual generation tasks only considered the gain domain. Participants were asked to imagine receiving a job offer, winning a free vacation trip, or tasting food for free, but they did not consider counterfactuals in the context of losing or failing a goal. Indeed, ruminating on desirable but unobtained outcomes has significant implications in mental health and is associated with anxiety and depression (Davis et al., 1995; Rachman et al., 2000; Roese et al., 2009). Interestingly, people generate both upward (more desirable) and downward (less desirable) counterfactuals when they experience a loss (Markman et al., 1993), which seems to depend on one's cognitive style (e.g., Kasimatis & Wells, 2014; Sanna et al., 1999). Future work could test the role of desirability on counterfactual retrieval in the loss domain. Here, we expect our memory model to provide substantial insight by formally specifying the effect of domain type on counterfactual retrieval tendencies.

Relatedly, it would also be useful to examine the effect of the type or domain of the event on counterfactual generation. We expect that participant responses would change with the task. So, if instead of countries for job offers participants have to imagine having traffic accidents, they may list countries where they would prefer having traffic accidents or countries where such accidents are especially likely. We believe that our model would be able to capture this through the desirability and likelihood of occurrence variables, both of which are task-specific (e.g., desirability in Experiment 3A corresponds to desirability of a job in a country whereas desirability in Experiment 3B corresponds to the desirability of a vacation in a country, so similarly one could in principle elicit desirability ratings for a traffic accident in a country). Of course, our model would also predict that semantic similarity with previously retrieved items and with the target item would influence retrieval. Since this is task independent, the model would predict correlations across tasks. Exploring these properties of our model is quite important, especially if one wishes to develop a task-independent (or task-general) model of counterfactual retrieval.

Building on our model, future studies can investigate the effects of individual differences in counterfactual thinking. One promising domain for such an analysis involves aging. Researchers can fit separate models for elderly populations and parametrically specify the effects of aging on counterfactual retrieval. This will shed light on the specific set of memory mechanisms that are damaged with age. Prior work has found that older people are more likely to confuse counterfactual simulations for remembered events (Gerlach et al., 2014). By examining the parameters that increase the perceived similarity between counterfactuals and experiences in memory, researchers may be able to obtain new insights about age-related differences in cognition and behavior. The promise of our model extends beyond age-related cognitive impairments to other types of disorders. For example, by correlating individual-level model parameters inferred from counterfactual retrieval data with neural data, researchers can better understand brain regions implicated in disruptions to counterfactual simulation. These studies can facilitate the development of interventions that improve real-world cognition in impaired populations.

To conclude, we have presented a novel approach to modeling retrieval dynamics in counterfactual thought. By building a Markov model of counterfactual retrieval, we have applied insights from memory research to investigate the mental processes involved in counterfactual thinking. We have validated existing findings on counterfactual thinking and have found new evidence for the important role of core memory processes in counterfactual retrieval. Our work opens up many research questions with several applications in the cognitive and behavioral sciences, and we look forward to future work that uses our modeling framework to understand the determinants and consequences of counterfactual thought.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558–569. https://doi.org/10.1037/a0038693

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Aka, A., & Bhatia, S. (2021). What I like is what I remember: Memory modulation and preferential choice. *Journal of Experimental Psychology: General*, *150*(10), 2175–2184. https://doi.org/10.1037/xge0001034

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, *194*, Article 104057. https://doi.org/10.1016/j.cognition.2019.104057

Bhatia, S. (2019). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 627–640. https://doi.org/10.1037/xlm0000618

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Biderman, N., & Shohamy, D. (2021). Memory and decision making interact to shape the value of unchosen options. *Nature Communications*, *12*(1), Article 4648. https://doi.org/10.1038/s41467-021-24907-x

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, *30*(2), 149–165. https://doi.org/10.1080/00221309.1944.10544467

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*(1), 135–157. https://doi.org/10.1146/annurev-psych-122414-033249

Chapman, C. A., & Martin, R. C. (2022). Effects of semantic diversity and word frequency on single word processing. *Journal of Experimental Psychology: General*, *151*(5), 1035–1068. https://doi.org/10.1037/xge0001123

Cofer, C. N., Bruce, D. R., & Reicher, G. M. (1966). Clustering in free recall as a function of certain methodological variations. *Journal of Experimental Psychology*, *71*(6), 858–866. https://doi.org/10.1037/h0023217

Davis, C. G., Lehman, D. R., Wortman, C. B., Silver, R. C., & Thompson, S. C. (1995). The undoing of traumatic life events. *Personality and Social Psychology Bulletin*, *21*(2), 109–124. https://doi.org/10.1177/0146167295212002

De Brigard, F., Henne, P., & Stanley, M. L. (2021). Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility. *Cognition*, *209*, Article 104574. https://doi.org/10.1016/j.cognition.2020.104574

De Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science*, *28*(1), 59–66. https://doi.org/10.1177/0963721418806512

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gerlach, K. D., Dornblaser, D. W., & Schacter, D. L. (2014). Adaptive constructive processes and memory accuracy: Consequences of counterfactual simulations in young and older adults. *Memory*, *22*(1), 145–162. https://doi.org/10.1080/09658211.2013.779381

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, *78*(1–3), 111–133. https://doi.org/10.1016/0001-6918(91)90007-M

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*(1), 23–29. https://doi.org/10.1037/h0040561

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Gruenewald, P. J., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 225–240. https://doi.org/10.1037/0278-7393.6.3.225

Hall, J. F. (1954). Learning as a function of word-frequency. *The American Journal of Psychology*, *67*(1), 138–140. https://doi.org/10.2307/1418080

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research*, *67*(8), 1688–1699. https://doi.org/10.1016/j.jbusres.2014.02.015

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*(1), 85–98. https://doi.org/10.1006/jmla.2001.2798

Jaya, I. G. N. M., Tantular, B., & Andriyana, Y. (2019). A Bayesian approach on multicollinearity problem with an informative prior. *Journal of Physics: Conference Series*, *1265*(1), Article 012021. https://doi.org/10.1088/1742-6596/1265/1/012021

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, *23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Johnson, J. G., & Raab, M. (2003). Take the first: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes*, *91*(2), 215–229. https://doi.org/10.1016/S0749-5978(03)00027-X

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(1), 107–138. https://doi.org/10.1146/annurev-psych-010418-103358

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

Kasimatis, M., & Wells, G. L. (2014). Individual differences in counterfactual thinking. In N. J. Roese & J. M. Olson (Eds.), *What might have been* (pp. 81–101). Psychology Press.

Klein, G., Wolf, S., Militello, L., & Zsambok, C. (1995). Characteristics of skilled option generation in chess. *Organizational Behavior and Human Decision Processes*, *62*(1), 63–69. https://doi.org/10.1006/obhd.1995.1031

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1943–1946. https://doi.org/10.1037/a0033669

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language*, *64*(3), 249–255. https://doi.org/10.1016/j.jml.2010.11.003

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824. https://doi.org/10.2307/2232669

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology*, *29*(1), 87–109. https://doi.org/10.1006/jesp.1993.1005

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423–429. https://doi.org/10.1111/j.1467-9280.1997.tb00455.x

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26, pp. 3111–3119). Curran Associates, Inc.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*(6), 887–899. https://doi.org/10.3758/bf03209337

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*(18), 4649–4654. https://doi.org/10.1073/pnas.1619717114

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, *23*(12), 1026–1040. https://doi.org/10.1016/j.tics.2019.09.007

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Rachman, S., Grüter-Andrew, J., & Shafran, R. (2000). Post-event processing in social anxiety. *Behaviour Research and Therapy*, *38*(6), 611–617. https://doi.org/10.1016/S0005-7967(99)00089-3

Richie, R., Aka, A., & Bhatia, S. (2023). Free association in a neural network. *Psychological Review*, *130*(5), 1360–1382. https://doi.org/10.1037/rev0000396

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 56, pp. 1–79. Academic Press.

Roese, N. J., Epstude, K. A. I., Fessel, F., Morrison, M., Smallman, R., Summerville, A., Galinsky, A. D., & Segerstrom, S. (2009). Repetitive regret, depression, and anxiety: Findings from a nationally representative survey. *Journal of Social and Clinical Psychology*, *28*(6), 671–688. https://doi.org/10.1521/jscp.2009.28.6.671

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*(1), 28–34. https://doi.org/10.1111/j.1467-9280.1993.tb00552.x

Sanna, L. J., Turley-Ames, K. J., & Meier, S. (1999). Mood, self-esteem, and simulated alternatives: Thought-provoking affective influences on counterfactual direction. *Journal of Personality and Social Psychology*, *76*(4), 543–558. https://doi.org/10.1037/0022-3514.76.4.543

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory*, *117*, 14–21. https://doi.org/10.1016/j.nlm.2013.12.008

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223–247. https://doi.org/10.1146/annurev-psych-010814-015135

Stan Development Team. (2021). *RStan: The R interface to Stan*. https://mc-stan.org/rstan/

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Sumby, W. H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, *1*(6), 443–450. https://doi.org/10.1016/S0022-5371(63)80030-4

Wang, F., Aka, A., & Bhatia, S. (2023, September 8). *Memory modeling of counterfactual generation*. https://osf.io/497ct

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, *56*(2), 161–169. https://doi.org/10.1037/0022-3514.56.2.161

Wells, G. L., Taylor, B. R., & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology*, *53*(3), 421–430. https://doi.org/10.1037/0022-3514.53.3.421

Zhao, W., Richie, R., & Bhatia, S. (2022). Process and content in decisions from memory. *Psychological Review*, *129*(1), 73–106. https://doi.org/10.1037/rev0000318

Zultan, R. I., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Causality and counterfactuals in group attributions. *Cognition*, *125*(3), 429–440. https://doi.org/10.1016/j.cognition.2012.07.014

Appendix A

Inclusion and Exclusion of Invalid Counterfactuals

Experiment 2

Including Participants Who Have Listed Invalid Items as Counterfactuals

Participants who listed one or more invalid items as counterfactuals were included in the analyses. Note, however, that one participant in the List 20 condition only listed one valid counterfactual item and had to be excluded due to the modeling structure (Figure A1).

Excluding Participants Who Have Listed Invalid Items as Counterfactuals

Participants who listed at least one invalid items as counterfactuals were excluded from these analyses. The model fits improved compared to Figure 6A–6D (Figure A2).

Experiments 3A–3C

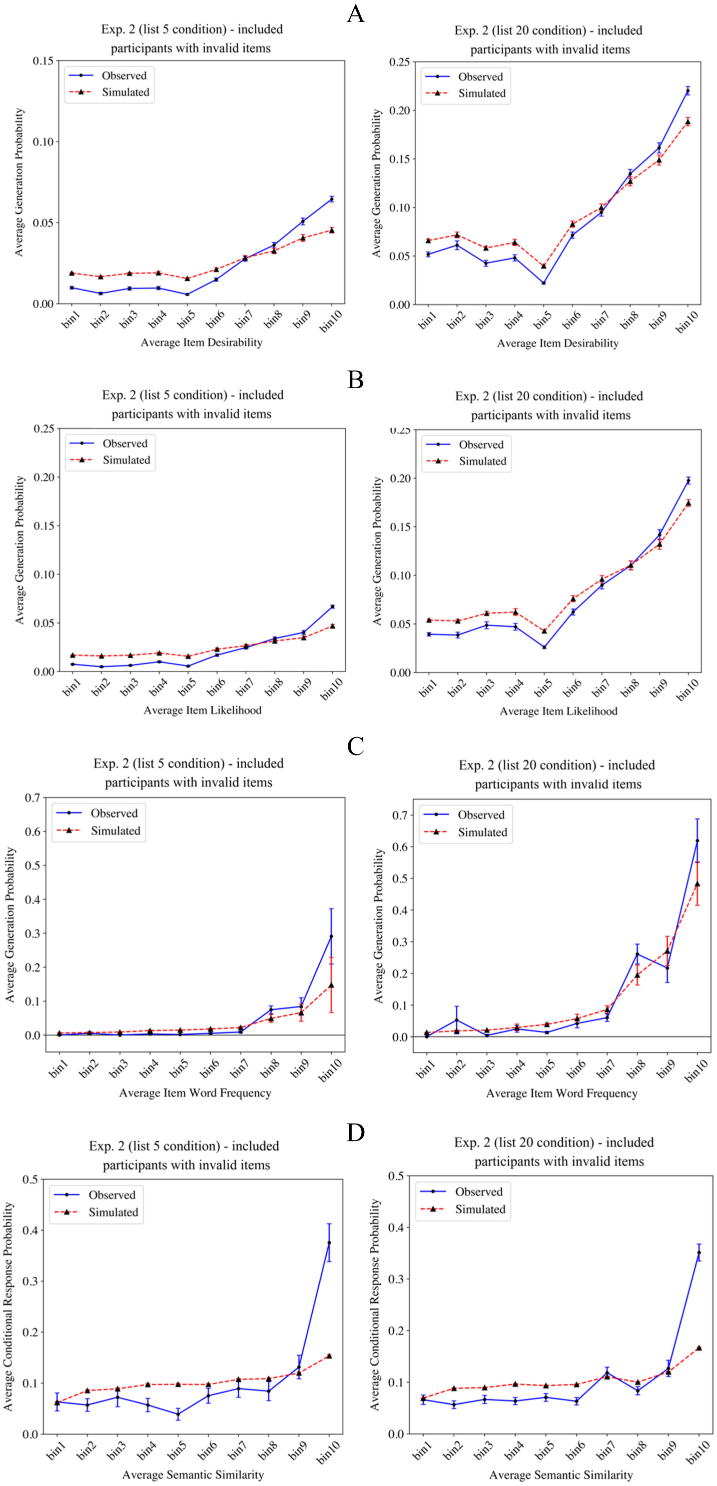
Excluding Participants Who Have Listed Any Invalid Items

Participants who listed at least one invalid items as counterfactuals are excluded. Again, the model was better able to capture the empirical patterns compared to Figures 8–11 (Figure A3).

(*Appendices continue*)

Figure A1

Experiment 2 Included Participants With Invalid Items

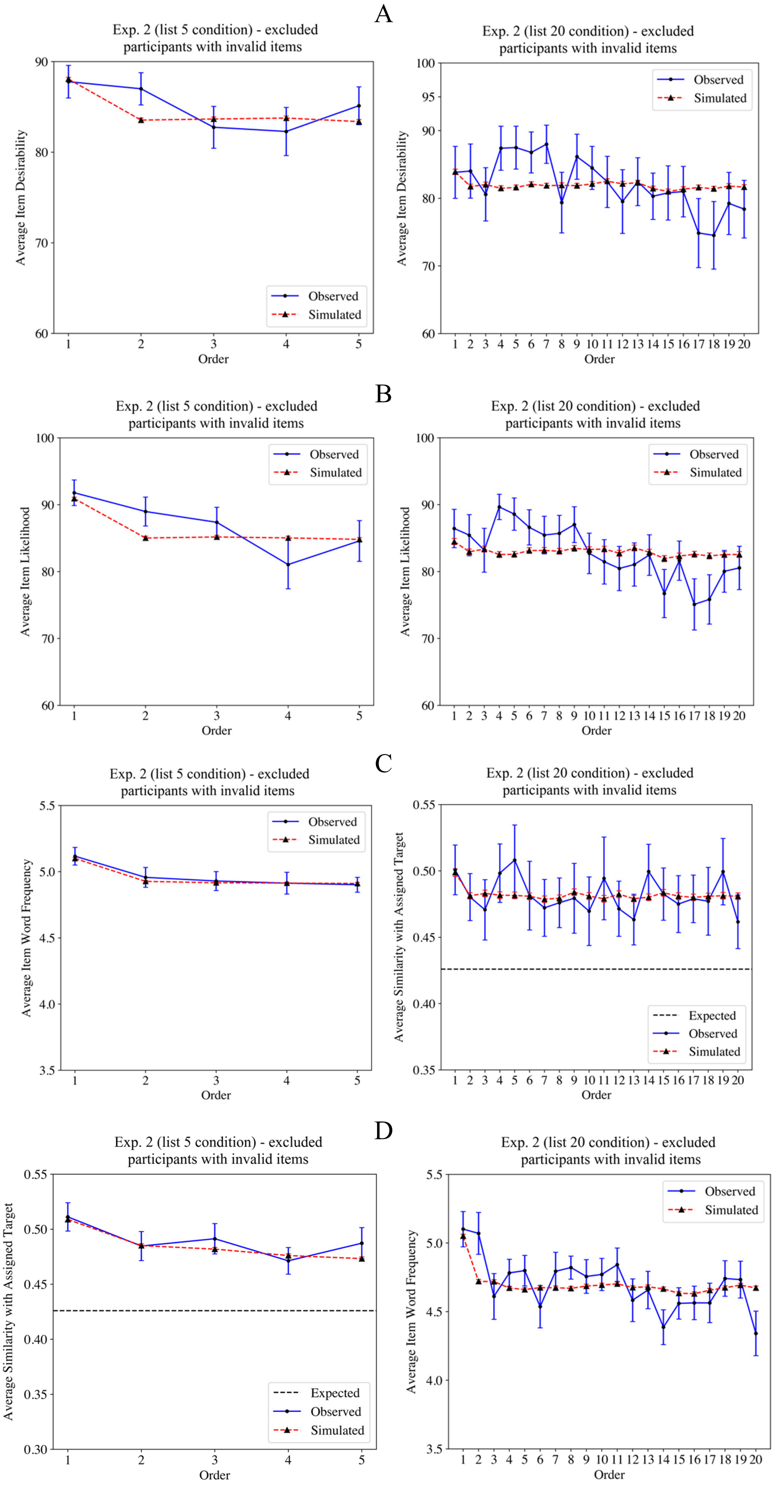


Note. Average observed and model-predicted probabilities of counterfactual generation as a function of item (A) desirability decile, (B) likelihood decile, and (C) word frequency (log-transformed) decile. (D) Observed and model-predicted average conditional response probabilities for 10 semantic similarity bins. Error bars display ±1 *SE*. See the online article for the color version of this figure.

(*Appendices continue*)

Figure A2

Experiment 2 Excluded Participants With Invalid Items

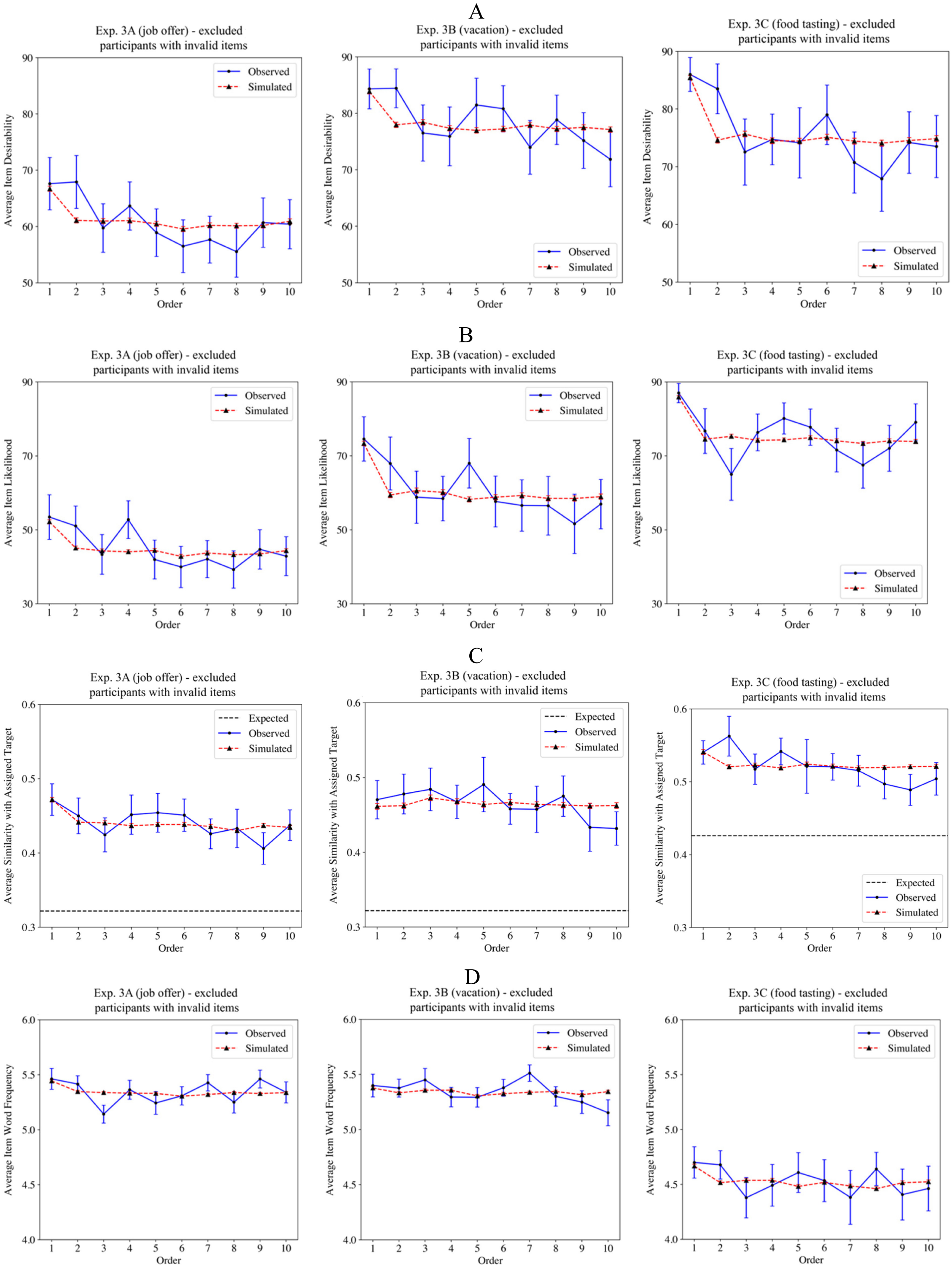


Note. (A) Average observed and predicted desirability, (B) likelihood, and (C) average observed and model-predicted cosine similarity with the assigned target as a function of order. (D) Average observed and model-predicted log-transformed word frequencies of items, plotted over the order in which counterfactual items were generated. Error bars display ±1 *SE*. See the online article for the color version of this figure.

(*Appendices continue*)

Figure A3

Experiments 3A–3C Excluded Participants With Invalid Items



Note. (A) Average observed and predicted desirability, (B) likelihood, and (C) average observed and model-predicted cosine similarity with the assigned target as a function of order. (D) Average observed and model-predicted log-transformed word frequencies of items, plotted over the order in which counterfactual items were generated. Error bars display ±1 *SE*. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix B

Survey Prompts

Prompts

Experiment 3A

Imagine that you are a senior and are looking for jobs after you graduate. You have a particular company in mind that you really want to work for. This company offers the exact kind of work that you like and provides a wonderful paycheck. You anticipate excellent career advancement opportunities at this company. On your previous visit there, people seemed nice and you had a great conversation with the boss, who was very supportive and spoke highly about your potentials.

You had an interview with this company two weeks ago, and have just received a call from the company giving you a job offer. They offer to match all the salary and benefits that you have previously discussed.

However, the location of the job is outside of the United States. If you accept the offer, you will be working in the company's main office in [X].

Experiment 3B (and Experiment 4)

Imagine that it's 2022. The pandemic has passed and everything is back to normal. You have signed up for the International Travel Sweepstake for a chance of winning a free trip. The top prize is worth $4,250 and it covers all of your travel expenses. If you win, you know exactly how you are going to spend that precious vacation. You will visit all the tourist spots, eat your favorite local food, and share lots of selfies with your friends back home. You just can't wait to go on that vacation trip!

You have just received a call from the organization, and--congratulations!–you have won a free five-day vacation trip in [X]!

Experiment 3C

Imagine that it's 2022. The pandemic has passed and everything is back to normal. Last week, your friend gave you a $50 ticket to the Annual Food Festival, a popular event where people can learn about and even taste food from all around the world! You have looked up this event online and got really excited about going there. In particular, you couldn't wait to check out the free tasting event where they offer fresh fruits and vegetables!

Today, the festival finally arrives! You rush straight to the free tasting event and find lots of tables with fruits and vegetables on top. While you are looking around, a staff member calls you up and offers you to taste what's on the table. You pick up the item and it's a [X]!

Received March 1, 2022
Revision received September 8, 2023
Accepted January 13, 2024 ■