Proceedings of SALT 30: 603–623, 2020

Counterfactuals and Undefinedness: Homogeneity vs Supervaluations*

Paul Marty
University College London

Jacopo Romoli
University of Bergen

Paolo Santorio
University of Maryland, College Park

Abstract Theories of counterfactuals agree on appealing to a relation of comparative similarity, but disagree on the quantificational force of counterfactuals. We report on two experiments testing the predictions of three main approaches: universal theories, homogeneity theories, and single-world selection theories (plus supervaluations over selection functions). The critical cases in our experiments were constructed so as to discriminate between the three theories. Our results provide empirical support for the selectional theories, while challenging the other two approaches.

Keywords: counterfactuals, homogeneity, supervaluations, undefinedness

1 Introduction

On mainstream theories, counterfactuals like (1) exploit a relation of comparative closeness between worlds (see Lewis 1973a,b, 1979). Via comparative closeness, we can determine a set of antecedent-verifying worlds that functions as the domain of quantification for the conditional. For example, (1) quantifies over the set of closest worlds to the actual world where ticket #37 is bought.

(1) If ticket #37 was bought, it would win a prize.

At the same time, different theories disagree about the quantificational force of counterfactuals. Three main views have gained prominence in the literature. One is the classical theory of Lewis (1973a; 1973b) and Kratzer (2012), on which counterfactuals are simply universal quantifiers over closest antecedent worlds. The second is Stalnaker's selectional theory (1968; 1981; 1984). On this theory, the semantics of counterfactuals requires them to select a single closest antecedent world, and cases

* For very helpful comments and suggestions, we would like to thank Simona Aimar, Moysh Bar-Lev, Itai Bassi, Lucas Champollion, Dean Hughes, Angelika Kratzer, Dan Lassiter, Matt Mandelkern, Todd Snider, the audiences at SALT30 and at ELM1.

where there is no such a single world are handled via supervaluations. On the third theory, counterfactuals are universal quantifiers, as on the Lewis-Kratzer view, but they also enforce a homogeneity requirement (see von Fintel 1997, Schlenker 2004). This requirement demands that either all antecedent worlds make the consequent true, or else that all of them make the consequent false.

All three theories capture some basic data about counterfactuals. In addition, the selectional theory and the homogeneity theory make analogous predictions for unembedded cases. But embeddings of counterfactuals under certain quantifiers pull apart the predictions of all three theories. This paper reports on two experiments that investigate some of these cases. Our results provide empirical support for the selectional theory and pose a challenge for the other two.

The rest of this paper is structured as follows. In §2, we discuss the three theories in more detail, as well as the novel predictions we target. In §3–4, we report on two experiments that tested these predictions. In §5, we discuss the results in relation to the theoretical predictions. Finally, §7 concludes. Throughout the paper, we use the traditional notation '$\mathsf{A} \boxrightarrow \mathsf{C}$' as shorthand for the counterfactual *If* A*, would* C.

2 Background: counterfactuals and undefinedness

2.1 Three theories of counterfactuals

Counterfactual conditionals are standardly treated as modalized sentences whose semantics appeals to a relation of comparative closeness (denoted by '$\preceq$'). Most accounts agree on the general form of their truth conditions: $\mathsf{A} \boxrightarrow \mathsf{C}$ is predicted to be true just in case C is true in some relevant range of close A-verifying worlds.[1] But different accounts make different claims about the quantificational force associated to counterfactuals. Here we review three prominent options.

2.1.1 Universal Theories

Accounts belonging in this family treat counterfactuals as universal quantifiers (see Lewis 1973a,b and Kratzer 1986, 2012 for some classical representatives). Simplifying somewhat, the schematic truth conditions that these accounts assign to a counterfactual are in (2).[2]

1 There are some well-known dynamic variants of the static accounts that we present: for discussion, see von Fintel 2001 and Gillies 2007. For our purposes, we can lump dynamic accounts with universal theories, since they make analogous predictions about the relevant sentences.

2 The simplification consists in making the so-called limit assumption, which Lewis overtly disavows (besides Lewis, see Kaufmann 2017 for discussion). Issues concerning the limit assumption are irrelevant for our purposes.

(2) $[\![\mathsf{A} \mathrel{\square\!\!\rightarrow} \mathsf{C}]\!]^{w,\preceq} = \text{true}$ iff $\forall w'$: $w' \in \textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]^{w,\preceq})$, $[\![\mathsf{C}]\!]^{w',\preceq} = \text{true}$
where $\textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]^{w,\preceq})$ is the set of maximally $\preceq$ −close worlds to w

To illustrate, suppose that Maria considered flipping a coin yesterday at noon, but didn't do it in the end, and suppose that we utter (3) in this context.

(3) If Maria had flipped the coin, it would have landed heads.

The truth conditions that the universal theories predict for (3) are in (4). On the assumption that the closest worlds to the actual world involve a mixture of heads and tails-worlds, (3) is thus predicted to be false in the suggested context.

(4) $[\![(3)]\!]^{w,\preceq} = \text{true}$ iff $\forall w'$: $w' \in \textsc{max}_{w,\preceq}([\![\textbf{flip}]\!]^{w,\preceq})$, $[\![\textbf{heads}]\!]^{w',\preceq} = \text{true}$

2.1.2 Selectional Theories

On selectional theories, counterfactuals select a single closest antecedent-verifying world (see Stalnaker 1968, 1981, 1984). A counterfactual is true if and only if the selected world also verifies the consequent. Following Stalnaker, we state the semantics using selection functions, i.e., functions of the form $s : W \times \mathscr{P}(W) \mapsto W$ mapping a pair of a proposition and an 'input' world to a selected world.[3] On this view, the truth conditions of a counterfactual are, schematically, the following:

(5) $[\![\mathsf{A} \mathrel{\square\!\!\rightarrow} \mathsf{C}]\!]^{w,s} = \text{true}$ iff $[\![\mathsf{C}]\!]^{s(w,[\![\mathsf{A}]\!]),s} = \text{true}$

The selectional theory does not appeal directly to a notion of comparative closeness, but talk of selection functions can be rephrased into talk of comparative closeness (*modulo* background assumptions about the properties of the comparative closeness relation): the selected world is the single closest world to the world of evaluation that makes true the antecedent of a counterfactual.[4]

Without supplementation, the selectional theory runs into a well-known difficulty. The selectional semantics requires that, for every antecedent and every world of

3 Here are the full conditions that Stalnaker imposes on selection functions:

i. If $[\![A]\!]$ is non-empty, $s(w,A) \in [\![A]\!]$

ii. If $s(w,A) = \lambda$, then $[\![A]\!] = \varnothing$
(where λ is the absurd world, i.e., a world where every sentence is true)

iii. If $w \in [\![A]\!]$, then $s(w,A) = w$

iv. For all A, A': if $s(w,A) \in [\![A']\!]$ and $s(w,A') \in [\![A]\!]$, then $s(w,A) \in [\![A']\!] = s(w,A') \in [\![A]\!]$

4 For discussion of this point, see Lewis 1973a, Chapter 2. The background assumption needed is that the relation of comparative closeness induces a linear order on worlds.

evaluation w, there is a single closest antecedent-world to w. As examples like (3) suggest, however, this assumption is highly implausible: in a situation where Maria is flipping a fair coin, it appears that some heads-worlds and some tails-worlds will be tied for closeness, no matter what specific construal of closeness we adopt. In fact, Stalnaker (1981; 1984) agrees that, in many cases, for some choice of antecedent A and some world w, there won't be a single closest A-world to w. He suggests that this problem should be handled not in the semantics proper, but rather in the metasemantics. Cases of this sort will be treated as cases where it is indeterminate which selection function is the 'right' one. This kind of predicament can be modeled via supervaluations.[5]

The idea behind supervaluations is the following. In cases like (3), there are several selection functions that are equally plausible candidates for being the selection function individuated by the context. Given this, we may define notions of determinate truth and determinate falsity by quantifying over these candidate selection functions. More specifically, we define determinate truth as truth at all the $\langle w, s\rangle$ pairs, where s is a candidate selection function at the relevant context; determinate falsity is defined in an analogous fashion. Finally, we say that a sentence is *undefined* just in case it is neither determinately true nor determinately false.[6]

> A is **determinately true (false)** at c iff, for all $\langle w_c, s\rangle$ such that s is a candidate selection function at c, $[\![\mathsf{A}]\!]^{w,s}$ is true (false).

Note that determinate truth and determinate falsity at a context replace the classical Kaplanian notions of truth and falsity at a context (see Kaplan 1989). These notions are not part of the compositional semantics proper. Rather, they apply after the compositional computation of semantic value is complete. This will play an important role in the way that the undefinedness of counterfactuals projects under embeddings. For now, let us see how the account works by considering again (3):

(3) If Maria had flipped the coin, it would have landed heads.

Plausibly, there are several candidate selection functions for an utterance of (3). On some of them, the selection function maps the world of evaluation and the antecedent of (3) to a heads-world and, on some others, to a tails-world. On these assumptions, (3) is thus predicted to be undefined. Therefore, we have a difference in predictions here between the universal and the selectional theory.

5 Supervaluations were introduced by Van Fraassen in 1969. We also refer the reader to Fine 1975 for a classical use of supervaluations to model vagueness.

6 On a number of views about indeterminacy, this terminology might be misleading, since A's not being determinately true is compatible with it being true (see Barnes & Williams 2011). We want to be clear that, despite the appeal to this terminology, we remain neutral on the underlying issue.

2.1.3 Homogeneity Theories

Homogeneity theories (von Fintel 1997, Schlenker 2004) have features in common with both universal and selectional theories. Like universal theories, they treat counterfactuals as universal quantifiers over closest antecedent worlds. Like selectional theories, they assume that some counterfactuals will be undefined. Crucially, however, in this case undefinedness results from a definedness condition that requires the domain of quantification of the counterfactual to be homogeneous with respect to the consequent. That is, the definedness condition requires that either all closest antecedent-worlds are consequent-worlds, or else that none are.[7] Below are the schematic truth conditions for a counterfactual on this account:

(6) $[\![\mathsf{A} \mathrel{\square\!\!\rightarrow} \mathsf{C}]\!]^{w,\preceq} = \begin{cases} \underline{\text{defined}} \text{ iff either } \forall w' \text{: } w' \in \textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]^{w,\preceq}), [\![\mathsf{C}]\!]^{w',\preceq} = \text{true} \\ \qquad\qquad \text{or } \forall w' \text{: } w' \in \textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]^{w,\preceq}), [\![\mathsf{C}]\!]^{w',\preceq} = \text{false} \\ \underline{\text{true}} \text{ iff } \forall w' \text{: } w' \in \textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]^{w,\preceq}), [\![\mathsf{C}]\!]^{w',\preceq} = \text{true} \end{cases}$

The appeal to a definedness condition is an attempt at reproducing some desirable logical features of selectional semantics. In particular, like the selectional theory (and unlike the universal theory), the homogeneity theory vindicates the negation swap inferences reported and illustrated by the pair in (7).[8]

(7) **Negation swap** $\neg(\mathsf{A} \mathrel{\square\!\!\rightarrow} \mathsf{C}) \dashv\models \mathsf{A} \mathrel{\square\!\!\rightarrow} \neg\mathsf{C}$

a. It's not the case that, if Maria had flipped the coin, the coin would have landed tails.

b. If Maria had flipped the coin, the coin would not have landed tails.

At the same time, homogeneity theories also capture some of the advantages of universal theories. For example, they correctly predict that *would*-counterfactuals are duals of *might*-counterfactuals. Thus for instance, they correctly predict the incompatibility of the counterfactuals in (8).

(8) a. If Maria had flipped the coin, the coin would have landed tails.

b. If Maria had flipped the coin, the coin might not have landed tails.

Given this background, consider again our benchmark example (3):

(3) If Maria had flipped the coin, it would have landed heads.

On the assumption that some heads-worlds and some tails-worlds are equally close, homogeneity theories predict that the homogeneity requirement is not satisfied

7 As Schlenker 2004 emphasizes, this proposal is motivated by a suggestive analogy with the behavior of plural definite phrases like *the girls*.

8 Given that the semantics of conditionals is trivalent, the relevant notion of entailment here is Strawson-entailment. See von Fintel 1999 for discussion.

for (3), and hence the sentence is undefined. Thus, selectional and homogeneity theories yield the same verdict for (3). In fact, provided that we make symmetrical assumptions about what worlds are tied for closeness and what selection functions are admissible, the two theories will always agree on their verdicts for unembedded conditionals. Importantly for our purposes, however, there is a key difference between the way the two families of theories treat undefinedness. On selectional theories, the compositional semantics is fully bivalent, and undefinedness only emerges when we define truth at a context. Conversely, homogeneity theories are trivalent. Hence we have constituents of sentences whose semantic value is undefined. As a result, on this theory, we need a projection algorithm which tells us in what way the undefinedness of simpler expressions affects the definedness of complex expressions.

The projection of homogeneity is a matter of live debate in the literature (see Križ 2015, Križ & Chemla 2015 among others for discussion). For our purposes, however, this debate is not central. As we point out below, the test case that we will focus on is one where all projection theories in the literature are in agreement.

2.2 Global vs Local undefinedness: diverging predictions

The key difference between supervaluational and homogeneity theories concerns the stage at which undefinedness emerges. On the supervaluational view, undefinedness emerges at the *global* level. A sentence A is undefined at c iff, at c, there are candidate selection functions s_1 and s_2 such that A is true relative to $\langle w, s_1 \rangle$ and false relative to $\langle w, s_2 \rangle$. But, at all compositional stages, the semantics is indistiguishable from a standard bivalent theory. Conversely, on the homogeneity view, homogeneity emerges at the *local* level. This means that clauses that are embedded in a complex sentence are undefined if the homogeneity requirement is not satisfied.

Two theories that differ in this respect will make identical predictions for unembedded counterfactuals.[9] However, we can pull apart their predictions when we consider embeddings under certain operators. For the purposes of this paper, we will focus on embeddings under negative universal determiner phrases like *no ticket*. Let us introduce our example. Consider the following scenario:

> There is a raffle where prize-winning tickets are selected via a random draw among all the tickets bought. Only some of the tickets among those bought will win a prize, and any ticket has the same chance of winning and losing.

Consider first a simple counterfactual about a random ticket in the lot, like (9):[10]

9 As we mentioned above, this holds provided that we make analogous assumption about comparative closeness and selection functions in both cases.

10 As Simon Goldstein and Angelika Kratzer have independently pointed out to us, (9) is not a contrary-

(9) If ticket #37 was bought, it would win a prize.

(9) is predicted to be undefined by both supervaluational and homogeneity theories, on plausible assumption about closeness. To explain why, let us first lay out the relevant assumptions about selection/closeness. We assume that, in the relevant scenario, all candidate selection functions map the world of evaluation and the antecedent of (9) to worlds where some but not all of the tickets win. Among the worlds in this set, some are worlds where ticket #37 wins, and some are worlds where ticket #37 loses. We can rephrase this point in terms of comparative closeness: all worlds in the relevant set of closest worlds are worlds where some but not all tickets win. Within this set, some of these worlds are worlds where #37 wins, and some are worlds where #37 loses.

It is easy to see how these assumptions lead to undefinedness. Here is the prediction of the supervaluational theory:

(10) (9) is undefined at c iff for some candidate selection functions s_1 and s_2 compatible with c, $[\![(9)]\!]^{w,s_1}$ = true and $[\![(9)]\!]^{w,s_2}$ = false

Since by assumption there are two such selection functions, (9) is undefined at the relevant context. And here is the prediction of the homogeneity theory:

(11) $[\![(9)]\!]^{\langle w,\preceq\rangle}$ = undefined iff

(i) $\exists w'$: $w' \in \textsc{max}_{w,\preceq}([\![\text{\#37 bought}]\!]^{w,\preceq})$, $[\![\text{\#37 win}]\!]^{w',\preceq}$ = true, and

(ii) $\exists w'$: $w' \in \textsc{max}_{w,\preceq}([\![\text{\#37 bought}]\!]^{w,\preceq})$, $[\![\text{\#37 win}]\!]^{w',\preceq}$ = false

That is, (9) is undefined if in some closest worlds where ticket #37 is bought the ticket wins, and in some closest worlds where ticket #37 is bought the ticket loses. Since this condition holds, (9) is again predicted to be undefined.

Let us now consider a more complex sentence. Holding fixed the raffle scenario above, consider the following sentence:

(12) No ticket would win a prize if it was bought.

Unlike (9), (12) pulls apart the predictions of the two theories. To explain why, let us again make some plausible assumptions about selection. We assume that, given the functioning of the raffle, worlds where some of the tickets win and some of the tickets lose are closer than all other worlds (let us call them 'win-some-lose-some' worlds). In terms of selection function, we assume that, all candidate selection

to-fact conditional strictly speaking, but rather a so-called future-less-vivid conditional, i.e., it concerns a future event. We are assuming here that *would* has analogous quantificational force in future-less-vivid and contrary-to-fact conditionals. We plan on running follow-up experiments on contrary-to-fact conditionals in the next phase of the project.

functions in the context will map counterfactuals to a win-some-lose-some world, unless the antecedent of a counterfactual explicitly contradicts this. Consider first the supervaluational theory. We have, again:

(13) (12) is undefined at c iff for some candidate s_1 and s_2 compatible with c, $[\![(12)]\!]^{w,s_1}$ = true and $[\![(12)]\!]^{w,s_2}$ = false

In this case, (12) is predicted to be defined. Indeed, all candidate selection functions take us to a win-some-lose-some world. Hence, on all of them, (12) is evaluated as false. As a result, the supervaluational theory predicts that (12) has a determinate truth value, and that it is false.

Holding fixed the assumptions about closeness, the homogeneity theory makes a different prediction. This theory exploits a trivalent compositional semantics, so we need to determine the definedness conditions for (12) on a compositional basis. First, note that (12) has the structure in (14):

(14) No ticket$_x$ [[if x was bought][x would win a prize]]

From our discussion of (9), we know that, for all values of x, the embedded clause *if x was bought, x would win a prize* is undefined. To get definedness conditions for the full sentence, we need to determine how undefinedness projects under negative determiner phrases like *no ticket*. The literature includes two main options (see Križ 2015, Križ & Chemla 2015; see also George 2008, Fox 2012, Mandelkern 2016 for a corresponding debate related to presupposition projection):

- **Existential projection**. $No_x[F(x)][G(x)]$ is defined iff, for at least one object o in the domain of quantification, $F(o) \wedge G(o)$ is defined.
- **Universal projection**. $No_x[F(x)][G(x)]$ is defined iff, for every object o in the domain of quantification, $F(o) \wedge G(o)$ is defined.

For our current purposes, this choice is irrelevant. As we pointed out, the open sentence embedded under *No ticket* in (12) is false for all objects in the domain. So, no matter what projection algorithm we choose, (12) is predicted to be undefined.

2.3 Interim summary

In this section, we have introduced three families of theories of counterfactuals: universal, selectional, and homogeneity theories. Both selectional and homogeneity theories predict that some counterfactuals are undefined. Moreovoer, the two yield analogous predictions for unembedded counterfactuals like (9), repeated below. At the same time, they disagree for at least some cases of embeddings. In particular, they disagree about examples where counterfactuals are embedded under negative

THEORY	Example (9)	Example (12)
Universal	false	true
Selectional	undefined	false
Homogeneity	undefined	undefined

Table 1 Predictions of the three approaches for the unembedded case in (9) and the negative quantifier case in (12).

determiner phrases, as in (12). Finally, universal theories predict that (9) is false, while (12) is true. These predictions are summarized in Table 1. In the next sections, we report on two experiments designed to test these predictions, and adjudicate between the three theories at hand.

(9) If ticket #37 was bought, it would win a prize.

(12) No ticket would win a prize if it was bought.

3 Experiment 1

Detecting by experimental means the failure of a sentence to be either true or false is not an easy task, and various experimental options have been explored in the previous literature toward this end (see Križ & Chemla 2015 for discussion).[11] The two experiments reported in this paper used for these purposes a graded acceptability task, much in the spirit of Ripley (2009). Participants were presented with items like the one in Figure 1. Each item involved a context, presented through a vignette, and a target sentence, in bold font. Participants had to assess the extent to which the sentence was true or false in the suggested context. They reported their judgments by setting a slider tooltip along a scale going from 'Completely false' (left anchor) to 'Completely true' (right anchor). In the critical conditions, sentences like (9) and (12) were paired with contexts in which only part of the tickets bought would win a prize, as in the example in Figure 1. We hypothesized that, if these items give rise to gap judgments, participants should set the slider toward the middle of the scale; conversely, if these items give rise to clearly true or false judgments, participants should move the slider away from the middle, closer to the extreme values.

11 Other experimental options include, among others, independent evaluation of truth and falsity (Križ & Chemla 2015: experiments A1-3), binary judgments supplemented with independent processing measures (Schwarz 2016), ternary judgments (Abrusán & Szendröi 2012; Alxatib & Pelletier 2009; Tieu, Bill & Romoli 2019) and multiple unordered choices (Serchuk, Hargreaves & Zach 2011).

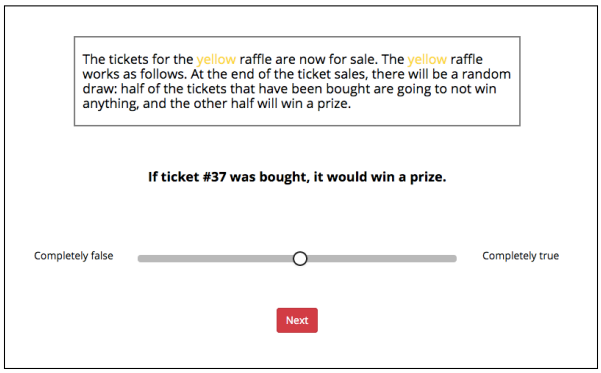

Figure 1 Example item illustrating the experimental display seen by the participants in our experiments. This item is an example of a POSITIVE target sentence in the *mixed*-context in Experiment 1.

3.1 Participants

100 participants were recruited through Prolific and were paid £1.2 for their participation. Of these, 1 was removed prior to analyses because they did not declare English as their native language. The data of the remaining 99 were included in the analyses (47 female, average age 35.9 years). All participants gave written informed consent to the processing of their personal information for the purposes of this study. All data were collected and stored in accordance with the provisions of Data Protection Act 2018, the UK's implementation of the General Data Protection Regulation.

3.2 Materials

Each item consisted of a short context followed by a test sentence (see Figure 1). Each context described the working of one of three kinds of raffles: (i) one in which all the tickets bought win a prize (*all*-context), (ii) one in which only half of the tickets bought win a prize (*mixed*-context), and (iii) one in which none of the tickets bought win a prize (*none*-context), as illustrated in (15)-(17).

(15) **All-context**
The tickets for the orange raffle are now for sale. It is the 50th anniversary of this raffle and the organizers want all participants to be content: at the end of the ticket sales, every ticket that has been bought is going to win a prize.

(16) **Mixed-context**
The tickets for the yellow raffle are now for sale. The yellow raffle works as follows. At the end of the ticket sales, there will be a random draw: half of

the tickets that have been bought are going to not win anything, and the other half will win a prize.

(17) **None-context**
The tickets for the red raffle are now for sale. But the red raffle is rigged: at the end of the ticket sales, none of the tickets that have been bought are going to win a prize.

Test sentences involved two types of targets: simple counterfactuals (POSITIVE) and counterfactuals embedded under *no ticket* (NEGATIVE), as shown in (18). For each target, a corresponding control was included in the study, (19). Crucially, these control sentences are not predicted on any approach to give rise to undefinedness, unlike our targets. Thus, they were expected to be judged as false in the critical conditions for the targets, i.e., when evaluated relative to the *mixed*-context.

(18) **Target sentences**

a. If ticket #37 was bought, it would win a prize. POSITIVE

b. No ticket would win a prize, if it was bought. NEGATIVE

(19) **Control sentences**

a. If ticket #37 was bought, it would have to win a prize. POSITIVE

b. No ticket could win a prize, if it was bought. NEGATIVE

Crossing contexts and sentence types gave rise to $3 \times 4 = 12$ test items. 12 filler items were further included in the study to diversify the content of the sentences presented to participants. Filler items involved contexts similar to those used in the test items, but were followed by non-counterfactual sentences.

3.3 Procedure

In the instructions, participants were told that they would read short stories, followed by a sentence, and that their task would be to assess the extent to which the sentence was true or false in the context of the story. They were next introduced to the response scale used in the study: they were instructed to move the slider to the right if they judged the sentence as completely true, to the left if they judged it as completely false, and to the middle if they found it neither completely false, nor completely true. Participants were encouraged to use all the flexibility of the slider to represent at best their intuitions about each sentence. After the instructions, the experiment started with 2 (unannounced) practice trials and then continued with the 24 experimental items (12 test+12 filler), which were presented in random order.

3.4 Data analysis

Participants' ratings were coded as the position of the slider on the scale, from 0% for 'Completely false' to 100% for 'Completely true'. We analysed the data by modeling ratings using linear mixed-effects models fit by restricted maximum likelihood. Analyses were conducted using the `lme4` (Bates, Maechler & Bolker 2011) and `lmerTest` packages, and languageR libraries for the R statistics program.

3.5 Results

Figure 2 shows the mean ratings to the test items. Responses to the control conditions were as expected: participants uniformly accepted the POSITIVE sentences in the *all*-context and the NEGATIVE ones in the *none*-context (all ratings $> 88\%$), and they uniformly rejected the POSITIVE sentences in the *none*-context and the NEGATIVE ones in the *all*-context (all ratings $< 7\%$). Turning now to the critical conditions, the POSITIVE target sentences gave rise in the *mixed*-context to intermediate ratings ($M = 47\%$, 95% CI[50,44]), closer to the midpoint of the scale than their corresponding controls ($M = 38\%$, 95% CI[42,34]). On the other hand, the NEGATIVE target sentences gave rise in this same context to very low ratings ($M = 12\%$, 95% CI[16,8]), just like their corresponding controls ($M = 13\%$, 95% CI[17,9]). To evaluate the differences between POSITIVE and NEGATIVE sentences, we examined the effects of sentence type and status on participants' ratings in the *mixed*-context. The model included Sentence (2 levels: Positive, Negative), Status (2 levels: Target, Control) and their interaction as fixed effects, a random effect for subject and a random slope for Sentence per subject.[12] The model showed a main effect of Sentence (Negative<Positive, $\beta = -35$, $p < .001$), a main effect of Status (Control<Target, $\beta = -9$, $p < .001$) as well as a significant interaction between both factors ($\beta = 10$, $p < .001$) such that the difference in ratings between Target and Control was greater for the POSITIVE than the NEGATIVE sentences in the *mixed*-context.

3.6 Discussion

Results show that, in the critical *mixed*-context, simple counterfactuals (POSITIVE) received intermediate ratings whereas counterfactuals embedded under *no* (NEGATIVE) received very low ratings. If our interpretation of the task is correct, these results indicate that the former gave rise to gap judgments while the latter gave rise to judgments of falsity. These findings are in line with the predictions of the selectional theories while they are unexpected on the universal and homogeneity theories.

12 R pseudo-code describing the model: `Rating~Sentence*Status+(1+Sentence|Subject)`.

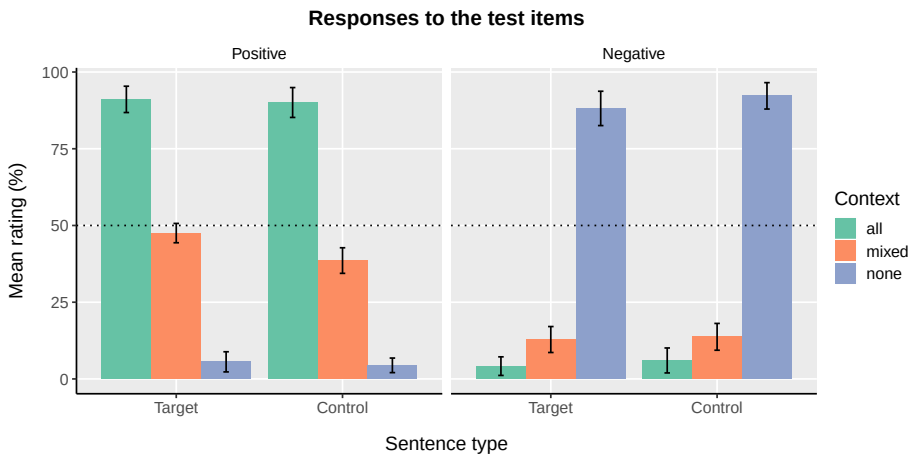

Figure 2 Mean rating to the test items in Experiment 1 as a function of the type of Context. The dotted line represents the midpoint of the response scale and error bars denote 95% confidence intervals.

4 Experiment 2

Experiment 2 aimed at assessing both the reliability of the results from Experiment 1 and the adequacy of our linking hypothesis.

4.1 Participants

80 participants were recruited through Prolific and were paid £1.2 for their participation. The data from all the participants were included in the analyses (36 female, average age 34.8 years). The consent and data collection procedures were the same as in Experiment 1.

4.2 Materials

The materials and method used in Experiment 2 were the same as in Experiment 1, except for the following two changes. First, the POSITIVE sentences in Experiment 2 were created using the frames in (20), where #X was a numeric value between 1 and 100 pseudo-randomly generated so as to be unique for each instance of these sentences. We made this modification to prevent participants from focusing on a particular ticket number as well as to make the choice of the ticket mentioned in these sentences look more random. In addition, we modified the formulation of the POSITIVE control in an attempt to provide a better baseline for the POSITIVE target.

(20) **Positive sentences**
Consider a random ticket, say ticket #X:

a. If ticket #X was bought, it would win a prize. Target

b. If ticket #X was bought, necessarily, it would win a prize. Control

Second, the content of the *mixed*-context was minimally altered so as to not make reference to a specific ratio (e.g., *half of the tickets*), as illustrated in (21). We made this modification to avoid an interpretation of intermediate ratings as matching the proportion of ticket bought (or the probability of a ticket being a winning ticket).

(21) **Mixed-context**
The tickets for the yellow raffle are now for sale. The yellow raffle works as follows. At the end of the ticket sales, there will be a random draw: only some of the tickets that have been bought will win a prize.

The rest of the design of Experiment 2 (number of test items, list of fillers, etc.) was identical to that of Experiment 1 in all relevant respects.

4.3 Procedure

The procedure was the same as in Experiment 1 (see Section 3.3 for details).

4.4 Data analysis

The data were analysed using the data analysis pipelines created to analyse the data from Experiment 1. The results from Experiments 1 & 2 are thus directly comparable.

4.5 Results

Figure 3 shows the mean ratings to the test items. The patterns of ratings for the control conditions (i.e., *none*-context and *all*-context) and the critical conditions (i.e., *mixed*-context) were essentially the same as those observed in Experiment 1. In particular, the POSITIVE target received a middle-range rating ($M = 46\%$, 95% CI[50,41]), closer to the midpoint of the scale than its control ($M = 35\%$, 95% CI[41,30]), while the NEGATIVE target received a low-range rating ($M = 13\%$, 95% CI[18,9]), just like its control ($M = 11\%$, 95% CI[15,6]). As in Experiment 1, we examined the effects and interaction of Sentence and Status on participants' ratings in the *mixed*-context. The model showed a main effect of Sentence (Negative<Positive, $\beta = -32$, $p < .001$), a main effect of Status (Control<Target, $\beta = -10$, $p < .001$) and a significant interaction between both factors ($\beta = 7.3$, $p < .05$).

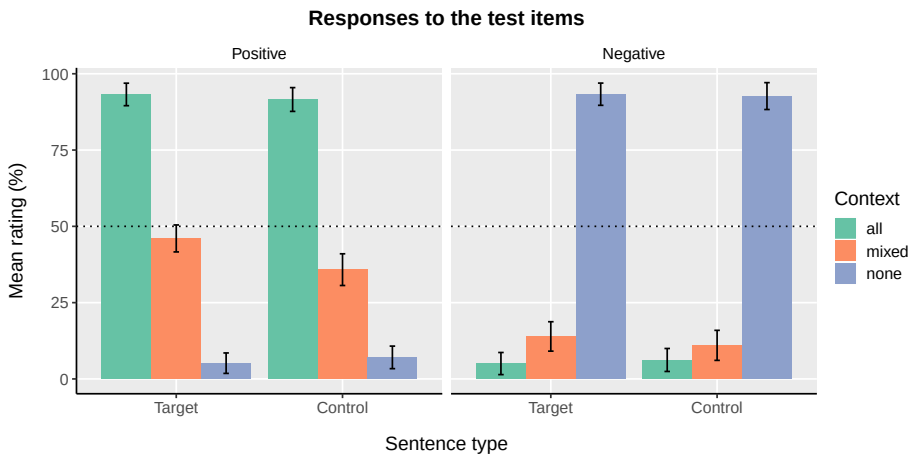

Figure 3 Mean rating to the test items in Experiment 2 as a function of the type of Context. The dotted line represents the midpoint of the response scale and error bars denote 95% confidence intervals.

4.6 Discussion

Experiment 2 yielded similar results as Experiment 1. Interestingly, we found that the POSITIVE target still received middle ratings in the novel *mixed*-context where no specific ratio was mentioned, unlike the *mixed*-context used in Experiment 1. We take this replication to support the hypothesis that the midpoint of the response scale was used by participants as a reference point for categorizing sentences that they perceived as neither completely true, nor completely false.

5 General discussion

In both experiments, we found clear middle ratings for the POSITIVE target in the *mixed*-context. These ratings were reliably different from those for their corresponding controls. This finding is in line with the predictions of selectional and homogeneity theories, while it is challenging for universal theories. In fact, it confirms the conclusion of several authors who have argued against universal theories for various kinds of conditionals (see for instance Klinedinst 2011).

In both experiments, we also found that the endorsement rate for the NEGATIVE target in the *mixed*-context was overall very low and no different from that of their corresponding control. This suggests that both target and control NEGATIVE sentences were essentially judged false in these contexts. This finding is consistent with selectional theories, but is a challenge for both homogeneity and universal theories.

In summary, our experimental results provide a clear argument for the selectional theory, and against both universal and homogeneity theories. At the very least, the latter have to supplement the semantics of counterfactuals with some additional mechanism to account for our data.

Before concluding, we should briefly emphasize a surprising aspect of our results. We found in both experiments that the POSITIVE control, involving either *have to* or *necessarily*, were rated significantly lower than the POSITIVE target. Nonetheless, ratings for these sentences were distinctly higher than what it should be if these sentences were robustly judged false by participants (compare with NEGATIVE control for instance). We leave this observation as an open puzzle for now.

6 Alternative theories

In this section, we briefly sketch two additional theoretical options which are in principle compatible with our results.

The implicature approach. Bassi & Bar-Lev (2016) propose an implicature-based account of bare conditionals. They don't discuss counterfactuals directly, but their approach can easily be extended to the latter. Bassi & Bar-Lev's idea is that conditionals have existential force on their basic meaning, as in (22), and that this meaning is strengthened to a universal one via implicature, as in (23). The details of how this implicature comes about are not important for us; we merely mark the strengthened sentence via an 'IMP' operator.

(22) a. If ticket #37 was bought, it would win a prize

b. $\exists w'$: $w' \in \textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]^{w,\preceq}),[\![\mathsf{C}]\!]^{w',\preceq}$ = true

(23) a. IMP [If ticket #37 was bought, it would win a prize]

b. $\forall w'$: $w' \in \textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]^{w,\preceq}),[\![\mathsf{C}]\!]^{w',\preceq}$ = true

In unembedded cases, the strengthened meaning tends to be the prominent one, if not the only possible one. But the basic meaning should resurface in environments where implicatures tend not to arise, like in downward entailing contexts. This predicts that sentences like (24) should mean that, for every ticket x, there is no closest world where ticket x is bought such that ticket x wins in that world. These truth conditions are false in our mixed contexts.

(24) a. None of the tickets would win a prize if it was bought.

b. $\neg\exists x[\exists w'$: $w' \in \textsc{max}_{w,\preceq}([\![\mathsf{P}x]\!]^{w,\preceq}),[\![\mathsf{Q}x]\!]^{w',\preceq}$ = true$]$

Hence, the implicature approach predicts our results for the NEGATIVE conditions. In addition, the intermediate ratings observed for the POSITIVE ones are in line with

what is generally found for scalar implicatures with this type of measures (see Tieu et al. (2019); Renans, Romoli, Makri, Tieu, de Vries, Folli & Tsoulas (2017); Marty, Chemla & Spector (2015) among others).[13]

In sum, just like the selectional approach, the implicature approach is compatible with our results. The two approaches could be distinguished via further embeddings, like for instance under disjunction, but we must leave an investigation of these other embeddings to future work.

Homogeneity and QUD. The second approach supplements the homogeneity theory with a pragmatic mechanism involving a question under discussion, making its predictions consistent with our data.[14]

This approach builds on an analogous idea in the literature on plural definites, which are the paradigmatic case for homogeneity-based accounts. The main idea is that propositions with extension gaps depend for their evaluation on the current Question Under Discussion (QUD). In particular, if the QUD of the context lumps the worlds where the proposition is undefined with those where it is true, then the proposition can be judged as 'true enough' (see Križ 2015; Križ 2016; Champollion, Bumford & Henderson 2019). To illustrate the point, consider (25) and assume that it is associated with a trivalent proposition which is true when all windows are open, false when none of them is, and undefined otherwise. The sentence is thus predicted to be undefined in a context in which only some of the windows are open. Now, imagine that the QUD of the context is whether *any* of the windows are open (e.g., imagine that a storm is coming, and we need to decide whether we should go back home to close any windows that might be open). The partition associated with this QUD lumps together the cases in which all windows are open and those in which only

13 One way to make this observation more precise is to hypothesize that, in cases where implicatures are possible (i.e., in upward entailing contexts) and the literal and strengthened meanings lead to conflicting responses, participants will tend to look for a middle ground, e.g., select an intermediate response (see Bar-Lev 2020 for similar discussion in relation to judgments about plural definites, for which he also provides an implicature-based approach).

14 Thanks to Lucas Champollion for extremely helpful discussion on this subsection.

some of them are. Given this QUD, (25) would be now judged as 'true enough'.[15]

(25) The windows are open.

At first blush, this idea could be extended to our scenario. The homogeneity theorist could hypothesize that, when judging our NEGATIVE target items, participants accommodated a QUD that lumps worlds where the counterfactual is undefined with worlds where the counterfactual is false. On this hypothesis, one could argue that our participants judged the sentence 'false enough' and marked it as false in the experiment. This is a conceivable strategy for the homogeneity approach. Yet it leaves open a number of substantial questions. First, why do almost all participants accommodate such a QUD (and not, say, a QUD that lumps undefinedness and true together)? Second, why is this QUD accommodated only for the *negative* conditions, and not for the POSITIVE ones? Overall, while this idea seems worth developing, it needs to be supplemented with a principled pragmatic story about why speakers accommodate the right QUD in the context. In principle, such an account could be tested experimentally, for instance by introducing a context where a QUD is stated explicitly. We must also leave an investigation of this idea to future work.

7 Conclusion

We reported on two experiments testing the predictions of three major families of theories of counterfactuals: universal, selectional, and homogeneity theories. The critical cases in our experiments were constructed so as to discriminate between the key predictions of these three theories. Our findings support selectional theories and challenge universal and homogeneity theories.

15 Another way whereby homogeneity can be removed has to do with the presence of certain items. Thus for instance, sentences like those in (i) exhibit the expected pattern given homogeneity: both (ia) and its negation, (ib), are judged neither true nor false in contexts in which some but not all of the students left. By contrast, their counterparts with *all* in (iia) and (iib) are both judged false in such contexts (Križ & Chemla 2015).

(i) a. The students left.

b. The students didn't leave.

(ii) a. All the students left.

b. All the students didn't leave.

We used in fact such a strategy to construct our control items for the positive cases, e.g., we used *necessarily*, which was supposed to play the same role as *all* (Schlenker 2004). Crucially, however, our targets did not contain any corresponding item that could serve as a homogeneity remover. In particular, we know from the literature on definites that quantifiers like *none* remove homogeneity only with respect to the argument position that they are filling. With a quantifier in subject position, a definite plural in object position still gives rise to homogeneity (see Križ 2015 for discussion).

References

Abrusán, Márta & Kriszta Szendröi. 2012. Experimenting with the king of france. In *Logic, Language and Meaning*, 102–111. Springer.

Alxatib, Sam & Jeff Pelletier. 2009. On the psychology of truth-gaps. In *International workshop on vagueness in communication*, 13–36. Springer.

Bar-Lev, Moshe. 2020. An Implicature account of Homogeneity and Non-maximality. *Linguistics and Philosophy* (to appear).

Barnes, Elizabeth & J. Robert G. Williams. 2011. A theory of metaphysical indeterminacy. In Karen Bennett & Dean W. Zimmerman (eds.), *Oxford studies in metaphysics volume 6*, 103–148. Oxford University Press.

Bassi, Itai & Moshe Bar-Lev. 2016. A unified existential semantics for bare conditionals. *Sinn und Bedeutung* 21(1). 125–142.

Bates, Douglas, Martin Maechler & Ben Bolker. 2011. Package 'lme4'. *Linear mixed-effects models using S4 classes* .

Champollion, Lucas, Dylan Bumford & Robert Henderson. 2019. Donkeys under discussion. *Semantics & Pragmatics* 12(1). 1–50.

Fine, Kit. 1975. Vagueness, truth and logic. *Synthese* 30(3-4). 265–300. doi:10.1007/BF00485047.

von Fintel, Kai. 1999. NPI licensing, strawson entailment, and context dependency. *Journal of Semantics* 16(2). 97–148.

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

Fox, Danny. 2012. Presupposition projection from quantificational sentences: trivalence, local accommodation, and presupposition strengthening. In Ivano Caponigro & Carlo Cecchetto (eds.), *From grammar to meaning: the spontenous logicality of language*, 201–232. Cambridge University Press.

van Fraassen, B. C. 1969. Presuppositions: Supervaluations and free logic. In K. Lambert (ed.), *The logical way of doing things*, 67–92. Yale University Press.

George, Benjamin. 2008. *Presupposition repairs: a static, trivalent approach to predicting projection*: UCLA MA thesis.

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

Kaplan, David. 1989. Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In Joseph Almog, John Perry & Howard Wettstein (eds.), *Themes from kaplan*, 481–564. Oxford: Oxford University Press.

Kaufmann, Stefan. 2017. The limit assumption. *Semantics and Pragmatics* 10(18). doi:10.3765/sp.10.18.

Klinedinst, Nathan. 2011. Quantified conditionals and conditional excluded middle.

Journal of Semantics 28. 149–170.
Kratzer, Angelika. 1986. Conditionals. In *Chicago Linguistics Society 22*, 1–15.
Kratzer, Angelika. 2012. *Modals and conditionals: New and revised perspectives*, vol. 36. Oxford University Press.
Križ, Manuel. 2015. *Aspects of homogeneity in the semantics of natural language*: University of Vienna PhD dissertation.
Križ, Manuel. 2016. Homogeneity, maximality, and all. *Journal of Semantics* 33(3). 493–539.
Križ, Manuel & Emmanuel Chemla. 2015. Two methods to find truth-value gaps and their application to the projection problem of homogeneity. *Natural Language Semantics* 23(3). 205–248.
Lewis, David K. 1973a. *Counterfactuals*. Cambridge, MA: Harvard University Press.
Lewis, David K. 1973b. Counterfactuals and comparative possibility. *Journal of Philosophical Logic* 2(4). 418–446.
Lewis, David K. 1979. Counterfactual dependence and time's arrow. *Noûs* 13(4). 455–476.
Mandelkern, Matthew. 2016. Dissatisfaction theory. In *Semantics and linguistic theory*, vol. 26, 1–50.
Marty, Paul, Emmanuel Chemla & Benjamin Spector. 2015. Phantom readings: The case of modified numerals. *Language, Cognition and Neuroscience* 30(4). 462–477.
Renans, Agata, Jacopo Romoli, Maria-Margarita Makri, Lyn Tieu, Hanna de Vries, Raffaella Folli & George Tsoulas. 2017. Abundance inference of pluralised mass nouns is an implicature: Evidence from Greek. *Glossa: A Journal of General Linguistics* 3(1). 103.
Ripley, David. 2009. Contradictions at the borders. In *International workshop on vagueness in communication*, 169–188.
Schlenker, Philippe. 2004. Conditionals as definite descriptions. *Research on language and computation* 2(3). 417–462.
Schwarz, Florian. 2016. False but slow: Evaluating statements with non-referring definites. *Journal of Semantics* 33(1). 177–214.
Serchuk, Phil, Ian Hargreaves & Richard Zach. 2011. Vagueness, logic and use: Four experimental studies on vagueness. *Mind & Language* 26(5). 540–573.
von Fintel, Kai. 1997. Bare plurals, bare conditionals, and only. *Journal of Semantics* 14(1). 1–56.
Stalnaker, Robert. 1968. A theory of conditionals. In N. Recher (ed.), *Studies in logical theory*, Oxford.
Stalnaker, Robert. 1984. *Inquiry*. MIT Press.
Stalnaker, Robert C. 1981. A defense of conditional excluded middle. In William

Harper, Robert C. Stalnaker & Glenn Pearce (eds.), *Ifs*, 87–104. Reidel.
Tieu, Lyn, Cory Bill & Jacopo Romoli. 2019. Homogeneity or implicature: An experimental investigation of free choice. *Semantics and Linguistic Theory* 29. 706–726.

Paul Marty
University College London
Psychology and Language Science
Chandler House, 2 Wakefield Street, London
United Kingdom
p.marty@ucl.ac.uk

Jacopo Romoli
University of Bergen
Department of Foreign Languages
HF-Bygget, Sydnesplassen 7
5020, Bergen
Norway
jacopo.romoli@uib.com

Paolo Santorio
Department of Philosophy
University of Maryland, College Park
4300 Chapel Drive, College Park, MD 20742
USA
paolosantorio@gmail.com