Undefined Counterfactuals: Experimental Evidence for Selection Semantics

Sonia Ramotowska
Heinrich Heine Universität Düsseldorf

Paul Marty
L-Università ta Malta

Jacopo Romoli
Heinrich Heine Universität Düsseldorf

Paolo Santorio
University of Maryland

Abstract Theories of counterfactuals agree on appealing to a relation of comparative similarity, but disagree on the quantificational force of counterfactuals. We report on three experiments testing the predictions of three main approaches: universal theories, homogeneity theories, and single-world selection theories (plus supervaluations over selection functions). To disentangle the predictions of these theories, we contrasted counterfactual sentences embedded under the scope of various quantifiers and tested speakers' intuitions about such sentences using graded truth value judgment tasks. Our results provide empirical support for selectional theories, while challenging the other two approaches. We end by discussing how a more recent alternative implicature-based approach to counterfactuals is also not in line with our results, and the connection between counterfactuals and related phenomena like plural definites.

Keywords: counterfactuals, homogeneity, supervaluations, undefinedness, semantics

Word count: 5349

Contents

1 Introduction

On all standard analyses, the semantics of counterfactual conditionals like (1) exploits a relation of comparative closeness between worlds (see Lewis 1973a,c, 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.

While there is agreement on the appeal to comparative closeness, there is less agreement on what kind of logical operator best models counterfactuals. The classical literature includes two competing views. On the analyses put forward by Lewis (1973a, 1973c) and Kratzer (2012), counterfactuals are universal quantifiers (roughly, over a domain of 'closest' antecedent-verifying worlds). On the selectional analysis put forward by Stalnaker (1968, 1981, 1984), counterfactuals are not strictly speaking quantifiers. Rather, they select the single closest antecedent-verifying world – somewhat analogously to singular definites. Cases where there is no such a single world are handled via supervaluations. More recently, a homogeneity-based analysis has emerged (see von Fintel 1997, Schlenker 2004, Križ 2015, Agha 2023; see also Willer 2022). On this third analysis, counterfactuals are universal quantifiers, as on the Lewis-Kratzer view, but they also come with a homogeneity requirement that demands that either all antecedent worlds make the consequent true, or that all of them make the consequent false.

All three theories capture some basic facts about counterfactual statements. But they also make different predictions regarding their truth conditions across contexts and/or across environments. The most obvious difference in predictions is between the universal theory and the other two. On the universal theory, counterfactuals like (1) are either true or false.[1] On both the selectional and the homogeneity theory, counterfactuals like (1) can be undefined in some contexts. Interestingly, the selectional and the homogeneity theory also differ, though in subtler ways. In particular, while they agree on all unembedded counterfactuals like (1), they disagree in their predictions about embeddings.

The goal of the present study is to single out cases where the three theories make different predictions, and test these cases experimentally. We report the results of three experiments. The first two experiments focus on unembedded counterfactuals and counterfactuals embedded under negative quantifiers. The third experiment

1 *Modulo* the presence of items that induce a failure of bivalence in the antecedent of the consequent, such as a definite phrase whose presupposition is not met in the context. The point is that counterfactuals themselves are never the source of undefinedness.

focuses on counterfactuals embedded under quantifiers of different strength, and addresses a potential confound in the first two experiments. As we discuss below, the results from these experiments offer clear empirical support for the selectional theory, whose predictions are fully compatible with the data. Conversely, they are challenging for both the quantificational and the homogeneity theory.

The rest of this paper is structured as follows. In §2, we discuss the three theories in more detail, as well as the predictions we target. After giving an overview of the experiments in §3, we report on the first two experiments in §4–5. In §6, we discuss a potential confound in the first two experiments, motivating our third experiment which we report in §7. In §8, we discuss the results in relation to the theoretical predictions. In §9, we briefly discuss the predictions of a fourth approach, the implicature approach. Finally, §10 concludes. Throughout the paper, we use the traditional notation ‘A $\Box\!\!\rightarrow$ C’ as shorthand for the counterfactual *If A, would C*.

2 Background: counterfactuals and undefinedness

2.1 Three theories of counterfactuals

All standard theories on counterfactuals agree on the general form of their truth conditions: A $\Box\!\!\rightarrow$ C is true just in case C is true in some relevant range of ‘closest’ or ‘most similar’ A-verifying worlds.[2] These truth conditions can be formalized by introducing a three-place relation of comparative closeness $\preceq_{w_0}$, which works as follows: $w_i \preceq_{w_0} w_k$ is true just in case w_i is closer than w_k to a ‘base’ world w_0.

While the foregoing is common ground among all main views, there is disagreement about the quantificational force of counterfactuals, and in fact on whether counterfactuals are quantificational at all. In this section, we review the three main options that have been developed in the literature.

2.1.1 The universal theory

On a first, classical theory, counterfactuals have universal quantificational force. Classical accounts such as Lewis’s (1973a, 1973c) and Kratzer’s (1986, 2012), as well as their descendants, adopt versions of this view. Abstracting away from some complications that are not relevant for us, the truth conditions we get on these theories are in (2).[3]

2 There are some well-known dynamic variants of the static accounts presented below: 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 sentences of interest.

3 Strictly speaking, neither Lewis nor Kratzer subscribe to the truth conditions in (2). This is because (2) involves the so-called limit assumption, i.e. the assumption that, for every world w and for every counterfactual antecedent A, there is a set of A-worlds that are closest to w. Famously, Lewis allows

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

For illustration, suppose that Maria considered flipping a coin yesterday at noon, but in the end didn't. 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).

(4) $[\![(3)]\!]^{w,\preceq}$ = true iff, for every w' s.t. w' is a $\preceq$-closest worlds where Maria flipped the coin, the coin landed heads in w'.

Suppose that the closest worlds to the actual world involve a mixture of heads and tails-worlds. (This is plausible on any construal of closeness, assuming that coin flips are chancy events.) Then (3) is predicted to be false in this context.

2.1.2 The selectional theory

A second classical theory denies that counterfactuals are quantificational. On this theory—which we call 'selectional'—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.[4] The schematic truth conditions of a counterfactual are in (5):

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

that, in some cases, there can be infinite chains of antecedent-worlds that are increasingly closer to the 'base' world. (Besides Lewis, see Kaufmann 2017 for discussion of the limit assumption.) For simplicity, here we consider only the version of the universal theory that includes the limit assumption. Dropping the limit assumption will not improve the predictions of the universal theory for all the cases we consider. Notice, moreover, that the two main competing theories we consider—the selectional theory and the homogeneity theory—both entail the limit assumption.

4 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 $s(w,A) \in [\![A']\!] = s(w,A') \in [\![A]\!]$

The selectional theory does not appeal directly to a relation of comparative closeness, but talk of selection functions can be rephrased in these terms (*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.[5]

Without supplementation, the selectional theory runs into an obvious difficulty. The selectional semantics requires that, for every antecedent A and every world w, there is a single closest A-world to w. As examples like (3) above 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.[6]

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.[7]

> A is **determinately true** at c iff, for all $\langle w_c, s\rangle$ such that s is a candidate selection function at c, $[\![\mathrm{A}]\!]^{w,s}$ is true.
>
> A is **determinately false** at c iff, for all $\langle w_c, s\rangle$ such that s is a candidate selection function at c, $[\![\mathrm{A}]\!]^{w,s}$ is false.
>
> A is **undefined** at c iff A is neither determinately true nor determinately false at c.

5 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.

6 Supervaluations were introduced in van Fraassen (1969). We also refer the reader to Fine 1975 for a classical use of supervaluations to model vagueness.

7 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.

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. As we shall latter see, this feature plays 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 the sentence in (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 first difference in predictions between the universal and the selectional theory.

2.1.3 Universal vs selectional theories: the classical debate

The classical debate on counterfactuals between Lewis and Stalnaker centered on the comparison between the universal and the selectional theory. The selectional theory yields a strictly stronger logic than the universal theories. Hence it vindicates some extra logical principles. In particular, the early debate focused on the conditional counterpart of the Excluded Middle principle:

Conditional Excluded Middle (CEM). $\vDash (\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}) \vee (\mathsf{A} \mathrel{\Box\!\!\rightarrow} \neg \mathsf{B})$

On the selectional theory, CEM is valid. (The reason, informally: supposing A takes us to a single world. In that world, either B is true, or B is false. In the former case, $\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}$ is true, and in the latter $\mathsf{A} \mathrel{\Box\!\!\rightarrow} \neg\mathsf{B}$ is true; either way, their disjuction is true.) Conversely, on universal theories CEM is invalid.

Some principles that are closely related to CEM will be especially important for our purposes. One of them is what we call 'Negation Swap':

Negation swap. $\neg(\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}) \mathrel{=\!\!|\!\vDash} \mathsf{A} \mathrel{\Box\!\!\rightarrow} \neg\mathsf{B}$

Negation swap says that we can bring negation in and out of counterfactuals and obtain equivalent sentences. Negation swap follows from CEM, via minor side principles. Accordingly, it is validated by the selectional theory, and invalidated by the universal theory.

As even Lewis conceded (1973a: p. 80), CEM and related principles appear to be valid. For a simple illustration, consider the sentences in (6):

(6) a. It it not the case that, if Mary had flipped the coin, the coin would have landed heads.

b. If Mary had flipped the coin, the coin would not have landed heads.

(6a) and (6b) sound equivalent. This is evidence for the selectional theory, and against the universal theory. (For reasons of space, we omit arguments for the selectional theory that don't directly relate to negation; see Higginbotham 2003, Klinedinst 2011, Khoo 2022, Santorio 2022, among many.)

On the other side, there are two main arguments for the universal theory, one empirical and one conceptual. The empirical argument is that counterfactuals are the duals of *might*-counterfactuals. Duality is the principle that *would* and *might*-counterfactuals are duals in the same way as necessity and possibility operators in modal logic:

Duality. $\vDash (\mathsf{A} \Diamond\!\!\rightarrow \mathsf{B}) \leftrightarrow \neg(\mathsf{A} \Box\!\!\rightarrow \neg\mathsf{B})$

By validating Duality, the universal theory correctly predicts that pairs of *would* and *might*-counterfactuals like the one in (7) are incompatible.

(7) 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.

Conversely, the selectional theory struggles with this prediction.

The conceptual argument is that the selectional theory imposes unrealistic requirements on the selection function. The selectional theory requires that, for every antecedent and for every world of evaluation, we should be able to find a single closest world that verifies the antecedent. For Lewis, this is exceedingly implausible. Any plausible relation of comparative similarity must allow for ties.

> [T]he cost of [validating CEM] is too much. However little there is to choose for closeness between worlds where Bizet and Verdi are compatriots by both being Italian and worlds where they are compatriots by both being French, the selection function still must choose. I do not think it can choose—not if it is based entirely on comparative similarity, anyhow. Comparative similarity permits ties, and Stalnaker's selection function does not. Lewis (1973a: p. 80)

Supervaluations might help with this worry, but they don't completely remove it. One way of understanding Lewis's concern is this: nothing in a concrete communicative situation fixes what single-world selection function we use to evaluate a counterfactual. Consider a context where (3) is uttered: if the coin is fair, there is no way of determining which selection function is used to evaluate the sentence. Supervaluations do allay this concern. It can be indeterminate which selection function is fixed by the context, hence a concrete situation of utterance doesn't have to determine a unique selection function. But Lewis's concern might be understood in a second way: why should grammar require that conditional antecedents latch onto a scenario that is determinate up to the last detail? That seems an unattainable demand. Supervaluations do not allay this second worry, since the basic semantics of counterfactuals does require that a single world is selected.

2.1.4 The homogeneity theory

The homogeneity theory (von Fintel 1997, Schlenker 2004, Križ 2015, Agha 2023; see also Willer 2022) is an attempt at resolving the tension between universal and selectional accounts. Accordingly, they share features with both. Like the universal theory, it treats counterfactuals as universal quantifiers over closest antecedent worlds. Like the selectional theory, it assumes that some counterfactuals will be undefined. Crucially, though, the undefinedness has a different origin. On the homogeneity theory, the semantics of counterfactuals includes a homogeneity requirement: on top of Lewis-style universal truth conditions, A $\Box\!\!\rightarrow$ B requires that either all closest A-worlds are B-worlds, or all closest A-worlds are non-B-worlds. For illustration, consider:

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

(3) is predicted to be true just in case all closest Maria-flipping worlds are heads-worlds, false just in case no closest Maria-flipping worlds are heads-worlds, and indeterminate otherwise. More generally, here are the schematic truth conditions for a counterfactual on this account:

(8) $$[\![\mathrm{A} \Box\!\!\rightarrow \mathrm{B}]\!]^{w,\preceq} = \begin{cases} \underline{\text{defined}} \text{ iff either } \forall w'\text{: } w' \in \textsc{max}_{w,\preceq}([\![\mathrm{A}]\!]^{w,\preceq}),\ [\![\mathrm{B}]\!]^{w',\preceq} = \text{true} \\ \qquad\qquad \text{or } \forall w'\text{: } w' \in \textsc{max}_{w,\preceq}([\![\mathrm{A}]\!]^{w,\preceq}),\ [\![\mathrm{B}]\!]^{w',\preceq} = \text{false} \\ \underline{\text{true}} \text{ iff } \forall w'\text{: } w' \in \textsc{max}_{w,\preceq}([\![\mathrm{A}]\!]^{w,\preceq}),\ [\![\mathrm{B}]\!]^{w',\preceq} = \text{true} \end{cases}$$

The homogeneity theory is in part motivated by a suggestive analogy with the behavior of plural definite phrases like *the girls* (as emphasized by Schlenker 2004). It is well-known that plural definites display homogeneity. Consider (9):

(9) The girls danced.

(9) is judged true if all the girls danced, false if none of them did, and has an undefined status when some girls danced and some girls didn't. The homogeneity theory suggests that the same phenomenon responsible for the trivalent behavior of (9) is responsible for the indeterminacy of (3).

The homogeneity theory appears to give us the best of both worlds. Thanks to its appeal to a definedness condition, it can vindicate the logical properties that are also vindicated by the selectional theory. Given that the homogeneity theory treats conditionals as trivalent, it needs to define a notion of consequence that makes room for undefined sentences. The natural option is to appeal to Strawson Entailment (see von Fintel 1999), defined informally as follows:

> **Strawson entailment**
> $A_1, \ldots, A_n \vDash B$ iff, for all contexts c such that $A_1 \ldots, A_n$ are defined and true at c, and B is defined at c, B is also true at c.

Both the Negation Swap inferences and Conditional Excluded Middle are Strawson-valid.[8] At the same time, the homogeneity theory also has the advantages of the universal theory. It correctly predicts that *would*-counterfactuals and *might*-counterfactuals are duals, and it escapes the conceptual worries raised by single-world selection.

2.2 Selection vs homogeneity: comparing predictions

Both the selectional and the homogenity theory are trivalent: they treat some counterfactuals as undefined. As we observed, they yield identical predictions for our basic example (3), repeated below:

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

In fact, this symmetry in predictions is quite general. 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 counterfactuals. $A \Box\!\!\rightarrow B$ is predicted to be true/false/indeterminate by the selectional theory just in case it is predicted to be true/false/indeterminate by the homogeneity theory. Crucially though, this symmetry only holds for unembedded counterfactuals. Selectional and homogeneity theories predict differences in embeddings, because of differences in the way they treat undefinedness.

On the selectional theory, the compositional semantics is fully bivalent. Undefinedness only emerges at a 'global' level, when we define truth at a context. Specifically, 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 the compositional stages, the semantics is indistiguishable from a standard bivalent theory.

On the homogeneity theory, undefinedness emerges at a 'local' level, i.e. during composition. There are subsentential constituents whose semantic value is undefined. As a result, on the homogeneity theory, we need a projection algorithm

8 Notice though that Conditional Excluded Middle will still have undefined instances; see Križ 2015 for discussion of this point with regard to sentences involving plural definites.

to determine in what way the undefinedness of simpler expressions affects the definedness status of complex expressions. As we will see, this produces some key differences in predictions. (This also means that, to make predictions about complex sentences, we need an algorithm to determine how homogeneity projects; see Križ 2015, and Križ & Chemla 2015 for discussion. However, as we discuss below, the issue that we're interested in is orthogonal to what particular projection algorithm we choose.)

For illustration of the issue, let us start with embeddings under negative universal determiner phrases like *no ticket*. To 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 (1):[9]

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

(1) 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 present scenario, all candidate selection functions map the world of evaluation and the antecedent of (1) 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) (1) is undefined at c iff for some candidate selection functions s_1 and s_2 compatible with c, $[\![(1)]\!]^{w,s_1}$ = true and $[\![(1)]\!]^{w,s_2}$ = false

Since there are two such selection functions here, (1) is predicted to be undefined at the relevant context. And here is the prediction of the homogeneity theory:

9 As *[names omitted for anonymous review]* have independently pointed out to us, (1) is not a contrary-to-fact conditional strictly speaking, but rather a so-called future-less-vivid conditional (see Iatridou 2000). The antecedent concerns a future event and is seemingly compatible with facts at the time of utteranc. We are assuming that *would* has analogous quantificational force in future-less-vivid and contrary-to-fact conditionals. As we discuss below, our experiments test both types of conditionals, without finding substantial differences between the two.

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

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

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

That is, (1) 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, (1) is again predicted to be undefined.

Holding fixed the lottery scenario above, let us now consider a more complex sentence like (12), where the conditional is embedded under a negative quantifier:

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

Unlike (1), (12) allows us to pull 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 lottery, 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 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 for all candidate selection functions take us to a win-some-lose-some world and, on all of them, (12) is evaluated as false. As a result, the supervaluational theory predicts that (12) has a determinate truth value: it is (determinately) false. Everything else being equal, the homogeneity theory makes a different prediction. Since this theory exploits a trivalent compositional semantics, we first need to determine the definedness conditions for (12) on a compositional basis. Note here that (12) has the structure in (14):

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

From our discussion of (1), 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.

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

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

THEORY	Example (1)	Example (12)
Universal	false	true
Selectional	undefined	false
Homogeneity	undefined	undefined

Table 1 Predictions of the three theories of counterfactuals that we presented for the unembedded case in (1) and the negative embedded case in (12) in the win-some-lose-some lottery scenarios.

- **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 undefined for all objects in the domain. So, no matter what projection algorithm we choose, (12) is predicted to be undefined.

3 Overview of the experiments

We have introduced three families of theories of counterfactuals: universal, selectional, and homogeneity theories. Selectional and homogeneity theories both predict that some counterfactuals are undefined. In particular, they both predict unembedded counterfactuals like (1) to be undefined in the win-some-lose-some lottery scenarios. But these two theories disagree for some embeddings. In particular, they disagree about examples where counterfactuals are embedded under negative determiner phrases: in the scenarios of interest, a sentence like (12) is predicted to be false on selectional theories, but undefined on homogeneity theories. Finally, universal theories make different predictions all the way in predicting (1) to be false and (12) to be true in the scenarios of interest. Under this approach, in the win-some-lose-some lottery scenarios, it is false that for each closest world where a ticket is bought, it wins a prize and it is true that there is no ticket that would win a prize in every such closest world. These predictions are summarized in Table 1.

In the following sections, we report on three experiments designed to test these predictions and adjudicate between the three theories at hand. All three experiments involved a graded truth value judgment task 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 (1) and (12), repeated below, were paired with contexts in which only part of the tickets bought would win a prize, as in the example item in Figure 1.

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

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

We hypothesized that, if these items give rise to gappy 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.[10]

Experiment 1 tested the predictions in Table 1 by investigating speakers' truth-value intuitions about unembedded and embedded future-less-vivid conditionals. Experiment 2 tested these same predictions using novel materials to assess both the reliability of the results from Experiment 1 and the adequacy of our linking hypothesis. Finally, Experiment 3 expanded the range of cases under investigation by testing real counterfactuals embedded under various quantifier phrases while manipulating the Question under Discussion associated with each raffle scenario to further control for potential 'non-maximality' effects (see Section 6 for discussion of these effects). The data and analysis scripts for all three experiments are available on the Open Science Framework Platform (OSF).

4 Experiment 1

4.1 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

10 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). Beyond graded acceptability judgments, 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 & Szendrői 2012, Alxatib & Pelletier 2009, Tieu et al. 2019) and multiple unordered choices (Serchuk et al. 2011).

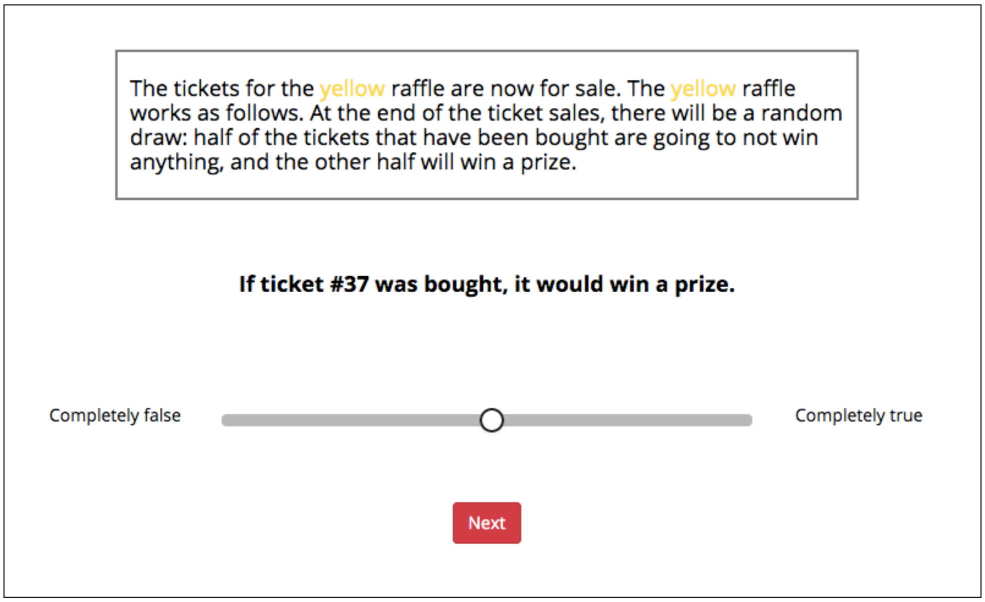


Figure 1 Example item illustrating the general display seen by the participants in our experiments (see Figure 4 for refinements regarding Experiment 3). This item is an example of a POSITIVE target sentence in the *mixed*-context in Experiment 1.

consent. All data were collected and stored in accordance with the provisions of the General Data Protection Regulation.

4.2 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). We used so-called 'future-less-vivid' counterfactuals as opposed to 'real' counterfactuals in order to keep the context simpler. We come back to this point below in Experiment 3, where we tested real counterfactuals. 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.

4.3 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 24 experimental items (12 test+12 filler), which were presented in random order.

4.4 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 et al. 2015) and `lmerTest` (Kuznetsova et al. 2017) packages for the R statistics program (R Core Team 2021).

4.5 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[44,50]), closer to the midpoint of the scale than their corresponding controls ($M = 38\%$, 95% CI[34,42]). On the other hand, the NEGATIVE target sentences gave rise in this same context to very low ratings ($M = 12\%$, 95% CI[8,16]), just like their corresponding controls ($M = 13\%$, 95% CI[9,17]). 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 by Subject.[11] The model showed an effect of Sentence ($\beta = -35$, $p < .001$), Status ($\beta = -9$, $p < .001$) as well as a significant interaction between the two ($\beta = 10$, $p < .001$) such that the difference in ratings between Target and Control was greater for the POSITIVE than the NEGATIVE sentences in the *mixed*-context.

4.6 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 theory, while they are unexpected on the universal and the homogeneity theory.

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

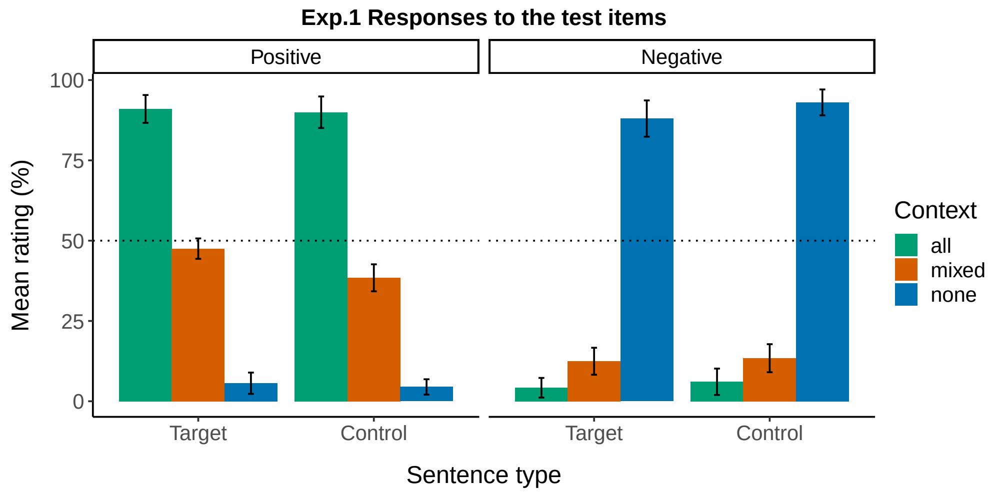


Figure 2 Mean rating to the test items in Experiment 1 as a function of the type of Context. The dotted line represents the midpoint of the response scale and error bars denote 95% confidence intervals.

Although the control POSITIVE sentences received significantly lower rating in *mixed*-context than POSITIVE counterfactuals, we should note that participants did not uniformly reject control sentences (38% acceptance rate). We address this issue by modifying the control sentences in Experiment 2.

5 Experiment 2

5.1 Participants

80 novel 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.

5.2 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, so as to make it less similar to the POSITIVE target, in an attempt to provide a better clearly false baseline.

(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.

5.3 Procedure

The procedure was the same as in Experiment 1 (see Section 4.3 for details).

5.4 Data analysis

The data were analysed using the data analysis pipelines from Experiment 1. The results from Experiments 1 and 2 are thus directly comparable.

5.5 Results

Figure 3 shows the mean ratings to the test items. The patterns of ratings for the control and target conditions were essentially the same as those observed in Experiment 1. In particular, the POSITIVE target received a middle-range rating ($M = 46\%$, 95% CI[41,50]), closer to the midpoint of the scale than its control ($M = 35\%$, 95% CI[30,41]), while the NEGATIVE target received a low-range rating ($M = 13\%$, 95% CI[9,18]), just like its control ($M = 11\%$, 95% CI[6,15]). As in Experiment 1, we examined the effects and interaction of Sentence and Status on participants' ratings in the *mixed*-context. The model showed an effect of Sentence ($\beta = -32$, $p < .001$), Status ($\beta = -10$, $p < .001$) and a significant interaction between the two in the same direction as before ($\beta = 7.3$, $p < .05$).

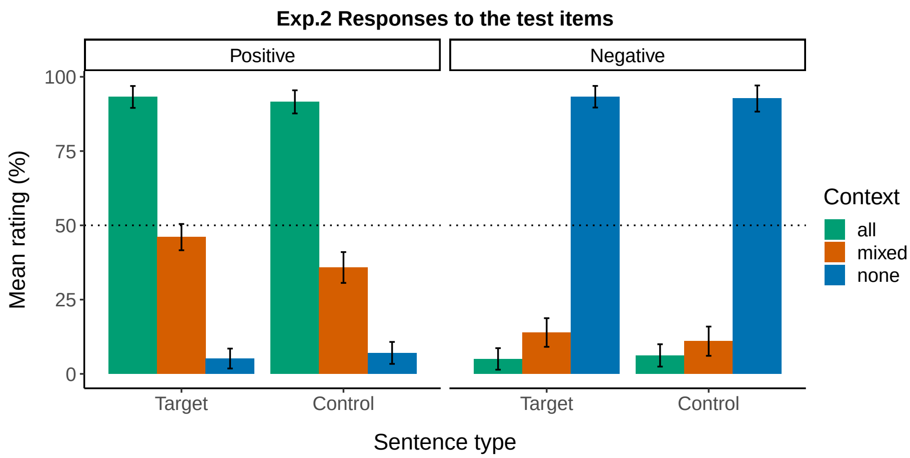


Figure 3 Mean rating to the test items in Experiment 2 as a function of the type of Context. The dotted line represents the midpoint of the response scale and error bars denote 95% confidence intervals.

5.6 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.

In both experiments, the POSITIVE controls, involving either *have to* or *necessarily*, were rated significantly lower than the POSITIVE target. Nonetheless, we should note that the ratings for these sentences were still distinctly higher than what it should be if these sentences were robustly judged false by participants (as compared with NEGATIVE controls for instance). In the next experiment, we move away from targetting judgments about undefinedness and hence from the need of false baselines of the type we used in Exp.1 and Exp.2. We turn to describe Exp.3 in the next section.

6 Summary and motivation for Experiment 3

6.1 Interim summary

In the first two experiments, we tested the predictions of three main approaches to counterfactuals. Specifically, we compared cases like (1) to cases like (12), repeated below for convenience, in contexts where tickets are selected by a random draw and only some tickets among those which were bought are winning—what we called a 'mixed lottery' scenario.

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

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

As discussed, the three approaches make distinct predictions for these cases, especially for (12). This sentence is predicted to be true by the universal approach, undefined by the homogeneity approach (since the prejacent of *no ticket* is undefined, for each ticket in the domain of the quantifier), and simply false by the selectional approach (since, for all choices of selection function, some tickets are going to win). Participants in the first two experiments overwhelmingly rejected sentences like (12) in such contexts, in line with the selectional approach and contra the other two approaches.

6.2 Homogeneity accounts and QuD

While we think that the challenge for the universal approach stands, the homogeneity approach could still be made compatible with our results if we supplement it with a pragmatic component sensitive to what is relevant in the context. To illustrate, consider plural definites like *the windows* in (22), which are often taken to be the paradigmatic case for homogeneity-based accounts. Under the homogeneity approach, analogously to counterfactual statements, (22) is predicted to be true if all windows are open, false if none of them are and undefined otherwise.

(22) The windows are open.

It has long been observed, however, that sentences like (22) can sometimes be judged true in a context where a few of the windows are closed and, similarly, can be judged false even if a couple of windows are open. Consider the following example:

> *Context.* There is a big storm coming and the speaker just realized that the windows in the two bedrooms of their house (though not the others) are open.

(23) The windows are open. We must go back home and close them!

This phenomenon, often referred to as ‘non-maximality’, is problematic for the homogeneity approach sketched above, which predicts that (23) is undefined unless all windows are open or all windows are closed.

One common strategy for homogeneity theorists involves including an extra layer of pragmatic interpretation, sensitive to relevance. This layer, which is generally modelled with Questions under Discussion (QuD) (Roberts 1987 and much following work), can push semantic undefinedness towards truth or falsity depending on the context. The idea goes as follows. The QuD introduces a relevant partition of epistemically live worlds. In some cases, the partition associated with the QuD of the context lumps the worlds where the proposition is undefined with those where it is true. In such cases, the proposition can be judged as ‘true enough’. The same holds for cases of falsity (see Križ 2015, Križ 2016, Champollion et al. 2019). To illustrate the point, consider (22) again, which is predicted to be undefined in a context in which only some of the windows are open. Take now a context where the relevant QUD is whether *any* of the windows are open. For example, take just the context of (23): a storm is coming, and we need to determine whether our home will get flooded with rain. The partition associated with this QuD lumps together the cases in which all windows are open and those in which only some of them are. Given this QuD, (23) would be now judged as ‘true enough’, thus accounting for the widespread intuition that the sentence is indeed acceptable in the given context.[12]

12 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 such a strategy to construct our control items for the positive cases in Experiment 2, i.e., we used *necessarily*, which was supposed to play the same role as *all* (see Schlenker 2004). Crucially, 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).

This idea can be extended to counterfactuals and to our scenario.[13] In particular, while semantically undefined, (12) may be judged as 'false enough' in a context where the partition associated with the QuD does not distinguish between false and undefined cases. For example, (12) may be judged 'false enough' in a context in which it is relevant whether each ticket bought *has a chance to win* (what we call an 'existential' QuD) in contrast to a context whether it is relevant whether each ticket *is guaranteed to win* (a 'universal' QuD). We did not control for potential QuD effects in the first two experiments, hence we cannot exclude the possibility that participants understood (12) against an implicit existential QuD, which is a plausible QuD type given how lotteries generally work, and judged it as false for this reason. This hypothesis allows the homogeneity approach to account for the asymmetry between the positive and negative cases observed in our first two experiments. Therefore, when supplemented with the QuD ingredient, the homogeneity approach remains compatible with our results thus far.

7 Experiment 3

In order to address this potential confound, we carried out a third experiment improving the design and materials of the first two experiments in three significant ways. First, we moved from future-less-vivid conditionals to real counterfactuals. Second, in addition to cases like (12), we tested the corresponding sentences with *every*, *not every* and *some*, as exemplified in (24). The labels used to refer to these four sentence types classify them along two dimensions, namely polarity (positive vs. negative) and strength (strong vs. weak).

(24)
a. Every one of these tickets would have won if it had been bought. POS-STRONG
b. None of these tickets would have won if it had been bought. NEG-STRONG
c. Not every one of these tickets would have won if it had been bought. NEG-WEAK
d. Some of these tickets would have won if they had been bought. POS-WEAK

Third, building on a recent study on plural definites by Augurzky et al. 2023, we experimentally manipulated the QuD associated with each scenario to control for the potential confound we described. The predictions tested in Experiment 3 are summarised in Table 2, focusing here on the selectional and homogeneity approaches. In the target 'mixed lottery' scenarios, where some randomly drawn tickets win, the homogeneity approach predicts semantic undefinedness for all four sentence types in (24). This prediction is a consequence of the fact that the homogeneity approach handles undefinedness locally, as shown in (14). We have

13 While the term 'non-maximality' is not much used in the literature on counterfactuals, Križ (2015) has argued that the phenomenon arises with counterfactuals just as well. In particular he points out that, intuitively, when evaluating counterfactuals, we tend to leave out far-fetched possibilities.

shown this above for the case of negative quantifiers: regardless of whether one assumes undefinedness to project universally or existentially, the whole sentence ends up undefined in our scenario. Given that the sentence is undefined even assuming the weakest possible projection of undefinedness, it is easy to see that the same prediction extends to all four quantifiers above, whether positive or negative.

Having said that, as soon as we take into account the effect of the QuD, the predictions of this approach change. In 'mixed lottery' scenario, *every*/*some* tickets are such that they have *a chance to win* is true but it is false that they are *guaranteed to win*. Therefore, the positive quantifiers are predicted to be true when the 'existential' QuD is relevant in the context and false if the 'universal' one is relevant. The opposite holds for negative quantifiers. *None*/*not every* ticket is such that it is *guaranteed to win* is true in 'mixed lottery' scenario, while *none*/*not every* ticket is such that it has *a chance to win* is false. This means that the negative quantifiers are predicted to be true when the universal QUD is relevant and false when the existential QUD is.

The selectional approach, on the other hand, expects no effect of QUD, but rather it predits an effect of quantifier strength: sentences with universal quantifiers like *every* and *no* are predicted to be false in the 'mixed lottery' scenarios. In both cases, all candidate selection functions map counterfactuals to a win-some-lose-some world, and on all of them universal quantifiers are evaluated as false. In contrast, sentences with non-universal quantifiers like *some* and *not every* are predicted to be true. In 'mixed lottery' scenarios such sentences are also mapped to a win-some-lose-some world but in all these worlds they are evaluated as true.

	Homogeneity Theory			
	every	*some*	*none*	*not every*
Universal QUD	false	false	true	true
Existential QUD	true	true	false	false

	Selectional Theory			
	every	*some*	*none*	*not every*
Universal QUD	false	true	false	true
Existential QUD	false	true	false	true

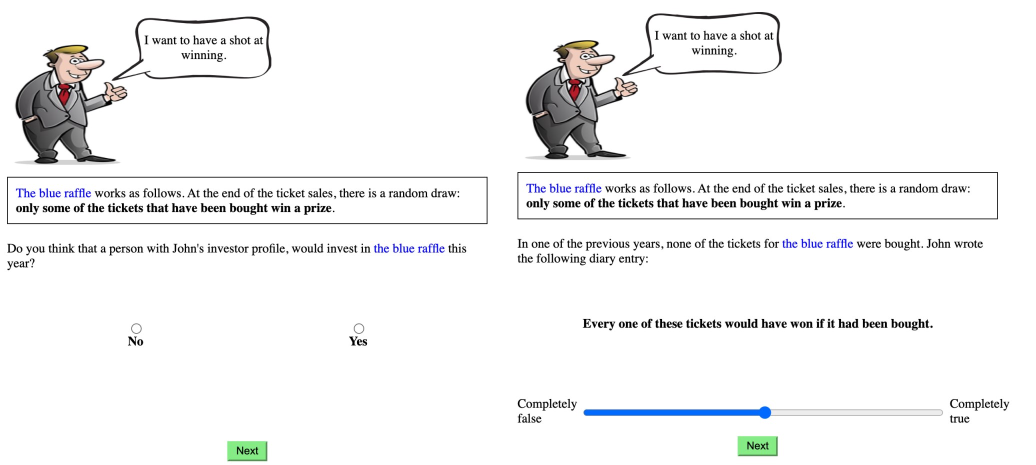
Table 2 Predictions of the homogeneity and selectional theories for the four sentence types tested in Experiment 3 in the win-some-lose-some lottery scenarios (i.e., the *mixed*-contexts).

7.1 Participants

100 novel participants were recruited through Prolific and were paid £12/*hr* for their participation. Due to technical issues, the data from 3 participants failed to be recorded. Thus, in effect, the data of 97 participants were considered for data treatment (48 female, mean age 42 years). The consent and data collection procedures were the same as in Experiment 1 and 2.

7.2 Materials and design

Each item displayed a lottery scenario together with the picture of a fictional character and involved a two-step response procedure: (i) a QuD check followed by (ii) a graded truth value judgment task, as shown in Figure 4.



Task 1: QuD check   Task 2: Graded TVJT

Figure 4 Example item illustrating the two-step response procedure in Experiment 3. This item is an instance of the EX-QuD condition for the POS-STRONG sentences in the *mixed*-contexts.

The lottery scenarios were *all*, *mixed* and *none* contexts, similar in essence to those previously used in Experiment 1 and 2, as exemplified in (25)-(27).

(25) **All-context**
The orange raffle works as follows. The organizers want all participants to be content: at the end of the ticket sales, every ticket that has been bought wins a prize.

(26) **Mixed-context**
The blue raffle works as follows. At the end of the ticket sales, there is a random draw: only some of the tickets that have been bought win a prize.

(27) **None-context**
The red raffle is rigged: at the end of the ticket sales, none of the tickets that have been bought win a prize.

The QuD was manipulated between subjects through the investor profile of the character displayed on the screen during the experiment. Specifically, each participant was introduced to one of two fictional characters, John or Bill, who had a different attitude toward winning in lotteries. In the universal QUD condition (henceforth, U-QuD), the character's goal was to win the lottery each and every single time, thus making relevant whether *each ticket is guaranteed to win.* This requirement could only be satisfied in the *all*-contexts, where every ticket was expected to win. By contrast, in the existential QuD condition (henceforth, EX-QuD), the character only wanted to have at least a shot at winning, thereby making relevant whether *each ticket has at least a chance to win.* This requirement was satisfied both in the *all*-contexts and in the *mixed*-contexts, where either some or all of the tickets were expected to win. To maintain the QuD salient throughout the experiment, the relevant character and his motto was displayed on every item, right above the lottery scenario (see Figure 4).

Sentences in the graded truth value judgment task were real counterfactuals embedded in the scope of one of four quantifiers (i.e., *every*, *none*, *not every*, and *some*) that differed from one another in terms of polarity (positive vs. negative), strength (strong vs. weak) or both. Target sentences involved the modal *would*, thus giving rise to the four sentence types exemplified in (24). The type of context, the polarity and the strength of the quantifier were all manipulated within subjects, resulting in a $3 \times 2 \times 2$ factorial design for each of the two QuD conditions. Crossing these three factors gave rise to 12 test items. 12 filler items were further included in the study to diversify the content of the sentences presented to participants. Filler items were similar in all respects to the test items except for the sentences presented in the graded truth value judgment task: filler sentences were obtained from the target sentences by replacing *would* with *could* (e.g., *Every on of these tickets could have won if it had been bought*).

7.3 Procedure

Upon entering the study, participants were randomly assigned one of the two QUD conditions, U-QuD or EX-QuD. In the instructions, participants were presented with a short background story describing the investor profile of the character associated with the QUD condition assigned to them. They were told that the character had been studying lotteries for some years and had written a diary entry for each of them. To make the counterfactual form of the sentences in the study prominent, participants were further told that the popularity of lotteries varied from one year to another and that, in some years, none of the tickets had been bought. Participants were next introduced to the two-step response procedure used in the study. They were told that they would read short stories, each of which would be followed by a question about the character's investment habits and then by a sentence coming from the character's diary entries. Instructions regarding the use of the graded response scale were the same as in Experiment 1 and 2.

After the instructions, the experiment started with 2 unannounced practice trials and then continued with 24 experimental items (12 test + 12 filler), presented in random order. At the beginning of each trial, participants were presented with a picture of the character and a lottery scenario. After reading the scenario, they had to complete two tasks successively (see Figure 4). First, they had to decide whether a person with the character's investor profile would invest in the lottery described in the scenario by clicking one of two response buttons labelled 'Yes' and 'No', respectively. Second, they were presented with a test sentence and had to assess the extent to which it is true in the given scenario using a slider going from 'Completely false' to 'Completely true' (as in Experiments 1 and 2). The character's picture and the relevant scenario remained on the screen throughout the trial.

7.4 Data treatment

We applied two exclusion criteria. First, we reasoned that, if participants understood the instructions correctly, they should generally answer 'No' in the QuD checks involving the *none*-contexts and 'Yes' in those involving the *all*-contexts, irrespective of the QuD manipulation. For these reasons, we excluded participants whose overall accuracy to the QuD checks in the control *all* and *none* contexts was lower than 80%. Following this procedure, the responses from 10 participants were excluded from all subsequent analyses. Of the remaining 87 participants, 43 completed the EX-QuD version of the experiment and 44 the U-QuD version.

The second criterion pertains to the analysis of participants' responses in the graded truth value judgment task. Consistent with our purposes, we excluded from this analysis the *all* and *none* trials in which participants responded incorrectly to the QuD checks and the *mixed* trials in which participants' responses to the QuD checks did not align with the QuD condition assigned to them. In total, 5% of the trials were excluded through this procedure.

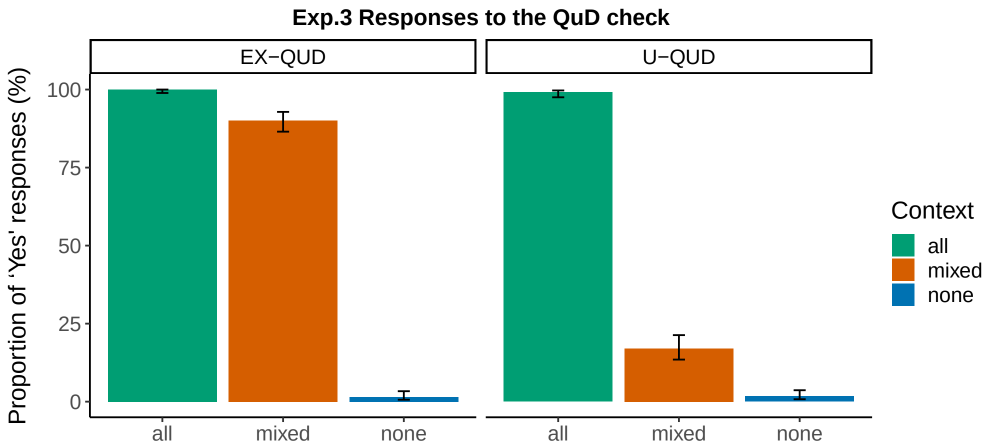


Figure 5 Mean proportion of 'Yes' responses in the QuD task as a function of the type of Context type and the QuD condition. Error bars denote 95% confidence intervals.

7.5 Results

7.5.1 QuD check

Figure 5 shows the results to the QuD check. As expected, irrespective of the QuD condition, participants uniformly answered 'Yes' in the *all*-contexts (*M*s of 'Yes' $\geq$ 99%) and 'No' in the *none*-contexts (*M*s of'Yes' $\leq$ 2%). To test the effect of the QuD manipulation on participants' responses in the critical *mixed*-contexts, we run a mixed-effects logit model with QuD as a predictor (EX-QUD vs. U-QUD; dummy coded), participants' response as a dependent variable ('No'=0 and 'Yes'=1) and a by-Subject random intercept[14]. The model analysis showed a significant effect of QuD ($\beta = -19.50, z = -8.14, p < .001$; model intercept: $\beta = 9.94, z = 6.73, p < 0.001$) indicating that, in the *mixed*-contexts, participants answered 'Yes' far more often in the EX-QUD condition (*M* of 'Yes'= 90%, 95% CI[87, 93]) than in the U-QUD condition (*M* of 'Yes'= 17%, 95% CI[13, 21]). These results show that the QuD manipulation was successful: in the EX-QUD condition, participants considered whether or not at least some tickets had a chance to win whereas, in the EX-QUD condition, they made sure that all tickets were guaranteed to win.

14 R pseudo-code describing the model: `Response~QuD+(1|Subject)`

7.5.2 Graded TVJT

Figure 6 shows the mean ratings in the critical *mixed*-scenarios. Participants' responses in the *mixed*-contexts were analyzed using a linear mixed-effects model fit by restricted maximum likelihood. The model included participant's ratings (range: 0-99) as a dependent variable, the factors QUD (universal; existential), POLARITY (positive; negative), STRENGTH (strong; weak) and their interactions as predictors and a by-Subject random intercept.[15] To facilitate the interpretation of main effects, all factors were contrast coded (level 1: 0.5; level 2:-0.5).

The model analysis showed a significant effect of STRENGTH ($\beta = -77.09, p < .001$) indicating that the acceptance of counterfactuals embedded under strong quantifiers (all Ms< 15) were overall much lower than those of counterfactuals embedded under weak quantifiers (all Ms> 84). Crucially, there was no effect of QUD nor any interaction between QUD and other factors: QUD ($\beta = -0.09$, $p = .97$), QUD × STRENGTH ($\beta = 0.51, p = .91$), QUD × POLARITY ($\beta = -3.53, p = .46$), QUD × STRENGTH × POLARITY ($\beta = -0.39, p = .97$). Finally, neither the effect of POLARITY ($\beta = -4.35, p = .07$) nor the interaction STRENGTH × POLARITY ($\beta = 4.10, p = .39$) was significant, while the intercept of the model was ($\beta = 50.46, p < .001$).

7.6 Discussion

Experiment 3 was designed so as to control for what is relevant in the context by manipulating the type of QuD that participants consider during the judgment task. As explained, the selectional approach predicted no effect of QuD, but a main effect of quantifier strength. On the other hand, the homogeneity approach predicted an interaction between QuD and quantifier polarity. Specifically, on this approach, it was expected that, in lose-some-win-some lottery scenarios, counterfactuals in the scope a positive quantifier should be judged as true on an existential QUD and that counterfactuals in the scope of a negative quantifier should be judged as true on a universal QUD. In our results, we found a strong effect of quantifier strength, as predicted by the selectional approach, but no interaction between QuD and quantifier polarity, contra the predictions of the homogeneity approach. It is worth emphasizing that, while the type of QuD had no reliable effect on participants' truth value judgments, it had a significant effect on their responses to the QuD check. These findings establish that participants were sensitive to our experimental manipulation of the QuD and suggest, in turn, that the lack of QuD effect in the

15 The model with by-Subject random slope for STRENGTH failed to converge while the inclusion of by-Subject random slope for POLARITY was not justified ($\chi(2) = 1.40, p = .50$). R pseudo-code describing the model: `Rating∼Strength∗Polarity∗QuD+(1|Subject)`.

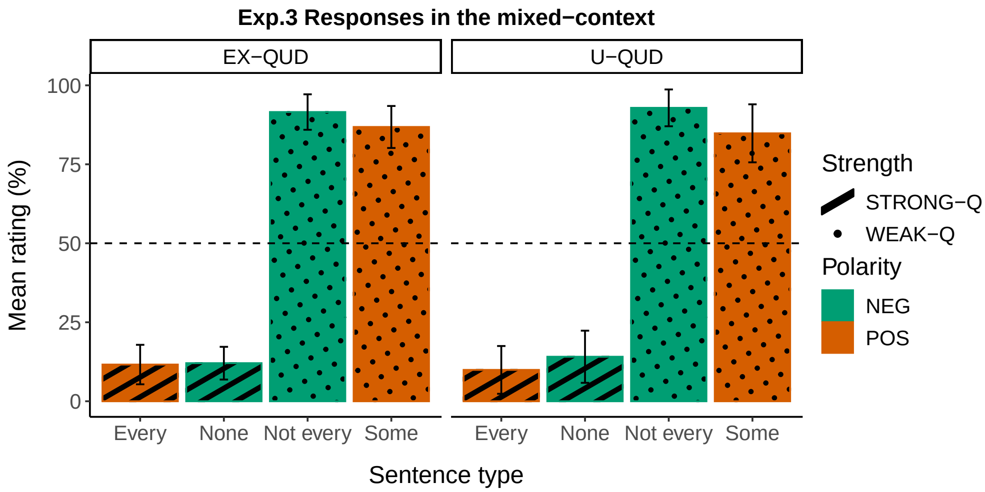


Figure 6 Mean rating to the target sentences in the *mixed*-scenarios by QuD condition as a function the Polarity and Strength of the embedding quantifier. The dotted line represents the midpoint of the response scale and error bars denote 95% confidence intervals.

truth value judgment task cannot simply be explained by saying that participants ignored context-relevant information.

8 General discussion

8.1 Main results

In Experiment 1 and 2, we found clear middle-range ratings for the POSITIVE target in the *mixed*-contexts. 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 (in addition to Stalnaker's classical arguments, see Higginbotham 2003, Klinedinst 2011, Khoo 2022, Santorio 2022, among many). In both experiments, we also found that the endorsement rate for the NEGATIVE target in the *mixed*-contexts 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 as false in these contexts. This finding is consistent with selectional theories, but challenging for both homogeneity and universal theories.

As discussed, the homogeneity theory is nonetheless compatible with these initial results, if we supplement it with a mechanism to pragmatically reinterpret undefinedness on the basis of relevance. In Experiment 3, we tested this theoretical possibility by manipulating the type of QuD while at the same time expanding the range of embedding environments under investigation. The results of this experiment showed no effect of QuD in the TVJT but instead an effect of quantifier strength, in line with the selectional approach but contra the homogeneity approach. Taken together, the present results provide compelling evidence for selectional theories, and against both universal and homogeneity theories.

8.2 A note on future-less-vivids vs real counterfactuals

Recall that we tested future-less-vivid conditionals like (28) in Experiment 1 and 2 before moving to real counterfactuals like (29) in Experiment 3.

(28) No ticket would win, if it was bought.

(29) None of the tickets would have won, if it had been bought.

We took the results of Experiments 1 and 2 to argue against the universal approach. But one might worry that these experiments do not really bear on the semantics of real counterfactuals, since the sentences tested are only future-less-vivids.[16]

Let us notice that it seems implausible that future-less-vivid conditionals and real counterfactuals, both of which share large amounts of modal and temporal morphology, could have different quantificational force.[17] But aside from considerations about theoretical unity, Experiment 3 gives us an empirical argument to respond to the worry. Participants in this experiment robustly rejected sentences like (29) in *mixed*-contexts. This runs against the predictions of the universal approach, on which these sentences are predicted to be straight-up true (as mentioned in Section 3). In addition, we found that participants robustly accepted the *some*-variants of these sentences, which are yet predicted to be false on the universal approach. Counterfactuals with *some* are false according to this approach because the *mixed*-contexts are examples of 'win-some-lose-some' worlds, in which not all bought tickets win. This means that the consequent of the counterfactual is not true in every possible closest world as the universal theory requires. Hence the universal approach is also incompatible with the results from Experiment 3.

16 Thanks to *[name omitted for anonymous review]* (p.c.) for very helpful discussion on this point.

17 *Pace* Lewis, who in the beginning section of *Counterfactuals* (1973b) explicitly lumps future-less-vivids with indicatives, which he took to be material conditionals.

9 The implicature approach

Let us briefly turn to a fourth analytical option and discuss how it fares with our results. This option is the implicature-based account of bare conditionals by Bassi & Bar-Lev (2016).[18] The authors do not directly discuss counterfactual conditionals, but their account can easily be extended to these cases as well. In a nutshell, Bassi & Bar-Lev propose that conditionals have existential force on their basic meaning, as exemplified in (30), and that this meaning is strengthened to a universal one via implicature, as exemplified in (31). The details of how this implicature comes about are not important for us; for clarity, we will simply mark the presence of this implicature via an 'EXH' operator.

(30) a. If ticket #37 was bought, it would win a prize
b. $\exists w'$: $w' \in \textsc{max}_{w,\preceq}(\llbracket \text{A} \rrbracket^{w,\preceq}), \llbracket \text{C} \rrbracket^{w',\preceq}$ = true

(31) a. EXH[If ticket #37 was bought, it would win a prize]
b. $\forall w'$: $w' \in \textsc{max}_{w,\preceq}(\llbracket \text{A} \rrbracket^{w,\preceq}), \llbracket \text{C} \rrbracket^{w',\preceq}$ = true

In unembedded cases, the strengthened meaning tends to be the prominent one, if not the only possible one. But the basic meaning is expected to resurface in environments where implicatures tend not to arise, like in downward entailing contexts. This predicts, for instance, that a sentence like (32) 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. As can be verified, these truth conditions are false in our *mixed*-contexts.

(32) a. None of the tickets would win a prize if it was bought.
b. $\neg\exists x[\exists w'$: $w' \in \textsc{max}_{w,\preceq}(\llbracket \text{P}x \rrbracket^{w,\preceq}), \llbracket \text{Q}x \rrbracket^{w',\preceq}$ = true]

The implicature approach is compatible with the results of Experiment 1 and 2. In particular, it predicts the results that we found for the NEGATIVE target conditions. In addition, the intermediate ratings observed for the POSITIVE ones are also in line with what is generally found for ambiguous scalar implicatures in truth value judgment tasks (see Tieu et al. (2019), Renans et al. (2017), Marty et al. (2015) among others).[19] The results of Experiment 3, however, are more challenging for

18 See also Herburger 2015b,a for a similar account, though not explicitly based on implicature. Herburger's account concerns bare conditionals and not counterfactuals, but in principle could be extended to the latter.

19 One way to make this observation more precise is to hypothesize that, in cases where implicatures are possible (e.g., 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 Bar-Lev also provides an implicature-based account).

this approach. To illustrate, consider first the sentence in (33a). Since the scope of *every* is an upward entailing environment, its meaning will be strengthened by an implicature, (33b). As Bassi & Bar-Lev (2016) argues, an implicature in the scope of this quantifier gives the intuitively correct (strong) meaning of such sentences, which is also in line with our results (i.e., (33a) is correctly predicted to be false in our *mixed*-contexts).

(33) a. Every one of the tickets$_x$ EXH[x would win a prize if it was bought].
b. $\forall x[\forall w'\colon w' \in \textsc{max}_{w,\preceq}([\![\mathrm{P}x]\!]^{w,\preceq}), [\![\mathrm{Q}x]\!]^{w',\preceq} = \text{true}]$

The problem is that the same reasoning should apply to *some*-sentences like (34), which is therefore predicted to be false in our *mixed*-contexts. This prediction, however, is not in line with our results. Furthermore, we note that this (incorrect) prediction obtains independently of the other potential implicature from *some* to *not all* (i.e., *not all of the tickets would have won a prize if it had been bought*).

(34) a. Some of the tickets$_x$ EXH[x would win a prize if it was bought].
b. $\exists x[\forall w'\colon w' \in \textsc{max}_{w,\preceq}([\![\mathrm{P}x]\!]^{w,\preceq}), [\![\mathrm{Q}x]\!]^{w',\preceq} = \text{true}]$

While the case of *some* could be accommodated by considering a global, rather than local derivation of implicatures, the case of *not every* remains challenging for the implicature approach.[20] The reason is that the implicature approach predicts such sentences to be false in our *mixed*-contexts (i.e., they are predicted to be equivalent to *some of the tickets could not win if it was bought*, which is false in our mixed scenario, as all tickets had a chance to win). However, our results show that they are in fact fully accepted in these contexts. Finally, we note that, given the sensitivity of implicatures to relevance, one would naturally expect the interpretation of sentences like (33) and (34) to be affected by the type of QuD being addressed. But this prediction, as we know, is not empirically borne out.

In sum, the results of Experiment 3 are challenging for the implicature approach for two main reasons: (i) we found no effect of QuD in the TVJT and (ii), most importantly, the results for the counterfactuals involving *some* and *not every* are not in line with the predictions of this approach.

10 Conclusion

In this paper, we reported on three experiments testing the predictions of three major families of theories of counterfactuals: universal, selectional, and homogeneity theories. The critical cases that we tested in our experiments were constructed

20 Thanks to *[name omitted for anonymous review]* for discussion of this point. We refer to Bar-Lev 2020 for relevant discussion of the corresponding case of free choice configurations embedded under *some*.

so as to discriminate between the key predictions of these three theories. The findings of all three experiments support selectional theories and challenge universal and homogeneity theories. In addition, we have briefly discussed a more recent, implicature-based approach to counterfactuals, and we have showed that its predictions do not align with the results in our third experiment.

To conclude, we want to hint at a more general theoretical issue, namely the similarities and differences between counterfactuals and plural definites. We mentioned in Section §6 that plural definites are often considered the paradigmatic case of homogeneity. Analyses of definites that employ homogeneity are however controversial and alternative theories based on scalar implicature have been proposed. What is interesting for us is that recent experimental work on plural definites has yielded very different results, despite using a paradigm that is quite similar to the one used in this paper. In particular, Augurzky et al. (2023) investigated plural definites like 'his presents' in the scope of *every*, *not every* and *no* (as in (35)–(37)) while manipulating the QuD in the context.

(35) Every boy opened his presents.

(36) Not every boy opened his presents.

(37) No boy opened his presents.

As the authors discuss, their results are challenging for both the homogeneity and the implicature approach in that they found an effect of QuD for both *every* and *not every* but much less for *no*. Aside from the theoretical implications for plural definites, what is most relevant for us here is that these results are very different from what we found for counterfactuals regarding both the effect of QuD and the response patterns for the different quantifier conditions, especially *every* and *not every*. Taken together, these results suggest that the analogy between conditionals and plural definites, which has been put forward by several authors in the literature, may be on the wrong track.

References

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

Agha, Omar. 2023. *How to ignore counterexamples: Homogeneity across times and worlds*: New York University (NYU) PhD dissertation.

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

Augurzky, Petra, Marion Bonnet, Richard Breheny, Alexandre Cremers, Cornelia Ebert, Clemens Mayr, Jacopo Romoli, Markus Steibach & Yasutada Sudo. 2023. Putting plural definites into context. In *Proceedings of Sinn und Bedeutung 27*, https://ling.auf.net/lingbuzz/007106?_s=aqzVP9HC_hzOFedF&_k=UkMJGTMoeuHZao_3.

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. In *Proceedings of sinn und bedeutung 21*, .

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

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. https://doi.org/10.1007/BF00485047.

von Fintel, Kai. 1997. Bare plurals, bare conditionals, and only. *Journal of Semantics* 14(1). 1–56.

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.

Herburger, Elena. 2015a. Conditional perfection: The truth and the whole truth. In *Semantics and linguistic theory 25*, .

Herburger, Elena. 2015b. Only if: If only we understood it. In Eva Csipak & Hedde Zeijlstra (eds.), *Proceedings of sinn und bedeutung 19*, 284–301.

Higginbotham, James. 2003. Conditionals and compositionality. *Philosophical Perspectives* 17(1). 181–194. https://doi.org/10.1111/j.1520-8583.2003.00008.x.

Iatridou, Sabine. 2000. The grammatical ingredients of counterfactuality. *Linguistic inquiry* 31(2). 231–270.

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 & Pragmatics* 18(10). https://doi.org/https://doi.org/10.3765/sp.10.18.

Khoo, Justin. 2022. *The meaning of "if"*. New York, USA: Oxford University Press.

Klinedinst, Nathan. 2011. Quantified conditionals and conditional excluded middle. *Journal of Semantics* 28(1). 149–170. https://doi.org/10.1093/jos/ffq015.

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.

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

Lewis, David K. 1973a. *Counterfactuals.* Cambridge, MA: Harvard University Press.

Lewis, David K. 1973b. *Counterfactuals.* Cambridge, MA: Harvard University Press.

Lewis, David K. 1973c. 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.

R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

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. Springer.

Roberts, Craige. 1987. *Modal subordination, anaphora and distributivity*: University of Massachusetts, Amherst PhD dissertation.

Santorio, Paolo. 2022. Path semantics for indicative conditionals. *Mind* 131(521). 59–98. https://doi.org/10.1093/mind/fzaa101.

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.

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. In *Semantics and linguistic theory*, vol. 29, 706–726.

Willer, Malte. 2022. Negating conditionals. In Ernest Lepore & David Sosa (eds.), *Oxford studies in philosophy of language volume 2*, 234–266. Oxford University Press.