Semantics & Pragmatics Volume 18, Article 6: 1–43, 2025
https://doi.org/10.3765/sp.18.6

Counterfactuals and quantificational force: Experimental evidence for selectional semantics *

Sonia Ramotowska
University of Amsterdam

Paul Marty
Universidade de Lisboa

Jacopo Romoli
Heinrich Heine Universität Düsseldorf

Paolo Santorio
University of Maryland, College Park

Submitted 2023-09-14 / First decision 2023-11-25 / Revision received 2025-01-31 / Accepted 2025-04-21 / Final version received 2025-05-15 / Published 2025-06-09 / Final typesetting 2025-10-13

Abstract Theories of counterfactuals agree on the use of a comparative similarity relation, but disagree about the quantificational force of counterfactual modals. This study reports findings from two experiments designed to evaluate the predictions of three prominent approaches: universal theories, homogeneity theories, and single-world selection theories (supplemented with supervaluations over selection functions). To differentiate the predictions of these theories, we examined counterfactual sentences embedded under various quantifiers and elicited graded truth-value judgments from speakers. The results provide empirical support for selection-based theories, while posing challenges to universal and homogeneity approaches. Additionally, we argue that a more recent implicature-based theory also fails to align with our findings. We discuss the broader implications of these results, including the similarities and differences between counterfactuals and plural definites.

Keywords: counterfactuals, homogeneity, supervaluations, undefinedness, semantics

* 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, Dean McHugh, Marco Degano and the audiences at SALT30 and SALT33, and at ELM1. SR was supported by the NWO OC project Nothing is Logical (grant no 406.21.CTW.023). PM was supported by a CEECIND grant from the Portuguese Foundation for Science and Technology (DOI: 2023.06820.CEECIND/CP2891/CT0008). JR was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projekt- nummer: RO6384/1-1

1 Introduction

On all standard analyses, the semantics of counterfactual conditionals like (1) exploits 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 had been bought, it would have won 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 three competing views. On the analyses put forward by Lewis (1973a,b) 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, 1981a, 1984), counterfactuals are not strictly speaking quantifiers. Rather, they select the single closest antecedent-verifying world. 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 account for some basic facts but diverge in their predictions. The most notable difference lies between the universal theory and the other two. On the universal theory, counterfactuals like (1) are either clearly true or clearly false.[1] In contrast, both the selectional and the homogeneity theory allow for a 'third status' — a notion that we will elaborate on later. However, the selectional and homogeneity theories also differ in more subtle ways, particularly in their predictions about embeddings.

The goal of this study is to identify cases where the three theories make distinct predictions and test these cases experimentally. We report results from two experiments that focus on counterfactuals embedded under quantifiers of varying strength. As we will discuss, the findings strongly support the selectional theory, whose predictions align well with the data. Conversely, the results pose significant challenges for both the quantificational and homogeneity theories.

1 *Modulo*, of course, the presence of items that would induce a failure of bivalence in the antecedent of the consequent. The point here is that counterfactuals themselves are never the source of undefinedness, or another type of third status.

The rest of this paper is structured as follows. In Section 2, we provide a detailed discussion of the three theoretical approaches and their predictions. In Section 3, we review a relevant previous study and highlight its shortcomings. Section 4 offers an overview of our experiments, the results of which are presented in Section 5 and Section 6. In Section 7, we analyze the results in relation to the theoretical predictions. Section 8, we briefly examine the implicature approach, a fourth theoretical option, and its incompatibility with our findings. Section 9 concludes the paper. Throughout the paper, we use the traditional notation '$\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}$' as shorthand for the counterfactual *If A, would B*.

2 Counterfactuals, indeterminacy and undefinedness

2.1 Three theories of counterfactuals

All standard theories on counterfactuals agree on the general form of their truth conditions: $\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}$ is true just in case B is true in some relevant range of 'closest' or 'most similar' A-verifying worlds. 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 at least as close as w_k to a 'base' world w_0. While this is common ground among all the main views, there is disagreement about the quantificational force of counterfactuals, and in fact on whether counterfactuals are quantificational at all. In what follows, we review the three main theories that have been developed in the literature.[2]

2.1.1 The universal theory

On a first theory, counterfactuals have universal quantificational force. Classical accounts such as Lewis's (1973a,b) and Kratzer's (1986, 2012), as well as their descendants, adopt versions of this view. Let us use the notation '$[\![\mathsf{A}]\!]_{\preceq}$' to denote the set of worlds such that A is true at those worlds, relative to $\preceq$ (so $[\![\mathsf{A}]\!]_{\preceq} = \{w : [\![\mathsf{A}]\!]^{w,\preceq} = \text{true}\}$). Abstracting away from some complications, the truth conditions we get on these theories are in (2).[3]

2 There are well-known dynamic variants of the accounts presented below: see von Fintel 2001 and Gillies 2007. These variants employ a simple strict conditional semantics, while relocating the appeal to closeness in mechanisms of context shift. 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). The reason is that (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.

(2) $[\![\mathsf{A} \Box\!\!\rightarrow \mathsf{B}]\!]^{w,\preceq} = \text{true} \quad \text{iff} \quad \forall w' \in \textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]_{\preceq})\text{: } [\![\mathsf{B}]\!]^{w',\preceq} = \text{true}$
where $\textsc{max}_{w,\preceq}([\![\mathsf{A}]\!]_{\preceq})$ is the set of maximally $\preceq$-close A-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 to w where Maria flipped the coin, the coin landed heads in w'.

Suppose that the closest worlds to the actual world involve a mix of heads and tails-worlds — a plausible assumption under any reasonable interpretation of closeness, given the inherently chancy nature of coin flips. Under these conditions, (3) is predicted to be false in this context.

2.1.2 The selectional theory

A second theory denies that counterfactuals are quantificational. Rather, counterfactuals select a single closest antecedent-verifying world (see Stalnaker 1968, 1981a, 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 an 'input' world and a proposition to a selected world.[4] Schematically, $\mathsf{A} \Box\!\!\rightarrow \mathsf{B}$ is true just in case B is true at the s-selected A-world.

Famously, Lewis allows that, in some cases, there can be infinite chains of antecedent-worlds that are increasingly closer to the 'base' world (see also 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 For completeness, we provide below the conditions that Stalnaker imposes on selection functions:

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

ii. $s(w, A) = \lambda$ iff $[\![A]\!] = \varnothing$
(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 $s(w, A) = s(w, A')$

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

The selectional theory does not appeal directly to comparative closeness, but talk of selection functions can be rephrased in these terms on the assumption that the relation of comparative closeness induces a linear order on worlds (for discussion of this assumption, see Lewis 1973a, Ch.2). To put it simply, the selected world is the single closest world to the world of evaluation that makes true the antecedent of a counterfactual.

Without supplementation, the selectional theory runs into an obvious difficulty. The theory requires that, for every antecedent A and every world w, there is a single closest A-world to w. But examples like (3) suggest that this assumption is implausible. In a situation where Maria is flipping a fair coin, some heads-worlds and some tails-worlds will be tied for closeness, no matter what specific construal of closeness we adopt.

Stalnaker (1981a, 1984) proposes that this problem should be handled not within the semantics proper, but in the metasemantics. He suggests that the selection function is one of the parameters fixed by the context. However, in some cases, the context does not determine a unique selection function. Instead, there is a range of candidate selection functions compatible with all contextual facts, leaving it indeterminate which function is the 'right' one. In such cases, a sentence will be judged 'clearly true' (or 'clearly false') if it's true on all (or no) possible choices of the contextual parameter. Otherwise, the sentence has a third status, which we characterize as 'indeterminate'. The notions of determinate truth, determinate falsity, and indeterminacy can be captured using supervaluations (van Fraassen 1969; see also Fine 1975 for a use of supervaluations to model vagueness).

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

Note that these notions are not part of the compositional semantics itself. Instead, they come into play after the compositional computation of semantic value is complete. As we will see later, this distinction is crucial to understanding how indeterminacy projects under embeddings. For now, let us consider 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 selected flip-world is a heads-world, on others a tails-world. Consequently, (3) is predicted to be indeterminate. This highlights a key difference in predictions between the universal theory and the selectional theory.

2.1.3 The homogeneity theory

The homogeneity theory (von Fintel 1997, Schlenker 2004, Križ 2015, Agha 2023; see also Willer 2022) shares features with both theories above. Like the universal theory, it treats counterfactuals as universal quantifiers over closest antecedent worlds. Like the selectional theory, it assumes that some counterfactuals have a 'third status'. Crucially though, this third status has a different origin.

On the homogeneity theory, counterfactuals include a homogeneity requirement: $\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}$ requires that either all closest A-worlds are B-worlds, or all closest A-worlds are non-B-worlds. Following a common view in the literature, we assume that counterfactuals failing this requirement are undefined, i.e., they are neither true nor false. On this view, a sentence like (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 undefined otherwise. Schematically:

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

The homogeneity theory is in part motivated by a suggestive analogy with the interpretation of plural definites like *the windows* (as emphasized by Schlenker 2004). Consider (7) for instance:

(7) The windows are open.

(7) is judged true if all the windows are open, false if none of them is, and triggers a hedged judgment when some windows are open and some aren't. On the homogeneity theory, the phenomenon responsible for the trivalent behavior of (7) occurs for (3). In the case of (3), it is evident that, if the coin is fair then, among

all closest Maria-flipping worlds, some are heads-worlds while others aren't. Consequently, (3) is predicted to be undefined. As a result, the homogeneity and selectional theories yield analogous predictions in unembedded cases like (3).

2.2 Comparing theories

2.2.1 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. One key point of contention is the principle of Conditional Excluded Middle:

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: suppose that A takes us to a single world then, in that world, either B is true, or B is false; in either case, one of the two disjuncts of $(\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}) \vee (\mathsf{A} \mathrel{\Box\!\!\rightarrow} \neg\mathsf{B})$ is true, hence the disjunction 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'. This principle follows from CEM, via minor side principles. Accordingly, it is validated by the selectional theory, and invalidated by the universal theory.

Negation Swap $\neg(\mathsf{A} \mathrel{\Box\!\!\rightarrow} \mathsf{B}) \mathrel{=\!|\!\vDash} \mathsf{A} \mathrel{\Box\!\!\rightarrow} \neg\mathsf{B}$ (for all consistent A)

As Lewis himself acknowledged (1973a: 80), CEM and related principles initially seem valid. For example, consider the sentences in (8): (8a) and (8b) appear equivalent. This observation provides prima facie support for the selectional theory and challenges the universal theory.[5]

(8) a. It is 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.

On the other side, the main argument for the universal theory is that counterfactuals are the duals of *might*-counterfactuals ($\mathsf{A} \mathrel{\Diamond\!\!\rightarrow} \mathsf{B}$). Duality is the principle that *might*-counterfactuals are equivalent to negations of *would*-counterfactuals, in the following way:

5 For reasons of space, we omit arguments for the selectional theory not directly related to negation; see Klinedinst 2011, Khoo 2022, Santorio 2022, among others.

Duality. $A \Diamond\!\!\rightarrow B \dashv\vDash \neg(A \Box\!\!\rightarrow \neg B)$

By validating Duality, the universal theory predicts that pairs of *would*- and *might*-counterfactuals, like the one in (9), are incompatible.

(9) a. If Maria had flipped the coin, the coin might have landed tails.
b. If Maria had flipped the coin, the coin would not have landed tails.

This seems a good prediction, as the sentences in (9) do indeed sound contradictory. Conversely, the selectional theory does not validate this prediction.[6]

The homogeneity theory, like the selectional theory, validates CEM and Negation Swap.[7] In addition, it can predict that *would*-counterfactuals and *might*-counterfactuals are duals.

In the next section, we go on to show that the predictions of these theories diverge further when counterfactuals are embedded under certain quantifiers.

2.2.2 Selection vs. homogeneity theories: The new debate

As discussed, both theories agree that some counterfactuals have a 'third status' — understood as indeterminacy in the selectional theory, and as undefinedness in the homogeneity theory. This symmetry in predictions is, in fact, quite general. Assuming symmetric conditions about which worlds are tied for closeness and which selection functions are admissible, the two theories consistently align in their verdicts for unembedded counterfactuals, in the following sense: $A \Box\!\!\rightarrow B$ is predicted to be determinately true/false/indeterminate by the selectional theory just in case it is predicted to be true/ false/undefined by the homogeneity theory. However, this symmetry breaks when counterfactuals are embedded. The divergence stems from the level of semantic composition at which the 'third status' arises. On the selectional theory, indeterminacy emerges only at a 'global' level, when we define truth relative to a context. At all stages of semantic composition, the selectional theory is otherwise indistinguishable from a standard bivalent theory. By contrast, the homogeneity theory allows undefinedness to arise at a 'local' level, i.e., during composition. This means that some subsentential constituents can have undefined semantic values, requiring

6 It is an open question how a selection theory might vindicate the data that vindicates Duality. For recent relevant discussion of *might*-counterfactuals, see Mizuno 2025 and Santorio 2024.

7 Since the homogeneity theory appeals to undefinedness, it needs to be equipped with a notion of consequence that allows for trivalence. The natural option, which we follow here, is to appeal to Strawson-Entailment (see von Fintel 1999). Both CEM and Negation Swap inferences are valid on Strawson-Entailment. We note though that CEM will still have undefined instances; see Križ 2015 for discussion of this point in relation to plural definites.

the homogeneity theory to include a projection algorithm to determine how the undefinedness of simpler expressions impacts the definedness of more complex expressions.

This fundamental difference between the two theories leads to key divergences in their predictions for embedded counterfactuals. To illustrate this, consider the following lottery scenario:

> **Win-some-lose-some lottery scenario (aka mixed context)**
> 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 that are 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 lottery:[8]

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

In this scenario, both the selectional and the homogeneity theory predict that (1) has a 'third status'. According to the selectional theory, this prediction arises because some candidate selection functions map the world of evaluation and the antecedent of (1) to a world where ticket #37 wins, and some other candidate selection functions map the same inputs to a worlds where ticket #37 loses. Similarly, on the homogeneity theory, the prediction arises because the set of closest antecedent worlds includes both winning and losing outcomes for ticket #37.

Now, holding the same scenario fixed, consider (10), where the conditional is embedded under a negative quantifier:[9]

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

The selectional theory predicts that (10) is determinately false. The reason is that, in the given scenario, all candidate selection functions select 'win-some-lose-some' worlds — worlds where at least some tickets (but not all of them) win. Since (10) asserts that no bought tickets win, it is false in all these worlds. Consequently, (10) is false relative to all candidate selection functions, hence

8 As Simon Goldstein and Angelika Kratzer 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). We use it simply for illustration here. See discussion of this point in Section 3.

9 Examples of this type were first used by Higginbotham (1986, 2003) to argue in favor of CEM. See also Klinedinst 2011 for detailed discussion of the argument for CEM from quantified conditionals.

determinately false. By contrast, the homogeneity theory predicts that (10) is undefined. To understand this, let us consider the structure of (10), represented in (11):

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

From our earlier discussion, we know that, for all values of x, the embedded clause *if x was bought, x would win a prize* is undefined. To determine the definedness condition for the entire sentence, we need to analyze how undefinedness projects under negative determiner phrases like *no ticket*. The literature offers two main options for this (see Križ 2015, Križ & Chemla 2015; see also George 2008, Fox 2012, Mandelkern 2016 for a corresponding debate on 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 purposes, this distinction is irrelevant. Since the open sentence embedded under *no ticket* in (11) is undefined for all objects in the domain, (10) is predicted to be undefined regardless of which projection algorithm is chosen.

The predictions of the three families of theories are summarized in Table 1. In sum, selectional and homogeneity theories agree that counterfactuals can have a 'third status', leading to hedged judgments. Additionally, when symmetric assumptions are made about selection functions and comparative closeness, both theories align in their predictions for unembedded counterfactuals. For example, sentences like (1) are predicted to have a 'third status' in 'win-some-lose-some' mixed scenarios. However, the theories diverge when considering embedded counterfactuals. In the scenarios of interest, a sentence like (10) is predicted to be false on selectional theories, but undefined on homogeneity theories. By contrast, universal theories make distinct predictions all the way in predicting (1) to be false (since it is false that, for each closest world where a ticket is bought, that ticket wins a prize) and (10) to be true (since it is true that there is no ticket that would win a prize in every such closest world).

In the next section, we review a previous study by Marty, Romoli & Santorio (2020) that tested these predictions. We summarize their methods and results and discuss the limitations of their study, which form the basis for the motivation of our research.

THEORY	Example (1)	Example (10)
Universal	false	true
Selectional	indeterminate	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 (10) in the 'win-some-lose-some' mixed scenarios.

3 Previous study: Marty, Romoli & Santorio 2020

3.1 Overview

Marty, Romoli & Santorio (2020) tested the predictions in Table 1 using a graded truth-value judgment task. In their task, participants were presented with a written context directly followed by a target sentence and had to assess the extent to which the sentence was true or false in the suggested context. Participants reported their judgments by setting a slider tooltip along a scale going from 'Completely false' (left anchor) to 'Completely true' (right anchor). The target sentences, illustrated in (12), were unembedded, positive counterfactuals (labeled POSITIVE) and counterfactuals embedded under *no* (labeled NEGATIVE), corresponding to the sentences in (1) and (10), respectively. In the critical conditions, these sentences were paired with win-some-lose-some lottery scenarios, dubbed *mixed*-contexts, in which only part of the tickets bought would win a prize. Each target sentence was associated with a corresponding control sentence, as shown in (13), expected on all theories to be judged as false in the critical conditions for the targets.

(12) **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

(13) **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

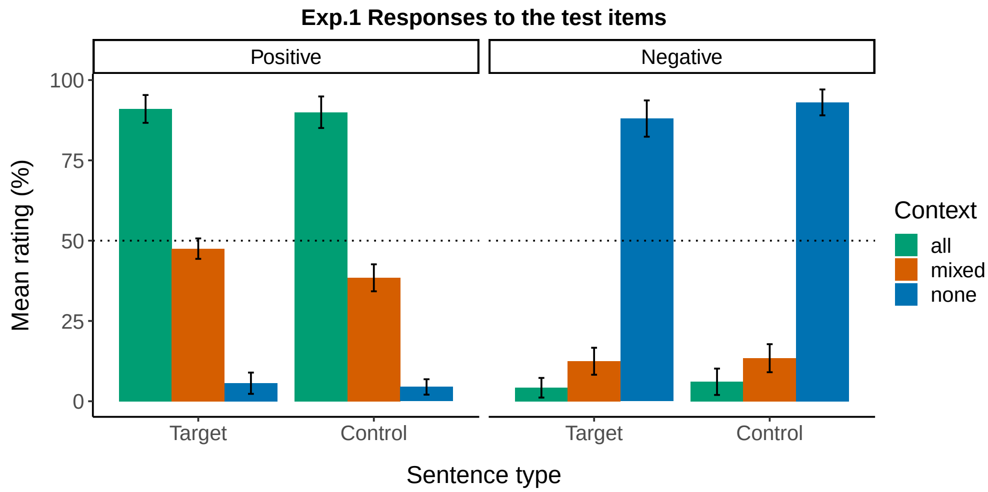


Figure 1 Mean rating to the test items in Marty, Romoli & Santorio 2020 as a function of the type of Context. The dotted line represents the midpoint of the response scale and error bars denote 95% confidence intervals.

The linking hypothesis was that, if the target sentences give rise to hedged judgments in *mixed*-scenarios, participants should set the slider toward the middle of the response scale; conversely, if these sentences give rise to clear judgments of truth or falsity, participants should move the slider away from the middle of the response scale, closer to the extreme values.[10]

The results obtained by Marty, Romoli & Santorio 2020 are presented in Figure 1. In summary, the authors found that, in the critical *mixed*-contexts, unembedded counterfactuals (POSITIVE) received intermediate ratings, closer to the midpoint of the response scale than their corresponding controls. In contrast, counterfactuals embedded under *no* (NEGATIVE) received very low ratings in the same contexts, comparable to their corresponding controls. These

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) or confidence level measure (Syrett & Koev 2015), ternary judgments (Abrusán & Szendröi 2012, Alxatib & Pelletier 2011, Tieu, Bill & Romoli 2019) and multiple unordered choices (Serchuk, Hargreaves & Zach 2011).

results suggest that the former elicited gap judgments, while the latter elicited clear judgments of falsity. Taken at face value, these findings are in line with the predictions of the selectional theory but are unexpected under both the universal and the homogeneity theory.

3.2 Limitations

While the findings of Marty, Romoli & Santorio 2020 are suggestive, they are subject to three important limitations.

First, one might argue that Marty, Romoli & Santorio's (2020) results do not necessarily bear on the semantics of true counterfactuals. The sentences tested in their study are so-called future-less-vivid conditionals, rather than contrary-to-fact conditionals (see Iatridou 2000). Future-less vivid conditionals (FLVs) have antecedents that concern future events and, at least on the face of it, seem compatible with the facts at utterance time. This contrasts with genuine counterfactuals, such as (14), where the antecedent describes an event that directly contradicts the facts at the time of utterance. Thus, it might just be that genuine counterfactuals have different quantificational force.[11]

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

Second, the embedded environments tested in Marty, Romoli & Santorio's (2020) study are limited in scope. Their study focuses exclusively on the diverging predictions of the selectional and homogeneity theories for conditionals embedded under the strong, negative quantifier *no*. However, the divergence in predictions between these theories extends to other positive and negative embeddings involving both weak and strong quantifiers. To illustrate, consider the paradigm of embedded counterfactuals presented in (15). The labels used to classify these sentences reflect two dimensions: the strength of the quantifier (strong vs. weak) and polarity (positive vs. negative).

(15) a. STRONG-POSITIVE
Every one of these tickets would have won if it had been bought.

11 Additionally, the *Strong Centering* principle (i.e., the principle that $A \wedge B$ entails $A \;\Box\!\!\rightarrow B$) may introduce noise in the judgments. To illustrate the issue, consider the future-less-vivid (FLV) conditional in (10). Assume, plausibly, that some tickets will be purchased and that some of these tickets will win a prize. Under these assumptions, at least one FLV of the form *if x was bought, x would win a prize* would hold true due to *Strong Centering*. Consequently, any negative universal claim would be rendered false. This problem critically hinges on the *Strong Centering* principle and does not arise in the case of genuine counterfactuals. We are grateful to an anonymous reviewer for bringing this point to our attention.

b. STRONG-NEGATIVE
None of these tickets would have won if it had been bought.

c. WEAK-POSITIVE
Some of these tickets would have won if they had been bought.

d. WEAK-NEGATIVE
Not every one of these tickets would have won if it had been bought.

In win-some-lose-some scenarios, the homogeneity approach predicts that all four sentences in (15) are undefined. This prediction arises because, as previously shown for (10), the embedded proposition in these examples is undefined in the scenarios of interest. Consequently, each sentence is predicted to be undefined at the global level, regardless of whether one assumes universal or existential projection. Therefore, the predictions of the homogeneity theory for *no* extend to all four quantifiers in (15), irrespective of quantifier strength. By contrast, in the same scenarios, the selectional approach predicts that the sentences in (15) are either true or false, depending on the strength of the quantifier. Specifically, WEAK sentences — which express non-universal quantification (e.g., *some*, *not every*) are predicted to be true whereas STRONG sentences — which express universal quantification (e.g., *every*, *no*) are predicted to be false, regardless of the polarity. The differing predictions of both approaches are summarized in Table 2. While Marty, Romoli & Santorio's (2020) findings provide preliminary evidence in favor of the selectional theory and challenge the homogeneity approach, we argue that a more systematic comparison of their predictions across a wider range of quantificational environments is necessary to draw a definitive conclusion.

THEORY	Strong Positive *every/all*	 Negative *none*	Weak Positive *some*	 Negative *not every/all*
Homogeneity	indeterminate	indeterminate	indeterminate	indeterminate
Selectional	false	false	true	true

Table 2 Predictions of the homogeneity and selectional theories for the true embedded counterfactuals in (15) in the 'win-some-lose-some' mixed scenarios by quantifier strength and polarity.

Third, we note that the homogeneity approach can be made compatible with Marty, Romoli & Santorio's (2020) results if it is supplemented with a pragmatic component sensitive to contextual relevance. For illustration, consider the sentence in (16), which involves the plural definite *the windows*. Plural definites are often regarded as the paradigmatic case of homogeneity. Accordingly, homogeneity theories (see e.g., Križ 2015) treat (16) as true if all windows are open, false if none of them are open, and undefined otherwise.

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

However, it has long been observed that sentences like (16) may be perceived as clearly true or clearly false in contexts where only some of the windows are open. Suppose, for instance, that there is a big storm coming and the speaker realizes that they left a few windows open. Intuitively, an utterance of (16) in this context feels acceptable. This phenomenon, often referred to as 'non-maximality', poses a challenge for the homogeneity approach, which predicts that (16) should result in hedged judgments in such cases.

Homogeneity theorists respond to this challenge by introducing an additional layer of pragmatic interpretation sensitive to relevance. This layer, often modeled using Questions under Discussion (QuD) (Roberts 1996 and subsequent work), can shift semantic undefinedness toward truth or falsity depending on the context. The core idea is that the QuD introduces a contextually relevant partition of epistemically possible worlds. In certain cases, the partition associated with the QuD merges worlds where the proposition is undefined with those where it is true, allowing the proposition to be judged as 'true enough'. Similarly, in cases leaning toward falsity, the same logic can be applied (see Križ 2015, 2016, Champollion, Bumford & Henderson 2019). To illustrate, consider (16) again. In the storm context we described, the speaker is likely concerned about whether their home will flood with rain. Thus, the relevant QuD would likely focus on whether *any* of the windows are open. The partition associated with this QuD combines the cases where all the windows are open with those where only some are. In this context, (16) would be judged as 'true enough', explaining why the sentence feels acceptable in such situations.

This idea can be extended to the embedded conditionals tested in Marty, Romoli & Santorio's (2020) study.[12] Specifically, while semantically undefined, (10) might be judged as 'false enough' in contexts where the partition associated with the QuD does not distinguish between false and undefined cases. For example, (10) could be interpreted as 'false enough' in a context where the relevant QuD is whether each ticket bought *has any chance of winning* — what we

12 While the term 'non-maximality' is not commonly used in the literature on conditionals, Križ (2015) has argued that the phenomenon applies to counterfactuals as well. Specifically, he suggests that, when evaluating counterfactuals, we often disregard far-fetched possibilities.

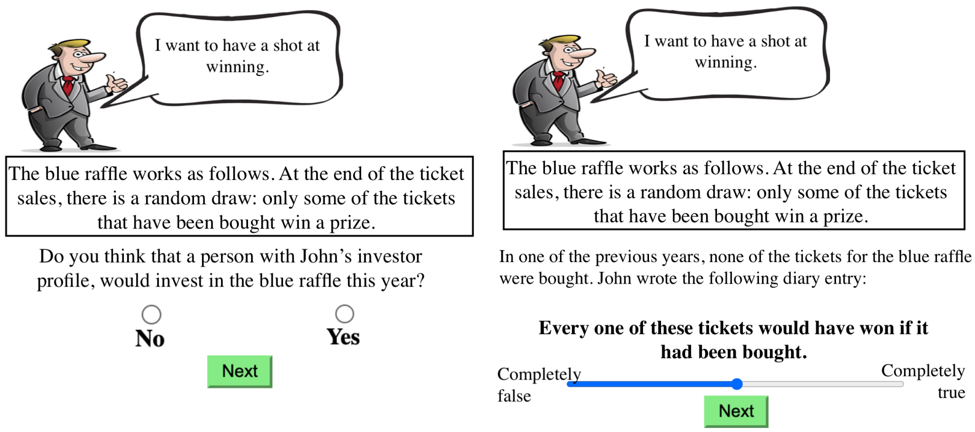
call an ‘existential’ QuD — as opposed to a context where the focus is on whether each ticket *is guaranteed to win* — a ‘universal’ QuD. Since Marty, Romoli & Santorio (2020) did not control for potential QuD effects in their study, it remains possible that their participants interpreted sentences like (10) against an implicit existential QuD, leading them to judge it as false. Such a QuD effect could explain, in particular, the asymmetry observed between unembedded and embedded cases in their findings.

In the next section, we outline the key features of our experiments and describe how we improved upon the design and materials in Marty, Romoli & Santorio (2020). Specifically, we addressed the limitations identified above by testing true counterfactuals, expanding the range of embedded cases, and manipulating the QuD associated with each scenario.

4 The present study

Building on Marty, Romoli & Santorio 2020, we conducted two experiments that addressed the main criticisms of their study through three significant improvements. First, we shifted from testing future-less-vivid conditionals to contrary-to-fact conditionals. Second, we expanded the range of embedded cases, incorporating sentences involving *every*, *not every* and *some*, thereby testing the full paradigm exemplified in (15). Third, drawing on insights from a recent study on plural definites by Augurzky et al. 2023, we manipulated the Question under Discussion (QuD) that participants should consider when judging the sentences presented to them. Specifically, in both experiments, we designed the background story and stimuli to bias participants toward either an existential QuD (emphasizing whether there was any chance of winning) or a universal QuD (emphasizing whether winning was guaranteed) during the judgment task.

Experiment 1 tested the sentences in (15) in the lottery scenarios devised by Marty, Romoli & Santorio 2020. As in their study, the target scenarios involved *lose-some-win-some* mixed contexts, where each ticket had a chance to win but none was guaranteed to win. In our experiment, however, participants were asked to evaluate the extent to which each sentence was true or false in this context, based on the perspective of a character whose attitude toward winning in lotteries was manipulated between subjects. This manipulation was intended to bias participants toward either an existential QuD (henceforth, E-QuD) or a universal one (henceforth, U-QuD) during the judgment task. In the E-QuD group, the character was interested merely in knowing whether there was a chance of winning a prize, making it relevant whether *each ticket has at least* *some chance to win*. In the U-QUD group, the character's goal was to win each and every single time, making it relevant whether *each ticket was guaranteed to win*.



Step 1: QuD check Step 2: Graded TVJT

Figure 2 Example item illustrating the two-step response procedure in Experiment 1. This item is an instance of the STRONG-POSITIVE sentences in the win-some-lose-some *mixed*-scenarios for the E-QUD group. Participants were reminded of the QuD by the character's motto shown in a speech bubble (e.g., for E-QUD: "I want to have a shot at winning.").

Factoring in the potential effect of the QuD on participants' responses in the graded truth-value judgments task, Experiment 1 focused on the predictions of the homogeneity and selectional theories summarized in Table 3. As discussed earlier, the selectional approach expects no effect of the QuD, but it predicts an effect of the quantifier strength. Specifically, in the *mixed*-contexts, the selectional approach predicts that in the case of WEAK sentences (*some*, *not every*) participants would pull the slider towards the end of the scale indicating that these sentences are 'completely true', whereas in the case of STRONG sentences (*every*, *no*) they would pull the slider towards the other end of the scale indicating the these sentences are 'completely false', irrespective of polarity or QuD type. In the results, this would manifest as a main effect of quantifier strength. In contrast, the homogeneity approach anticipates that the QuD manipulation will influence participants' judgments in a principled way. Specifically, in *mixed*-contexts, POSITIVE sentences (*every*, *some*) are expected to be perceived as 'true enough' when the existential QuD is salient than when the universal QuD is. As a result, participants would pull the slider towards the end of the scale, indicating that these sentences are 'completely true'. Conversely, in the case of NEGATIVE sentences (*none*, *not every*), participants would pull the slider towards the end of the scale, indicating that these sentences are 'completely false' under an existential QuD than under a universal QuD. In the results, this would manifest as an interaction between QuD type and polarity.

THEORY	QUD	Strong Positive *every/all*	 Negative *none*	Weak Positive *some*	 Negative *not every/all*
Homogeneity	E-QUD	true	false	true	false
	U-QUD	false	true	false	true
Selectional	E-QUD	false	false	true	true
	U-QUD	false	false	true	true

Table 3 Predictions of the homogeneity and selectional theories for the four sentence types in (15) tested in Experiment 1 in the win-some-lose-some *mixed*-contexts by quantifier strength, polarity and QuD type.

Experiment 2 aimed to further compare the predictions of the homogeneity and selectional theories in Table 3 but used other scenarios than the lottery scenarios from Marty, Romoli & Santorio 2020. Additionally, it sought to more directly assess the reliability of the QuD manipulation by further testing sentences with plural definites, whose interpretation has been shown to be QuD-sensitive (Augurzky et al. 2023).

In this experiment, participants were introduced to a multi-round, four-player card game and tasked with evaluating whether each round of the game met the acceptability criterion established by the game committee. Similar to Experiment 1, this criterion was manipulated between subjects to bias participants toward either an existential or a universal QuD during the judgment task. In the E-QUD group, the committee's criterion was that *at least one player has a chance to win*. In the U-QUD group, the committee's goal was for all players to be satisfied, so the criterion was that all *all players are guaranteed to win*. Participants in both groups were told that only games satisfying the relevant criterion would be deemed acceptable by the committee. The task mirrored that of Experiment 1, with the inclusion of a QuD check designed to verify participants' alignment with the committee's criterion, as illustrated in Figure 3.

Step 2: Graded TVJT

Step 1: QuD check

Figure 3 Example item illustrating the two-step response procedure in Experiment 2. This item is an instance of the WEAK-NEGATIVE sentences in the win-some-lose-some *mixed*-scenarios for the E-QUD group. Participants were reminded of the QuD through a statement of the committee's rule (e.g., for E-QUD: "Remember, the committee wants to know whether at least one player had a chance to win.").

The items involving counterfactuals replicated the conditions in Marty, Romoli & Santorio 2020. The sentences in (15) were adapted by modifying non-essential content words to align with the new background story and setup (i.e., *All/None/Some/Not all of the players would have won if they had played and finished this round*). In the target scenarios, these sentences were presented in a round where the cards had not been assigned to any player because the players did not participate in that round. In these scenarios, it was true that some, but not all, players could have won; however, it was false that any player was guaranteed to win or guaranteed to lose. The predictions of the two theories remained as outlined in Table 3, and thus our hypotheses for these sentences were unchanged.

The items involving plural definites used the sentence *The players won this round*. In the target scenarios, this sentence was presented at the end of a round where the players had participated, but only two of the four players had won. In these scenarios, the homogeneity reading of the sentence (i.e., that all players won) is false. However, as explained earlier, this sentence may be acceptable in a context where the salient QuD is existential and biases toward a non-maximal reading. Therefore, consistent with Augurzky et al. 2023, we predicted this sentence to be more accepted in the E-QuD group than in the U-QuD group. We reasoned that if participants were sensitive to our experimental manipulation of the QuD for plural definite expressions, the absence of a similar effect for counterfactuals could not be attributed to participants ignoring contextually relevant information.

In the following sections, we report on Experiment 1 and Experiment 2. The results of both experiments demonstrate, through different means, that participants were sensitive to context-relevant information pertaining to the QuD manipulation. Furthermore, the results from Experiment 2 confirm that the QuD manipulation had a direct and predictable effect on participants' judgments of sentences involving plural definites. However, findings from both experiments indicate that participants' judgments of counterfactuals were influenced solely by the strength of the quantifier in these sentences, aligning with the predictions of the selectional theory.

5 Experiment 1

5.1 Participants

A total of 100 English monolingual participants were recruited through Prolific and compensated at a rate of £9 per hour for their participation. Due to technical issues, data from three participants were not recorded, resulting in a final dataset of 97 participants (48 female, mean age: 42 years). All participants provided written informed consent, and all data were collected and stored in compliance with the General Data Protection Regulation (GDPR).

5.2 Materials and design

Each item presented a lottery scenario alongside the image of a fictional character and involved a two-step response procedure: (i) a QuD check and (ii) a graded truth-value judgment task, as illustrated in Figure 2. The lottery scenarios were categorized into three types: *all*, *mixed* and *none*, as exemplified below. These scenarios were conceptually similar to those previously used in Marty, Romoli & Santorio 2020.

(17) **All-scenario**
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.

(18) **Mixed-scenario**
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.

(19) **None-scenario**
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 differing attitudes toward winning in lotteries (see Appendix B on OSF for the full instructions). In the E-QUD group, the character merely wanted to have a shot at winning, thereby making it relevant whether *each ticket has at least a chance to win*. This criterion was met in both the *all*-scenarios and the *mixed*-scenarios, where some or all of the tickets were expected to win. By contrast, in the U-QUD group, the character's goal was to win the lottery every single time, thus making it relevant whether *each ticket is guaranteed to win*. This criterion could only be met in the *all*-scenarios, where every ticket was expected to win. To maintain the salience of the QuD throughout the experiment, the relevant character and their motto were displayed at the top of every item, directly above the lottery scenario (see Figure 2).

The QuD check consisted of a Yes/No question asking participants whether, based on the character's investor profile, they believed that this character would invest in the lottery described in the scenario. The question was formulated as follows: *Do you think that a person with [character]'s investor profile, would invest in the [raffle] this year?*, where *[character]* was the name of the character (i.e., John or Bill) and *[raffle]* was the name of the raffle (e.g., the blue raffle). Participants were expected to answer 'Yes' in the *all*-contexts and 'No' in the *none*-contexts, regardless of the QuD group. In the *mixed*-contexts, however, responses were expected to vary based on the QuD manipulation. Specifically, in these contexts, participants in the E-QUD group — where the character's goal is simply to have a shot at winning (John's motto: "I want to have a shot at winning.") — were predicted to answer 'Yes' much more frequently than participants in the U-QUD group — where the character's goal is to secure a guaranteed win (Bill's motto: "I care about winning each and every single time.").

The QuD check was immediately followed by a graded truth-value judgment task modeled after Marty, Romoli & Santorio 2020. The test sentences in this task were contrary-to-facts *would*-conditionals embedded in the scope of one of four quantifiers (*every*, *none*, *some*, and *not every*) that varied in terms of strength (strong vs. weak), polarity (positive vs. negative) or both, resulting in the four sentence types in (15). The type of lottery scenario, the strength and the polarity of the quantifier were all manipulated within subjects, creating a $3 \times 2 \times 2$ factorial design for each of the two QuD groups. Crossing these three factors resulted in 12 test items. To diversify the sentence content, an additional 12 filler items were further included in each survey. Filler items were similar to the test items in all respects except for the sentences presented in the graded truth-value judgment task: in filler sentences, *would* was replaced with *could* (e.g., *Every one of these tickets could have won, if it had been bought*).

5.3 Procedure

Upon entering the study, participants were randomly assigned to one of two QuD groups: E-QuD or U-QuD. In the instructions, participants were presented with a brief background story describing the profile of the character associated with their assigned QuD group (see Appendix B on OSF). They were informed that the character had been studying lotteries for several years and had written a diary entry for each lottery. To emphasize the counterfactual nature of the sentences in the study, participants were also told that the popularity of lotteries varied from year to year, and in some years, no tickets had been purchased.

After the background story, participants were introduced to the two-step response procedure employed in the study. They were told that they would read short stories, each followed by a Yes/No question about the character's investment habits, and then a sentence extracted from the character's diary entries. They were asked to answer the first question and then to evaluate the extent to which the sentence was true or false in the context of the story and from the perspective of the character. Participants were then introduced to the response scale used in the study. They were instructed to move the slider to the right if they judged the sentence to be 'completely true', to the left if they judged it to be 'completely false', and to the middle if they found the sentence to be neither completely true nor completely false. Participants were encouraged to use the full range of the slider to best represent their intuitions about each sentence.

Following the instructions, the experiment began with two unannounced practice trials and then proceeded with 24 experimental items (12 test + 12 filler), presented in random order. At the start of each trial, participants were shown a picture of the character along with a description of a lottery scenario. After reading the scenario, they completed the QuD Check and the graded truth-value judgment task in succession (see Figure 2). For the QuD check, participants responded by clicking one of two buttons labeled 'Yes' or 'No'. For the graded truth-value judgment task, They used a slider ranging from 'Completely false' to 'Completely true', as in Marty, Romoli & Santorio 2020. The character's picture and the relevant lottery scenario remained visible on the screen throughout the trial.

5.4 Data accessibility

The data and analysis scripts for Experiment 1 are available on the Open Science Framework (OSF) Platform: https://osf.io/3jywr/.

5.5 Software

The experiment was implemented using PennController for Internet Based Experiments (Zehr & Schwarz 2018). Data analysis was conducted in RStudio (RStudio Team 2020). Regression models were run using the `lmerTest` package (Kuznetsova, Brockhoff & Christensen 2017), and data visualization was carried out with the `ggplot2` package (Wickham 2016).

5.6 Data preparation

We applied two exclusion criteria to ensure the reliability of our data. First, we reasoned that participants who understood the instructions correctly should consistently answer 'No' in the QuD checks for the *none*-contexts and 'Yes' for the *all*-contexts, regardless of the QuD manipulation. Based on this, we excluded participants whose overall accuracy in the QuD checks for these contexts was below 80%. This criterion resulted in the exclusion of responses from 10 participants. Thus, the data from 87 participants were included in the subsequent analyses, with 43 participants in the E-QuD group and 44 in the U-QuD group.

Second, consistent with our purposes, we excluded from analysis the *all* and *none* trials in which participants responded incorrectly to the QuD checks as well the *mixed* trials in which their responses to the QuD checks did not align with the QuD group assigned to them. Overall, this procedure excluded 5% of the trials.

5.7 Results

5.7.1 QuD check

Figure 4 illustrates the mean proportion of 'Yes' responses to the QuD check across conditions. As anticipated, participants consistently answered 'Yes' in the *all*-scenarios (Ms$\geq$ 99%) and 'No' in the *none*-scenarios (Ms$\leq$ 2%), irrespective of the QuD group. To evaluate the effect of the QuD manipulation on participants' responses in the critical *mixed*-contexts, we conducted a mixed-effects logit model analysis. In the model, the QuD condition (E-QuD vs. U-QuD; dummy coded) served as the predictor, participants' responses as a dependent variable ('No'= 0 and 'Yes'= 1) and a by-subject random intercept was included. The analysis revealed a significant effect of QuD ($\beta = -19.50, z = -8.14, p < .001$; model intercept: $\beta = 9.94, z = 6.73, p < .001$). In the *mixed*-contexts, participants in the E-QuD group answered 'Yes' far more often ($M = 90\%$, $CI_{95\%}$[87, 93]) compared to participants in the U-QuD group ($M = 17\%$, $CI_{95\%}$[13, 21]). These findings demonstrate that participants in the E-QuD group focused on whether some tickets had a chance to win, while participants in the U-QuD group ensured that all tickets were guaranteed to win. This aligns with the character's respective profile in each QuD group.

5.7.2 Graded TVJT

Figure 5 presents the mean ratings for the test sentences in the *mixed*-scenarios. Participants' responses in these scenarios 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 (E-QuD vs. U-QuD), polarity (positive vs. negative), strength (strong vs. weak) and their interactions as predictors. A by-Subject random intercept was also included to account for individual variability. To facilitate the interpretation of the main effects, all factors were contrast-coded, with level 1 coded as 0.5 and level 2 coded as -0.5.

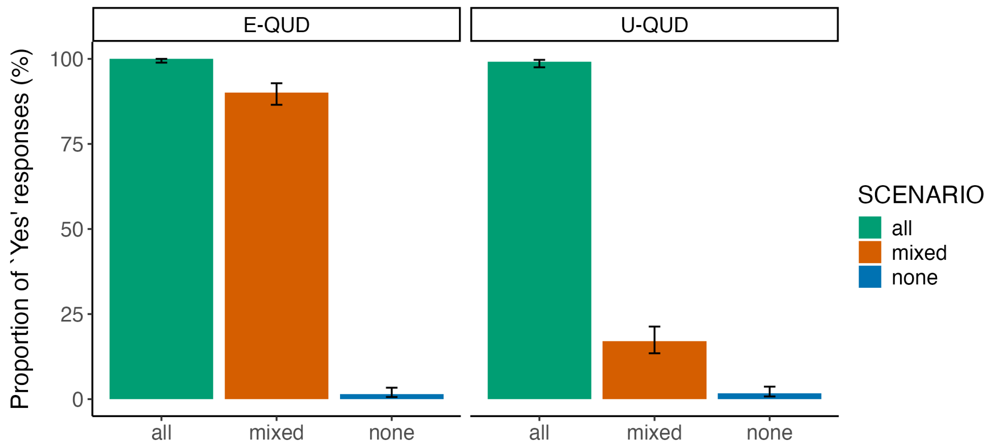


Figure 4 Mean proportion of 'Yes' responses to the QuD check as a function of the type of lottery scenario and QuD group. Error bars denote 95% confidence intervals.

The model analysis revealed a significant effect of STRENGTH ($\beta = -77.09, p < .001$), indicating that counterfactuals embedded under strong quantifiers (all Ms< 15) were rated significantly lower than those embedded under weak quantifiers (all Ms> 84). Critically, there was no significant 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 ($\beta = -0.39, p = .97$). Additionally, neither the effect of POLARITY ($\beta = -4.35, p = .07$) nor the interaction STRENGTH × POLARITY ($\beta = 4.10, p = .39$) was significant. However, the model intercept was significant ($\beta = 50.46, p < .001$).

5.8 Discussion

Experiment 1 sought to address key criticisms of Marty, Romoli & Santorio's (2020) study, particularly by controlling contextual relevance through manipulation of the type of QuD participants considered during the judgment task. The selectional approach predicted no effect of QuD but anticipated a main effect of quantifier strength. In contrast, the homogeneity approach predicted an interaction between QuD and quantifier polarity. Specifically, in *mixed* lottery scenarios, counterfactuals under positive quantifiers were expected to be judged true on an existential QuD, whereas those under negative quantifiers were ex-

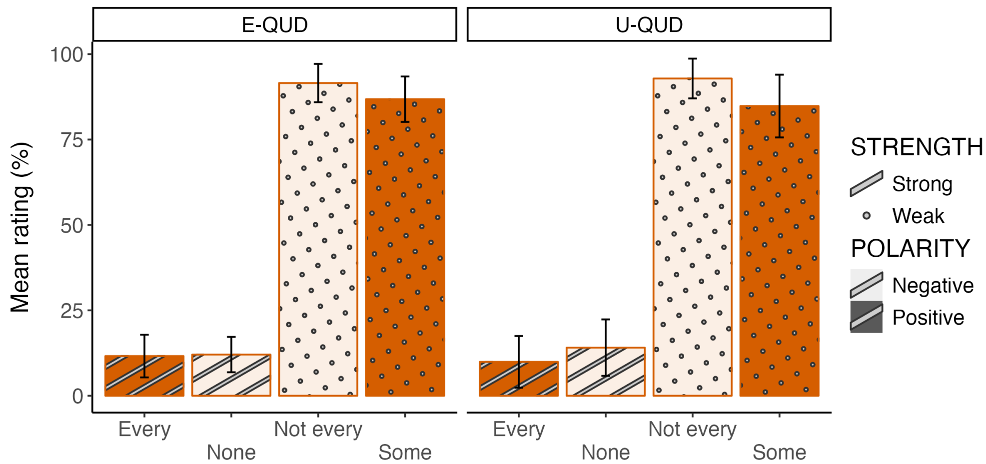


Figure 5 Mean rating for test sentences in *mixed*-scenarios by QuD group. The pattern of the bars represents the quantifier's strength and the transparency indicates its polarity. The error bars denote 95% confidence intervals.

pected to be judged true on a universal QuD. The results revealed a robust effect of quantifier strength, aligning with the predictions of the selectional approach, but no interaction between QuD and quantifier polarity. Notably, while the type of QuD did not reliably influence participants' truth-value judgments, it significantly impacted their responses to the QuD check. These findings indicate that participants were sensitive to the experimental QuD manipulation, suggesting that the absence of a QuD effect in truth-value judgments is unlikely attributable to a disregard for context-relevant information.

While these findings seem to support the selectional approach, we cannot entirely dimiss the possibility that the QuD manipulation was not sufficiently robust to elicit the effect predicted by the homogeneity approach. Specifically, the predominant relevance of actually winning in the lottery scenario may have diminished the impact of the QuD in the E-QuD group, which merely focused on the possibility of winning. This imbalance may have weakened the manipulation's ability to counteract the salience of actual wins. To address this concern, we conducted a second experiment incorporating several key modifications. First, we revised the method of introducing the QuD manipulation and added novel control conditions in which the putative homogeneity requirement is satisfied and the assertive component is true. Second, and most importantly, in addition to counterfactual sentences, we tested sentences involving plural definite expressions, which prior research has shown to be sensitive to QuD (Augurzky et al. 2023). We hypothesized that, if the results of Experiment 1 hold, a successful QuD manipulation would influence truth-value judgments for plural definite sentences, while leaving the truth-value judgments for counterfactual sentences unaffected.

In addition to these key modifications, we implemented a few less significant changes. Specifically, we increased the number of repetitions for each condition to improve the estimation of the by-subject random slopes. Furthermore, we introduced a training phase with feedback before the experiment to minimize data loss due to incorrect responses in the QuD check.

6 Experiment 2

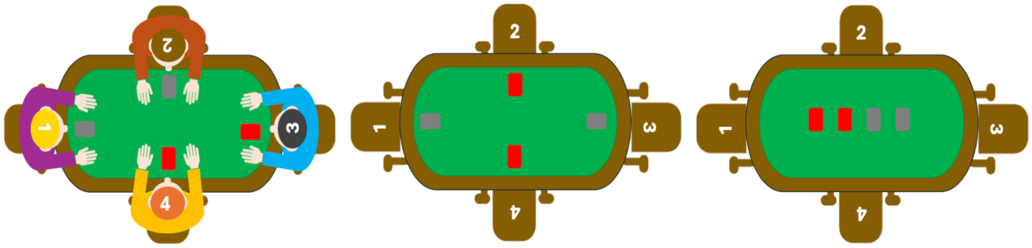
6.1 Participants

We recruited 100 new English monolingual participants (43 female, mean age 39 years) through the Prolific platform, all with approval ratings between 99% and 100%. Participants were compensated at a rate of £9 per hour. The consent and data collection procedures were identical to those used in Experiment 1.

6.2 Materials and design

Experiment 2 employed a mixed design, similar to Experiment 1, with a between-subject manipulation of QuD and within-subject manipulations of all other factors. The procedure involved the same two-step process: a QuD-check followed by a graded truth-value judgment task. The scenarios were parallel to the lottery scenarios used in Experiment 1 but involved a novel setup based on a multi-round card game devised for the study (see Appendix C on OSF for the instructions).

The general narrative of the scenarios was as follows. A group of 4 players participates in a card game. Each round, they are assigned to a game table and given a seat number (1, 2, 3, or 4). At any point during the game, there are always four cards placed on each table, waiting for the players. The cards can be of two different colors, red or gray. The round begins once all the players are seated. The four cards placed on the table are then randomly distributed among the players so that each player receives one card. A player wins if they receive a red card and loses if they receive a gray card.



Cards: Assigned. **Game:** Finished.

Cards: Assigned. **Game:** Disrupted.

Cards: Not assigned. **Game:** On hold.

Figure 6 Example scenarios illustrating the three possible states of the game. These scenarios are instances of the *mixed*-contexts.

Scenarios were presented to participants through visual depictions, with each picture illustrating one of three possible game states: ‘Finished’, ‘Disrupted’, or ‘On hold’, as shown in Figure 6 (see Appendix C on OSF for descriptions of these pictures). In the ‘Finished’ state, the round had both started and concluded: players were seated, and cards were visibly assigned, revealing the winners and losers. In the ‘Disrupted’ state, the round had started but was interrupted before completion: players had left their seats without knowing the results, and the cards were placed at their now-empty seats. In the ‘On hold’ state, the players were on a break and did not play: the seats were empty, and the cards remained unassigned in the middle of the table. Each game state was associated with three possible scenarios — *all*, *mixed* and *none* — paralleling Marty, Romoli & Santorio’s (2020) lottery scenarios. In the *all*-contexts, all four cards were red, meaning all players would win. In the *none*-contexts, all of them were gray, meaning all players would lose; finally, in the critical *mixed*-contexts, two of the four cards were red, ensuring that each player had a chance to win, regardless of the game state.

The judgment task included three groups of sentences: Plural Definite (PD), Control Counterfactuals (CC) and Target Counterfactuals (TC), illustrated below in (20), (22) and (21), respectively. Sentences within each group were consistently paired with a single game state (i.e., Finished, Disrupted or On Hold).

(20) **PD sentences — Finished**
The players won this round.

(21) **CC sentences — Disrupted**
All/None/Some/Not all of the players would have won
if they had finished this round.

(22) **TC sentences — On Hold**
All/None/Some/Not all of the players would have won
if they had played and finished this round.

PD sentences were simple declarative sentences in the past tense involving the plural definite *the players*. These sentences were paired with the 'Finished' state and tested across all three contexts. CC and TC sentences were *would*-counterfactuals embedded under one of four quantifiers — *all*, *none*, *some* and *not all* — differing in terms of strength or polarity or both (as in Experiment 1). CC sentences were paired with the 'Disrupted' state whereas TC sentences were paired with the 'On hold' state. Crucially, the putative homogeneity presupposition that *every player is either guaranteed to win or guaranteed to lose* was true in 'Disrupted' scenarios but false in 'On hold' scenarios. CC and TC sentences were tested with all four quantifiers in the critical *mixed*-scenarios. However, due to possible implicature readings, two quantifier-scenario combinations for TC sentences were excluded from the design: *some* in *all*-scenarios and *not all* in *none*-scenarios. To maintain a balanced number of (expected) true and false sentences in the study, we similarly excluded two quantifier-scenario combinations for CC sentences: *none* in *all*-scenarios and *all* in *none*-scenarios (see Table 4).

	all		*mixed*		*none*	
	CC	TC	CC	TC	CC	TC
all	true	+	false	target	—	+
not all	false	+	true	target	—	—
some	—	—	true	target	false	+
none	—	+	false	target	true	+

Table 4 Possible combinations of quantifier and game scenarios for CC and TC sentences. A minus (—) indicates that the combination was not included in the study and a plus (+) that it was included, but not analyzed.

The QuD was manipulated between subjects using a background story about the game committee. Participants were instructed to verify, on behalf of the committee, whether each round adhered to a specified rule. The rationale for this manipulation was the same as in Experiment 1. In the E-QuD group, the rule established by the committee was that *at least one player has a chance to win*. In the U-QuD group, the rule was that *all players are guaranteed to win*. Each trial included a Yes/No QuD check phrased as follows: *Does this table satisfy the rule?*. To reinforce the QuD throughout the experiment, this question was preceded by a reminder of the committee's rule.

As in Experiment 1, the QuD check was immediately followed by a graded truth-value judgment task. Each PD, TC, and CC sentence was repeated three times across all relevant factors (CONTEXT for PD and CONTEXT, STRENGTH, and POLARITY for TC and CC), resulting in a total of 9 PD, 30 TC, and 24 CC items. Each survey also included filler sentences corresponding to the PD, TC, and CC items. These fillers matched the same game states and contexts as their respective counterparts. PD fillers were of the form *Player [#n] won the round*, TC fillers of the form *Player [#n] would be guaranteed to win this round* and CC filler of the form *Player [#n] was guaranteed to win this round*, where [#n] was a seat number (1, 2, 3 or 4). Each survey included 12 fillers for each sentence group (PD, TC and CC), for a total of 36 filler sentences. Thus, in total, each survey contained 99 items.

6.3 Procedure

The procedure mirrored that of Experiment 1. Upon entering the study, participants were randomly assigned to one of the two QuD groups. The instructions introduced participants to the card game and provided examples of the three possible game states. The QuD manipulation was established through a background story about the game committee and its rule: that *at least one player has a chance to win* in the E-QuD group and that *all the players are guaranteed to win* in the U-QuD group. Participants were tasked with verifying, on behalf of the committee, whether the game satisfied the relevant rule in each round.

Following the background story, participants were introduced to the study's two-step response procedure. They were informed that they would view game scenarios, each followed by a Yes/No question about whether the scenario satisfied the committee's rule, and a sentence related to the scenario. Participants were instructed to first answer the Yes/No question and then evaluate the extent to which the sentence was true or false based on the presented scenario. The response scale was explained using the same instructions as in Experiment 1.

Following the instructions, participants proceeded to a training block consisting of 4 trials presented in a fixed order. These trials aimed to ensure that participants correctly interpreted the committee's rule and understood the 'Disrupted' and 'On Hold' states. The training trials had the same structure of the trials in the main experiment and included 2 TC fillers presented in *mixed* and *all*-scenarios and 2 CC fillers presented in *mixed* and *none*-scenario. During the QuD-check, participants received feedback aligned with their QuD group. In the judgment task, they were provided with explanations clarifying why a given sentence was true or false for the scenario. After completing the training block, participants advanced to the main experiment.

As in Experiment 1, each trial consisted of two steps completed in succession: the QuD check and the graded truth-value judgment task (see Figure 3). In the QuD check, participants were reminded of the committee's rule (e.g., *Recall that the committee wants to know whether all players were guaranteed to win*) and asked to respond by selecting one of two buttons labeled 'Yes' or 'No'. In the graded truth-value judgment task, the reminder of the rule remained on the screen. Participants were instructed to use a slider to judge the sentence provided. The picture depicting the game situation was visible throughout the trial to ensure participants could reference it while completing the tasks.

6.4 Data availability

The data and analysis scripts for Experiment 2 are available on the Open Science Framework (OSF) Platform: https://osf.io/3jywr/.

6.5 Software

The software used for running the experiment and data analyses were the same as in Experiment 1 (see Section 5.5 for details).

6.6 Data preparation

We applied the same exclusion criteria as in Experiment 1 (see Section 5.6 for details). Based on the first criterion, we excluded 6 participants, leaving 94 participants with an accuracy rate of 99%. Our final sample consisted of 94 participants (41 females, mean age 40 years old), with 43 participants in the E-QuD group and 51 in the U-QuD group. In accordance with the second criterion, we excluded 73 responses, which accounted for less than 1% of the data. Overall, the number of incorrect responses to the QuD checks in this experiment was lower than in Experiment 1, suggesting that the materials and method in Experiment 2 enhanced the effectiveness of the QuD in both groups.

6.7 Results

6.7.1 QuD check

Figure 7 illustrates the mean proportion of 'Yes' responses to the QuD check across conditions. Participants performed at ceiling in all scenarios in both QuD groups. In the E-QuD group, the proportion of 'Yes' responses was 99% $CI_{95\%}$[99.6; 98.7], 99% $CI_{95\%}$[99.4; 98.4], and 1.8% $CI_{95\%}$[2.7; 1.2] for *all*, *mixed*, *none* scenarios, respectively. In the U-QuD group, the corresponding rates were 98.7% $CI_{95\%}$[99.2; 98], 0.2% $CI_{95\%}$[0.5; 0.08], and 0% $CI_{95\%}$[0.3; 0] for *all*, *mixed*, *none* scenarios, respectively. The very high ratings in the *mixed* scenario for the E-QuD group, contrasted with the very low ratings for the U-QuD group, demonstrate an improved effectiveness of the QuD manipulation compared to the previous experiment.

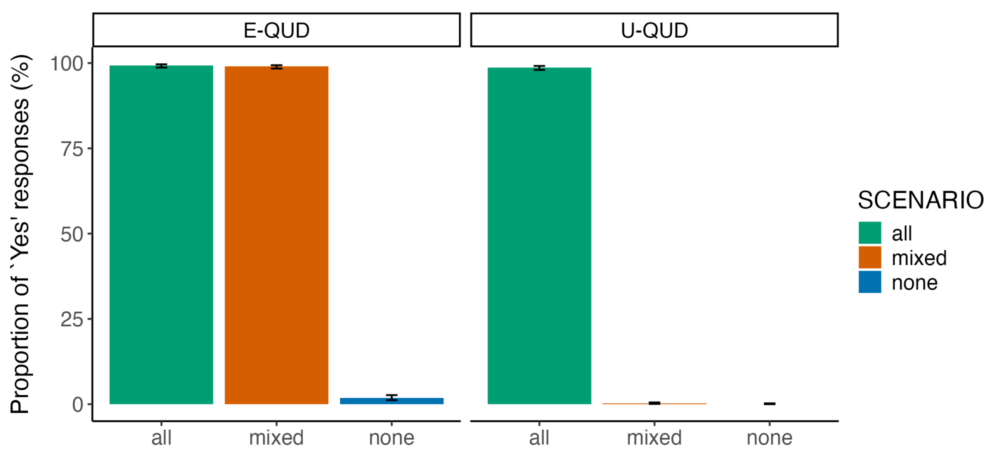


Figure 7 Mean proportion of 'Yes' responses to the QuD check as a function of the scenario and QuD group. Error bars denote 95% confidence intervals.

6.7.2 Plural definite sentences

To further evaluate the success of the QuD manipulation, we analyzed the truth-value slider ratings for the plural definite sentences. For plural definites, we predicted a QUD effect: higher ratings in the E-QUD group compared to the U-QUD in the *mixed* scenario. We fit a linear mixed effects model to slider ratings in the *mixed*-scenario with the QUD group as a predictor. We found a significant effect of the QuD ($\beta = -12.6, p = 0.01$), indicating that the ratings were higher in the E-QUD group ($M = 42.2$, $CI_{95\%}[47.6; 36.9]$) than in the U-QUD group ($M = 29.6$ $CI_{95\%}[29.6; 25.5]$), as shown in Figure 8. These findings, together with the results of the QuD check task, confirm that the QuD manipulation was effective.

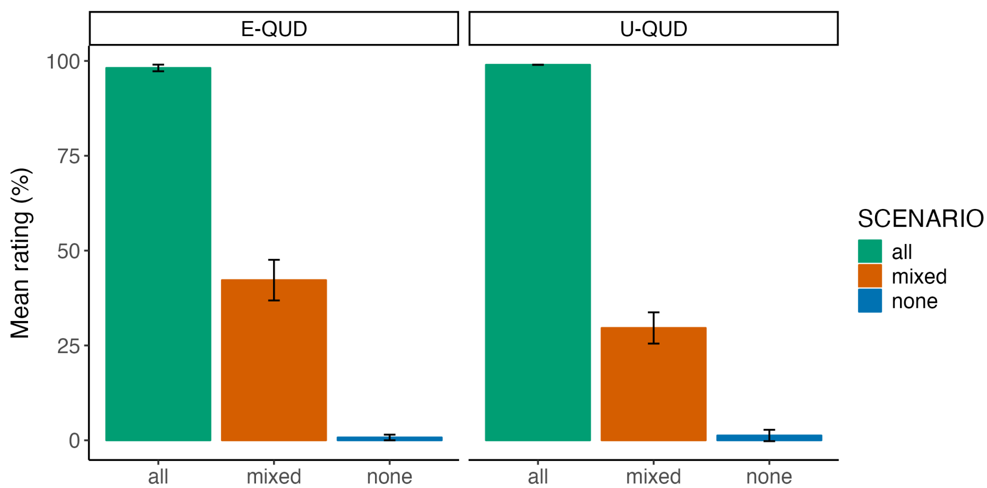


Figure 8 Mean rating for plural definite (PD) sentences as a function of the type of scenario and the QuD group. Error bars denote 95% confidence intervals.

6.7.3 Target counterfactual sentences

To examine the effect of the QuD manipulation on the target counterfactual sentences, we fit a linear mixed-effects model with slider value as the dependent variable and POLARITY, STRENGTH, and QUD, and their interactions as predictors (coding schema: 0.5 and −0.5) using a maximal random effect structure. The analysis revealed significant effects of STRENGTH, and POLARITY as well as their interaction (see Table 5). However, neither the main effect of QuD nor the interaction effects — QuD×STRENGTH, QuD×POLARITY and the three-way interaction — were significant.

Neither the POLARITY effect, the QuD effect, nor the interaction between these predictors was significant for strong quantifiers (see Table 5). The mean slider ratings for the U-QuD group were 3.3 ($CI_{95\%}$[5.9; 0.8]) for *none* and 1.5 ($CI_{95\%}$[3.2; -0.3]) for *all*. For the E-QuD group, the ratings were 0.9 ($CI_{95\%}$[1.8; -0.004]) for *none* and 1.2 ($CI_{95\%}$[2.4; 0.2]) for *all*. In contrast, ratings for weak quantifiers showed differences. For the U-QuD group, the mean slider ratings were 82.1 ($CI_{95\%}$[87.8; 76.5]) for *not all* and 96.1 ($CI_{95\%}$[97.8; 94.4]) for *some*. For the E-QuD group, the ratings were 86.1, ($CI_{95\%}$[91.9; 80.4]) for *not all* and 97.2 ($CI_{95\%}$[98.5; 95.9]) for *some*. Thus, in the case of counterfactual sentences with positive quantifier *some*, participants move the slider further towards the 'completely true' end of the scale than in the case of the negative quantifier *not all*. Crucially, however, the effects of QuD and its interaction with POLARITY and STRENGTH did not yield significant results (see Table 5).

	Whole		*Strong*		*Weak*	
Effect	β	p-value	β	p-value	β	p-value
INTERCEPT	46.1	$< .001$	1.74	< 0.01	90.41	$< .001$
STRENGTH	−88.7	$< .001$	—		—	
QuD	−0.6	0.7	1.3	0.3	−2.6	0.4
POLARITY	5.9	$< .001$	−0.7	0.4	12.5	$< .001$
QuD:POLARITY	0.3	0.9	−2.3	0.2	3.0	0.6
STRENGTH:QuD	3.9	0.2	—		—	
STRENGTH:POLARITY	−13.2	$< .001$	—		—	
STRENGTH:POLARITY:QuD	−5.3	0.4	—		—	

Table 5 Results of the mixed-effects models for the target counterfactual sentences: the *whole* model includes all sentence types; *Strong* includes strong quantifiers only, and *Weak* includes weak quantifiers only.

To sum up, the QuD manipulation did not have an effect on the truth value judgments of the target counterfactual sentences. Additionally, sentences with the positive weak quantifiers were rated higher than those with the negative weak quantifier in the *mixed*-scenario. We also found that the results for the target counterfactual sentences were replicated in the control counterfactual sentences,

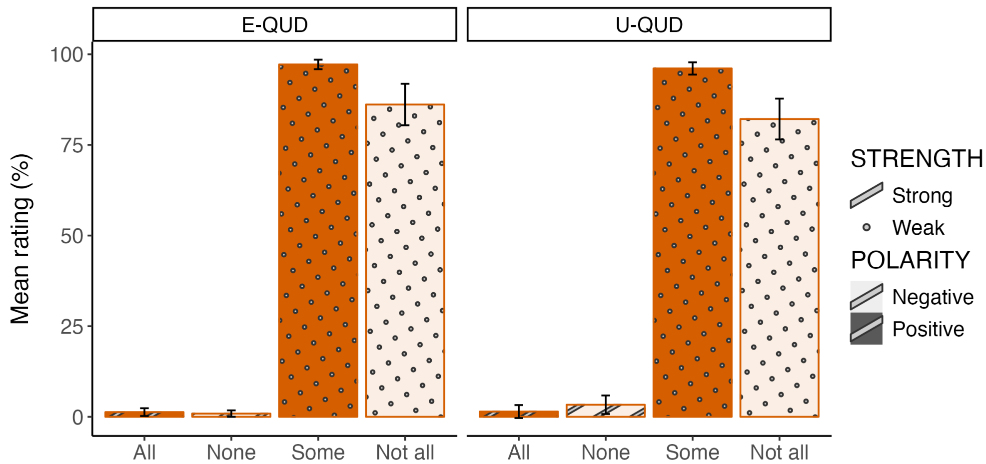


Figure 9 Mean rating for target counterfactual (TC) sentences in *mixed* scenarios by QuD group. The pattern of the bars represents the quantifier's strength and the transparency indicates its polarity. Error bars denote 95% confidence intervals.

indicating that the QUD effect was not significant either ($\beta = -0.9, p = .48$; see full model statistics in Appendix A on OSF).

6.8 Synthesis

Experiment 2 aimed to follow up on the findings of Experiment 1 by investigating whether the absence of a QuD effect on truth-value judgments for counterfactuals was due to an insufficiently robust QuD manipulation. To this end, sentences containing plural definites were tested alongside counterfactuals in a novel setup designed to enhance the QuD manipulation. The results reveal that plural definites were sensitive to the QuD manipulation, while counterfactuals were not. Specifically, in the *mixed*-scenarios, plural definites were significantly more accepted when the relevant QuD was existential than when it was universal. By contrast, the results for counterfactuals replicate those of Experiment 1, showing that counterfactuals embedded under weak quantifiers were significantly more accepted than those embedded under strong quantifiers in the *mixed*-scenarios.

7 General discussion

The findings of Marty, Romoli & Santorio 2020 aligned with selectional theories while posing challenges for both homogeneity and universal theories. As discussed, however, the homogeneity theory can accommodate these findings if supplemented with a mechanism that pragmatically reinterprets undefinedness based on relevance. In this study, we tested this theoretical possibility by manipulating the type of QuD and broadening the range of embedding environments examined. The results of Experiment 1 revealed no effect of QuD on the truth-value judgment task. Instead, an effect of quantifier strength was observed, consistent with selectional theories but contrary to the predictions of the homogeneity approach. Furthermore, Experiment 2 demonstrated that while the QuD manipulation influenced judgments in the case of plural definites, it had no effect on the evaluation of counterfactual sentences. Taken together, these findings provide compelling evidence in support of selectional theories, while challenging both universal and homogeneity theories.

The findings of Experiment 2 also directly address the similarities and differences between counterfactuals and plural definites. Indeed, plural definites are often regarded as the paradigmatic case of homogeneity. However, analyses of definites based on homogeneity remain controversial, with alternative theories, such as those grounded in implicature, being proposed. What is particularly noteworthy is that our results for plural definites align with prior experimental literature (Augurzky et al. 2023 a.o.) but differ markedly from what we observed for counterfactuals, both in terms of the effect of QuD and the response patterns across quantifier conditions. The absence of a QuD effect for counterfactuals cannot easily be attributed to a lack of statistical power or a failed QuD manipulation, as the experiment successfully detected a QuD effect in sentences with plural definites. These findings suggest that the analogy between conditionals and plural definites, suggested by several authors in the literature, may be on the wrong track.

8 A note on the implicature approach

The literature also includes a variant of the homogeneity approach that involves implicatures (Bassi & Bar-Lev 2018).[13] Our experiments are also relevant for this approach.

13 Bassi and Bar-Lev don't directly discuss counterfactuals, but their account can be seamlessly extended to them. See also Herburger 2015b,a for a similar account, though not explicitly based on implicature.

Bassi & Bar-Lev suggest that conditionals have a basic existential meaning (23), which is strengthened to a universal one via implicature (24). The mechanisms behind this implicature are not central to our discussion. For simplicity, we represent the presence of this implicature using the 'EXH' operator.

(23) a. If ticket #37 was bought, it would win a prize
b. $\exists w'$: $w' \in \text{MAX}_{w,\preceq}(\llbracket \mathsf{A} \rrbracket_{\preceq}), \llbracket \mathsf{B} \rrbracket^{w',\preceq} = \text{true}$

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

In unembedded cases, the strengthened meaning is the dominant interpretation. But the basic meaning is predicted to reemerge in environments where implicatures tend to not arise, such as in downward entailing environments. This prediction is compatible with the results in Marty, Romoli & Santorio 2020.

The results of our experiments, however, are challenging for this approach. To illustrate, consider (25).

(25) a. Some of the tickets$_x$ EXH[x would win a prize if it was bought].
b. $\exists x [\forall w'$: $w' \in \text{MAX}_{w,\preceq}(\llbracket \mathsf{P}x \rrbracket_{\preceq}), \llbracket \mathsf{Q}x \rrbracket^{w',\preceq} = \text{true}]$

The nuclear scope of *some* is an upward entailing environment. Hence we expect the implicature to be computed. This yields the prediction that (25) should mean, roughly, that some ticket is guaranteed to win a prize if bought. These truth conditions are false in our *mixed*-contexts. But this prediction is not in line with our results.

The case of *some* may be accommodated via a global, rather than a local, derivation of implicatures. But the case of *not every* in (26) is more challenging.

(26) a. Not every ticket would win a prize if it was bought.
b. $\neg \forall x [\exists w'$: $w' \in \text{MAX}_{w,\preceq}(\llbracket \mathsf{P}x \rrbracket_{\preceq}), \llbracket \mathsf{Q}x \rrbracket^{w',\preceq} = \text{true}]$

The reason is that now we have a downward entailing environment. Hence the implicature account predicts that the conditional will contribute its basic existential meaning. As a result, (26) should mean roughly: *some of the tickets would not win a prize, if they were bought*. These truth conditions are, again, false in our *mixed* scenarios. But the results of our experiments show that these sentences are fully accepted in these contexts.

Finally, given that implicatures are sensitive to relevance, on the implicature approach we would expect that interpretation of sentences like (25) should be affected by the QuD. This prediction is, again, not borne out.

In sum, Experiments 1 and 2 present challenges for the implicature approach for two key reasons: (i) we observe no effect of QuD in the judgment task and (ii), more importantly, the results for counterfactuals involving *some* and *not every/all* are not in line with the predictions of this approach.

9 Conclusion

In this paper, we reported on two experiments designed to test the predictions of three major theoretical approaches to counterfactuals: universal, selectional, and homogeneity theories. The critical cases examined in these experiments were specifically constructed to differentiate the key predictions of these theories. The findings from both experiments support selectional theories, while challenging the predictions of universal and homogeneity theories.

While our focus has been on counterfactuals, it is worth considering what results might emerge from parallel experiments on indicative conditionals. Theoretically, it is both desirable and plausible that the logics of counterfactuals and indicatives are either identical or closely related (following Stalnaker 1968, and *pace* Lewis 1973a). So, if counterfactuals are selectional, we should expect that indicatives should be too. We conjecture that experimental findings would support this analogy. But the validation of this conjecture must be left to future research.

References

Abrusán, Márta & Kriszta Szendröi. 2012. Experimenting with the king of France. In Maria Aloni, Vadim Kimmelman, Floris Roelofsen, Galit W. Sassoon, Katrin Schulz & Matthijs Westera (eds.), *Logic, language and meaning: 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19-21, 2011, Revised selected papers* (Lecture Notes in Computer Science (LNCS) 7218), 102–111. Berlin: Springer. https://doi.org/10.1007/978-3-642-31482-7_11.

Agha, Omar. 2023. *How to ignore counterexamples: Homogeneity across times and worlds*. New York, NY: New York University (NYU) dissertation. https://omaragha.org/wp-content/uploads/2023/09/agha_dissertation_final.pdf.

Alxatib, Sam & Jeff Pelletier. 2011. On the psychology of truth-gaps. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *Vagueness in communication: International workshop, VIC 2009, held as part of ESSLLI 2009, Bordeaux, France, July 20-24, 2009. Revised selected papers*

(Lecture Notes in Computer Science (LNAI) 6517), 13–36. Berlin: Springer. https://doi.org/10.1007/978-3-642-18446-8_2.

Augurzky, Petra, Marion Bonnet, Richard Breheny, Alexandre Cremers, Cornelia Ebert, Clemens Mayr, Jacopo Romoli, Markus Steibach & Yasutada Sudo. 2023. Putting plural definites into context. *Sinn und Bedeutung (SuB)* 27. 19–32. https://doi.org/10.18148/sub/2023.v27.1050.

Bassi, Itai & Moshe Bar-Lev. 2018. A unified existential semantics for bare conditionals. *Sinn und Bedeutung (SuB)* 21(1). 125–142. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/128.

Champollion, Lucas, Dylan Bumford & Robert Henderson. 2019. Donkeys under discussion. *Semantics & Pragmatics* 12(1). 1–50. https://doi.org/10.3765/sp.12.1.

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. https://doi.org/10.1093/jos/14.1.1.

von Fintel, Kai. 1999. NPI licensing, Strawson entailment, and context dependency. *Journal of Semantics* 16(2). 97–148. https://doi.org/10.1093/jos/16.2.97.

von Fintel, Kai. 2001. Counterfactuals in a dynamic context. In Michael Kenstowicz (ed.), *Ken hale: A life in language* (Current Studies in Linguistics Series 36), 123–152. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/4056.003.0006.

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. https://doi.org/10.1017/CBO9781139519328.011.

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

George, BR. 2008. *Presupposition repairs: A static, trivalent approach to predicting projection*. Los Angeles, CA: University of California, Los Angeles MA thesis. https://linguistics.ucla.edu/general/matheses/George_UCLA_MA_2008.pdf.

Gillies, Anthony S. 2007. Counterfactual scorekeeping. *Linguistics and Philosophy* 30(3). 329–360. https://doi.org/10.1007/s10988-007-9018-6.

Herburger, Elena. 2015a. Conditional perfection: The truth and the whole truth. *Semantics and Linguistic Theory (SALT)* 25. 615–635. https://doi.org/10.3765/salt.v25i0.3079.

Herburger, Elena. 2015b. *Only if*: If only we understood it. *Sinn und Bedeutung (SuB)* 19. 284–301. https://doi.org/10.18148/sub/2015.v19i0.235.

Higginbotham, James. 1986. Linguistic theory and Davidson's program in semantics. In Ernest LePore (ed.), *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*, 29–48. Cambridge: Blackwell.

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. https://doi.org/10.1162/002438900554352.

Kaufmann, Stefan. 2017. The limit assumption. *Semantics & Pragmatics* 10(18). 1–29. https://doi.org/10.3765/sp.10.18.

Khoo, Justin. 2022. *The meaning of "if"*. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190096700.001.0001.

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. *Chicago Linguistics Society (CLS)* 22(2). 1–15.

Kratzer, Angelika. 2012. *Modals and conditionals: New and revised perspectives* (Oxford Studies in Theoretical Linguistics 36). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199234684.001.0001.

Križ, Manuel. 2015. *Aspects of homogeneity in the semantics of natural language*. Vienna: University of Vienna dissertation. https://manuel.kriz.fr/wp-content/uploads/Kriz-2015-Aspects-of-Homogeneity.pdf.

Križ, Manuel. 2016. Homogeneity, maximality, and *all*. *Journal of Semantics* 33(3). 493–539. https://doi.org/10.1093/jos/ffv006.

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. https://doi.org/10.1007/s11050-015-9114-z.

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: Harvard University Press.

Lewis, David K. 1973b. Counterfactuals and comparative possibility. *Journal of Philosophical Logic* 2(4). 418–446. https://doi.org/10.1007/BF00262950.

Lewis, David K. 1979. Counterfactual dependence and time's arrow. *Noûs* 13(4). 455–476. https://doi.org/10.2307/2215339.

Mandelkern, Matthew. 2016. Dissatisfaction theory. *Semantics and Linguistic Theory (SALT)* 26. 1–50. https://doi.org/10.3765/salt.v26i0.3804.

Marty, Paul, Jacopo Romoli & Paolo Santorio. 2020. Counterfactuals and undefinedness: Homogeneity vs. supervaluations. *Semantics and Linguistic Theory (SALT)* 30. 603–623. https://doi.org/10.3765/salt.v30i0.4842.

Mizuno, Teruyuki. 2025. On the structures and meanings of *might*-counterfactuals: A view from Japanese. *Linguistics and Philosophy* 48(3). 449–504. https://doi.org/10.1007/s10988-025-09433-z.

Roberts, Craige. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics, 1998 revision. In Jae Hak Yoon & Andreas Kathol (eds.), *Papers in semantics* (Working Papers in Linguistics 49), 91–136. The Ohio State University, Department of Linguistics. A revised version was published in Roberts 2012.

Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics* 5(1). 1–69. https://doi.org/10.3765/sp.5.6.

RStudio Team. 2020. *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA. http://www.rstudio.com/.

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

Santorio, Paolo. 2024. Bounds for metaphysical modality. *Amsterdam Colloquium* 24. 316–322. https://platform.openjournals.nl/PAC/article/view/21894.

Schlenker, Philippe. 2004. Conditionals as definite descriptions. *Research on Language and Computation* 2(3). 417–462. https://doi.org/10.1007/s11168-004-0908-2.

Schwarz, Florian. 2016. False but slow: Evaluating statements with non-referring definites. *Journal of Semantics* 33(1). 177–214. https://doi.org/10.1093/jos/ffu019.

Serchuk, Phil, Ian Hargreaves & Richard Zach. 2011. Vagueness, logic and use: Four experimental studies on vagueness. *Mind & Language* 26(5). 540–573. https://doi.org/10.1111/j.1468-0017.2011.01430.x.

Stalnaker, Robert C. 1968. A theory of conditionals. In Nicholas Rescher (ed.), *Studies in logical theory*, 98–112. Oxford: Blackwell. Reprinted as Stalnaker 1981b.

Stalnaker, Robert C. 1981a. A defense of conditional excluded middle. In William Harper, Robert C. Stalnaker & Glenn Pearce (eds.), *Ifs: Conditionals, belief, decision, chance and time* (The Western Ontario Series in Philosophy of Science 15), 87–104. Dordrecht: Reidel. https://doi.org/10.1007/978-94-009-9117-0_4.

Stalnaker, Robert C. 1981b. A theory of conditionals. In William L. Harper, Robert C. Stalnaker & Glenn Pearce (eds.), *Ifs: Conditionals, belief, decision, chance and time* (The Western Ontario Series in Philosophy of Science 15), 41–55. Dordrecht: Reidel. https://doi.org/10.1007/978-94-009-9117-0_2.

Stalnaker, Robert C. 1984. *Inquiry*. Cambridge: MIT Press.

Syrett, Kristen & Todor Koev. 2015. Experimental evidence for the truth conditional contribution and shifting information status of appositives. *Journal of Semantics* 32(3). 525–577. https://doi.org/10.1093/jos/ffu007.

Tieu, Lyn, Cory Bill & Jacopo Romoli. 2019. Homogeneity or implicature: An experimental investigation of free choice. *Semantics and Linguistic Theory (SALT)* 29. 706–726. https://doi.org/10.3765/salt.v29i0.4631.

Wickham, Hadley. 2016. *Ggplot2: Elegant graphics for data analysis*. New York: Springer. https://ggplot2.tidyverse.org.

Willer, Malte. 2022. Negating conditionals. In Ernest Lepore & David Sosa (eds.), *Oxford studies in philosophy of language*, vol. 2, 234–266. Oxford University Press. https://doi.org/10.1093/oso/9780192844613.003.0008.

Zehr, Jeremy & Florian Schwarz. 2018. *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832.

Sonia Ramotowska
Institute for Logic, Language and Computation
University of Amsterdam
Science Park 107
1098 XG, Amsterdam
The Netherlands
ramotowska.or.s@gmail.com

Paul Marty
Centro de Linguística
Universidade de Lisboa
Alameda da Universidade
1600-214 Lisbon
Portugal
pmarty@edu.ulisboa.pt

Jacopo Romoli
Institute für Sprache und Information
Heinrich Heine Universität Düsseldorf
Universitätsstraße 1
40225 Düsseldorf
Germany
jacopo.romoli@hhu.de

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