Counterfactuals and undefinedness: homogeneity vs. supervaluations

Overview. Theories of counterfactuals agree on appealing to a relation of comparative similarity, but disagree on the quantificational force of counterfactuals. We report on an experiment testing the predictions of three main approaches: universal theories, homogeneity theories, and single-world selection theories (plus supervaluations over selection functions). Our results provide empirical support for the selectional/supervaluational theories while disconfirming the other two competing approaches.

Three theories. On standard approaches, counterfactuals are modalized sentences that appeal to a relation of comparative similarity and whose truth conditions conform to the schema in (1):

(1) A $\Box\!\!\rightarrow$ C is true at w relative to $\preceq$ iff $\underline{\text{QUANT } w'}$: $[w' \in \textsc{max}_{w,\preceq}([\![\text{A}]\!])][w' \in [\![\text{C}]\!]]$

Yet standard theories differ in their analysis of the quantificational force of counterfactuals. On UNIVERSAL THEORIES (U-theories; Lewis 1973, Kratzer 1981 a.o.), counterfactuals have universal force. On HOMOGENEITY THEORIES (H-theories; von Fintel 1997, Schlenker 2004 a.o.), they have universal force, supplemented by a definedness condition that requires that either all or no relevant worlds satisfy the consequent. On SELECTIONAL THEORIES (S-theories; Stalnaker 1968, 1980), counterfactuals exploit a selection function that takes as argument a world and an antecedent and that returns a 'selected' world; a counterfactual is true iff the consequent is true at that world. Since context is often insufficient to fix a value for the selection function, S-theories are accompanied by a supervaluational definition of determinate truth and determinate falsity. To illustrate, consider (2) together with the truth conditions predicted by each theory.

(2) If ticket #37 were bought, it would win a prize.

Universal.	$[\![2]\!]^{w,\preceq}$ =	TRUE iff $\forall w'$: $w' \in \textsc{max}_{w,\preceq}([\![\text{\#37 is bought}]\!])$, #37 wins in w'
Homogeneity.	$[\![2]\!]^{w,\preceq}$ =	DEFINED iff $\forall w'$: $w' \in \textsc{max}_{w,\preceq}([\![\text{\#37 is bought}]\!])$, #37 wins in w', or $\forall w'$: $w' \in \textsc{max}_{w,\preceq}([\![\text{\#37 is bought}]\!])$, #37 doesn't win in w';
		TRUE iff $\forall w'$: $w' \in \textsc{max}_{w,\preceq}([\![\text{\#37 is bought}]\!])$, #37 wins in w'
Selectional.	$[\![2]\!]^{w,s}$ =	TRUE iff $[\![\text{\#37 wins}]\!]^{s(w,[\![\text{\#37 is bought}]\!]),s}$
	(2) is DETERMINATELY TRUE[FALSE] at c iff, $\forall s$ compatible with c, $[\![(2)]\!]^{w,s}$ = true[false]	

Undefinedness projection. H-theories and S-theories yield different predictions about undefinedness in complex sentences. On H-theories, undefinedness projection follows the strong Kleene algorithm.[1] For connectives and quantifiers at least, this yields that whether A is undefined is fixed by the truth status of the constituents of A. S-theories use supervaluations, on which the truth status of a complex sentence is not determined by the truth status of the constituents. E.g., some disjunctions with undefined disjuncts are also undefined, but instances of Excluded Middle (A or not A) are determinately true even when A and $\neg$A are undefined. This difference leads to key different predictions for the cases we focused on.

Key predictions. Consider a raffle where prize-winning tickets are selected via a random draw among all the tickets bought. Simple counterfactuals like (2) are predicted to be false by U-theories and undefined by H-theories and S-theories. In addition, embedding a similar clause under a negative quantifier, as in (3), distinguishes between the three theories. (3) is predicted to be true by U-theories, undefined by H-theories (since the prejacent of *no ticket* is undefined, for each ticket in the domain of the quantifier), and false by S-theories (since, for all choices of selection function, some tickets are going to win prizes). Our experiment tested these key predictions.

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

Experiment. We ran an experiment testing the predictions of the three approaches above, building on previous work by Križ & Chemla 2015 and Tieu et al. 2019, among others.

<u>PARTICIPANTS.</u> 99 self-identified native speakers of English, recruited through Prolific, participated in the study (mean age $=$ 36 yrs, 47 female) and were paid \$1.5 for their time ($\approx$ 10 min).

<u>MATERIALS.</u> Each item consisted of a short context followed by a test sentence. Each context described the working of one of three kinds of raffles: (i) one in which all the tickets bought win a prize (*all*-contexts),

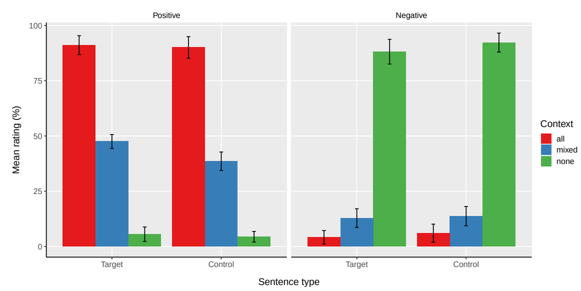
[1] Križ 2015 suggests modifications of the strong Kleene algorithm, based on data with definites in non-monotonic quantifiers. Those modifications do not affect our argument, so we put them aside here.

(ii) one in which only half of the tickets bought win a prize (*mixed*-contexts), and (iii) one in which none of the tickets bought win a prize (*none*-contexts). Test sentences involved two types of TARGET: simple counterfactuals (POSITIVE) and counterfactuals embedded under *no* (NEGATIVE). Each TARGET was paired with a corresponding CONTROL (see (4)). Those CONTROL are not predicted on any approach to give rise to undefinedness, and they are thus expected to be judged as false in the critical conditions for the TARGET, i.e. with the *mixed*-contexts. Crossing Context and Sentence types gave rise to 12 test items. 12 filler items were further included to disguise the purpose of the experiment. Filler items involved contexts similar to those used in the test items, but were followed by non-counterfactual sentences.

(4)
a. If ticket #37 was bought, it would win a prize. TARGET-POSITIVE
b. If ticket #37 was bought, it would have to win a prize. CONTROL-POSITIVE
c. No ticket would win a prize, if it was bought. TARGET-NEGATIVE
d. No ticket could win a prize, if it was bought. CONTROL-NEGATIVE

TASK. Participants were asked to read each story and then assess the extent to which the test sentence was true in the context of that story. They reported their judgements by setting the position of a slider tooltip along a line going from 'Completely false' to 'Completely true.' They were instructed to move the slider to the right if they judged the sentence as true, to the left if they judged it as false, and to the middle of the line if they found the sentence neither completely false, nor completely true. Participants started with 2 (unannounced) practise trials to get familiar with the experimental display. The 24 (12 test+12 filler) items were then presented in random order.

RESULTS. Mean ratings to test items are shown in the graph on the right. Participants' responses to POSITIVE and NEGATIVE sentences were analyzed using LMER models (see description and outputs in the table below the graph). For the NEGATIVE sentences, there was only a main effect of Context: ratings to the TARGET and the CONTROL were overall very low in the *mixed*-contexts (≈13%), and only slightly higher than in the *all*-contexts (≈5%). Most importantly, there was no difference in the patterns of responses for TARGET-NEGATIVE and CONTROL-NEGATIVE. For the POSITIVE sentences, there was a main effect of Context (*mixed*>*none*), a main effect of Status (TARGET>CONTROL) and a significant interaction between these factors. In a nutshell, TARGET-POSITIVE sentences, unlike TARGET-NEGATIVE ones, received an intermediate rating in the MIXED-contexts (≈48%) and this was significantly higher than that of its corresponding control in the same contexts (≈38%).



Sentence	Fixed Effect	β	sd	t-value	χ^2	p-value
Positive	Status	9	1.8	4	14	$< .0005$
	Context	-34	2.1	-15	486	$< .0001$
	Status:Context	-8	2.6	-3	8	$< .005$
Negative	Status	-1.8	2	-0.9	0.9	0.3
	Context	7.7	2.1	3.5	24	$< .0001$
	Status:Context	0.9	2.8	0.3	0.1	0.7

`Formula:Response~Status*Context+(Context|Subject)`

Discussion. In POSITIVE sentences, we observed clear intermediate ratings for both TARGET and CONTROL items. This suggests that the latter were not an ideal baseline for removing undefinedness after all. Importantly, however, there is still a reliable difference between TARGET and CONTROL items for those sentences. These results are in line with S-theories and H-theories, as the intermediate endorsement of the target sentences can be taken to be indicative of the undefinedness predicted by those approaches, while they are challenging for U-theories, which predicted them to be judged false. In NEGATIVE sentences, which were the crucial case for discriminating among the predictions of the three approaches, the endorsement rate of the TARGET in *mixed*-contexts, while slightly higher than that in the *all*-contexts, was overall very low and no different from that of the corresponding CONTROL items. This suggests that both TARGET and CONTROL sentences were essentially judged false in those contexts, in line with the predictions of the S-theories and against those of the other two approaches.

Selected References: Klinedinst, N. 2011. Quantified Conditionals and Conditional Excluded Middle. • Kratzer, A. 1981. Partition and Revision: The Semantics of Counterfactuals • Križ, M. 2015. Aspects of Homogeneity in the Semantics of Natural Language. • Lewis, D.K. 1973. *Counterfactuals* • Schlenker, P. 2004. Conditionals as Definite Descriptions