# *Cause*, *make*, and *force* as graded causatives

Angela Cao, Aaron Steven White, & Daniel Lassiter*

**Abstract.** We investigate the semantics of the causal verbs *cause*, *make*, and *force* as used in the construction *X* {*caused/made/forced*} *Y (to) Z*. The predominant approach to analyzing verbs of causing has been to argue that they convey some version of SUFFICIENCY, but it has also been suggested that INTENTION or possible ALTERNATIVES may also factor into the semantics of the verbs. Using sequences of tic-tac-toe states as experimental stimuli, we measure the three possible contributing factors in each stimuli and ask participants whether each verb is appropriate for describing the sequence. We find experimental support for a differentiating semantics of these verbs, in which no single predictor is the sole factor in when each verb is appropriate.

**Keywords.** semantics; causatives; causal models; psycholinguistics

**1. Introduction.** The predominant approach to analyzing verbs of causing has been to argue that they convey some version of SUFFICIENCY, which is measured given parameters of a causal situation (Nadathur & Lauer 2020, Lauer & Nadathur 2018, Glass 2023, Schulz 2011). Notably, some of this work has leveraged structural causal models (SCMs; Pearl, 2009) to model and make predictions about how we use these verbs (Baglini & Siegal 2021, Nadathur & Siegal 2022, Schulz 2011). In this paper, we argue that the semantics of causing verbs encode not only sufficiency but also intention and the number of feasible alternative actions. To support this argument, we provide experimental evidence for a differentiating semantics of three causing verbs using explicitly-defined causal models, which enable us to calculate measures derived from the stimuli. We are thus able to quantify concepts including SUFFICIENCY and use them as predictors of judgements.

Our objects of study are the three English periphrastic causatives *cause*, *make*, and *force*. We investigate the meaning of these three verbs when used in the linguistic constructions of the form in (1).

(1) $X \left\{ \begin{array}{c} \text{caused} \\ \text{made} \\ \text{forced} \end{array} \right\} Y \textit{ (to) } Z.$

where *X* and *Y* refer to entities that can be agents, and *Z* describes an action—e.g. (2).

(2) [The pirate]$_X$ forced [the prisoner]$_Y$ to [walk down the plank]$_Z$.

These verbs are of interest because they are clearly not interchangeable, in spite of the fact that they all seem to express a similar kind of causation. Consider the following examples, which were identified via pre-existing datasets (Cao et al. 2022, Williamson et al. 2023, Davies 2008–) and modified (where [] indicates insertion/deletion/replacement) for our purposes:

(3) a. A CAT [...] caused himself [to] look as much as possible like a doctor...

*This research underwent ethical review by the LEL Research Ethics panel at the University of Edinburgh. We would like to thank Lelia Glass, Julian Grove, Joshua Knobe, and the participants at ELM for useful feedback. Corresponding author: Angela Cao, University of Rochester (acao9@ur.rochester.edu); also, Aaron Steven White, University of Rochester and Daniel Lassiter, University of Edinburgh.

b. A CAT [...] made himself look as much as possible like a doctor... (Fables[1])
c. A CAT [...] forced himself [to] look as much as possible like a doctor...

(4) a. He caused cancer in one woman. (SPOK: THE FIVE 5:00 PM EST; 2013)
b. *He made cancer [happen] in one woman.
c. *He forced cancer [to happen] in one woman.

That these verbs appear not to be mutually replaceable is of interest, especially since some prior experimental work has analyzed them as have similar meanings, e.g. that *cause*, *make*, and *force* indicate that the causee did not have a tendency towards the result, the causer and causee were not in concordance, and that the result actually occurred (Klettke & Wolff 2003, Wolff et al. 2005). There is also extensive work arguing that there are important semantic distinctions between these three verbs. For example, Nadathur & Lauer (2020), Lauer & Nadathur (2018) argue that *make* denotes causal sufficiency, while *cause* denotes causal necessity. Furthermore, Shibatani (1976) proposes that periphrastic verbs lay on a continuous scale of directness, with varying degrees of control and agency exerted by the causer over the causee. Similarly, Childers (2016) argues that periphrastic causatives can be ordered on a single causee inclination continuum, in which *force* denotes the most direct compulsion. This is described as when the causee is non-cooperative, but has no right of refusal. Evidently, both the concepts of causee inclination and direct compulsion are related to our previous discussion of sufficiency, since a greater causee inclination requires a smaller degree of compulsion from the causer to bring about the result, and directly compulsing the causee to bring about an intended result is completely sufficient for bringing about the result.

Notably, much of this aforementioned work makes use of the logics of structural causal models (SCMs) from Pearl (2009), which has previously been used to model causal relations between events as well as their counterfactual values. In our paper, we focus on the constructions *X caused/made/forced Y (to) Z* and argue that the relationship between the verbs *cause*, *make*, and *force* is structured not by sufficiency, intentionality, or alternatives alone, but by some interactions of (at least) these three. In order to support our argument, we run an experiment in which participants' judgements of when the three verbs are appropriate in describing tic-tac-toe sequences is predicted by measures defined using the logics of structural causal models (Pearl 2009).

**2. Possible scales.** Consider examples (5)–(7).

(5) a. I **caused** Martha to go to the gym by mentioning how the habit has helped me.
b. *I **made** Martha go to the gym by mentioning how the habit has helped me.
c. *I **forced** Martha to go to the gym by mentioning how the habit has helped me.

(6) a. I **caused** Martha to go to the gym by criticizing her physical appearance.
b. ?I **made** Martha go to the gym by criticizing her physical appearance.
c. *I **forced** Martha to go to the gym by criticizing her physical appearance.

(7) a. I **caused** Martha to go to the gym by holding her child hostage.
b. I **made** Martha go to the gym by holding her child hostage.
c. I **forced** Martha to go to the gym by holding her child hostage.

[1] https://www.gutenberg.org/cache/epub/21/pg21.txt

It appears that the acceptability of the causal verb being used is modulated by several attributes of the causal relata. Specifically, the examples give rise to the question – what are the significant differences between the causing events of (5-a) "mentioning how the habit has helped me", (5-b) "criticizing her physical appearance", and (5-c) "holding her child hostage"? There are multiple possible analyses – one approach is that (5-c) affects the patient's space of safe alternatives in a way that (5-a) and (5-b) do not. That is, in the worlds of (5-a) and (5-b), Martha can choose not to go to the gym without drastic consequences. In contrast, in the worlds of (5-c), Martha's child might be killed (if the narrator is truthful). Relatedly, then, the causing events of (5-a), (5-b), and (5-c) can also be characterized by varying on how sufficient each was in bringing about the effect of *Martha going to the gym*. Intuitively, (c) is most sufficient in bringing about the effect. Finally, a possible characterization of (5-a)–(5-c) is that the agent of each varies in how *intentional* they were for the occurrence of the effect. Naturally, the agent in (5-c) seems drastically committed to causing Martha to go to the gym, moreso than the narrators of (5-a) and (5-b).

Based on the aforementioned examples and literature, we postulate that these causatives have a semantics built around threshold values on a continuous scale. We consider three measures that are relevant features of causal relationships based on our previous discussion: number of alternatives (ALT), intention (INT), and sufficiency (SUF).

2.1. STRUCTURAL CAUSAL MODELS. How can we quantify these concepts? A natural choice is to use games – particularly, the highly constrained, zero-sum game of tic-tac-toe. At every sequence of consecutive moves, the second player had some number of alternatives to the move they ended up taking, the first player intended the second player to make the move it did to some degree, and the first player's move was, to some degree, sufficient for bringing about the second player's action. Previous work such as Hammond et al. (2023) has instantiated examples of games in the framework of structural causal models (SCMs; termed *structural causal games*) for the purpose of formalizing agents and their interactions within a grounded, incentivized situation; furthermore, other work such as Halpern & Kleiman-Weiner (2018), Nadathur & Lauer (2020), Lauer & Nadathur (2018) and Pearl (2019) has defined concepts such as sufficiency and intention within this framework. Thus, we can use SCMs to measure values of alternatives, intention, and sufficiency in tic-tac-toe sequences, which we then use as predictors in participant judgements of when *cause*, *make*, and *force* are accurate in describing the sequences.

We define structural causal models in the sense of Pearl (2009). SCMs carve up causal relationships into a discrete set of independent and dependent variables, with defined mechanisms that structurally define variables' relationships with one another.[2]

**Definition 1** (Structural Causal Models)**.** We define a **time-indexed causal model** $\mathcal{M}$ to consist of:

- **Exogenous Variables** ($\mathcal{U}$) where each variable $X_t$ has an associated set of **values** it can take on $\mathsf{Val}(X_t)$. Exogenous variables have no parents.
- **Endogenous Variables** ($\mathcal{V}$) where each variable $Y_t$ has an associated set of **values** it can take on $\mathsf{Val}(Y_t)$ and a timestep $t \in \{0, 1, 2, \dots\}$. Endogenous variables have parents.

[2]The following definitions are also used by Cao et al. (2023) for developing a semantics of causing, enabling, and preventing verbs.

- **Causal Structure** ($\mathcal{F}$) represented by arrows running from "parent" variables to "child" variables, which also encode a node's value based on the value of its parents. We require that all parents immediately precede their children. Equivalently, if $P_t$ is a parent of $C'_t$, then $t = t' - 1$.

The relevant operation of causal models is an *intervention*, which fixes the value(s) of some variable(s). This action may have downstream changes, but can not affect upstream variables. This term is useful for our later definition of sufficiency.

**Definition 2** (Interventions)**.** An intervention $\mathbf{I} \leftarrow \mathbf{i}$ is a partial setting $\mathbf{i}$ of variables $\mathbf{I}$. A proposition $\phi$ is true under an intervention, written $\mathbf{I} \leftarrow \mathbf{i}\varphi$, if $\varphi$ is true in the model identical to $\mathcal{M}$ except the causal mechanisms of $\mathbf{I}$ are set to be constant functions mapping to the values in $\mathbf{i}$.

Thus, a probabilistic SCM is a vanilla SCM with a probability distribution across exogenous variables ($\mathcal{P}$). In this treatment of introducing probability into a causal model, the uncertainty is "pulled out" (Halpern 2016) of the endogenous variables and inserted into exogenous variables, such that the result is a distribution over possible deterministic settings of the model.

2.1.1. A CAUSAL MODEL OF TIC-TAC-TOE. For our purposes, consider a basic probabilistic causal model that describes the machinations of tic-tac-toe between two agents, Player $X$ and Player $O$. We have $\mathcal{M}_{\text{ttt}} = (\mathcal{U}, \mathcal{V}, \mathcal{F}, \mathcal{P})$, where every setting of $\mathcal{U}$ delineates possible unfoldings of a tic-tac-toe game based on external factors (e.g., player knowledge) and $\mathcal{V} = \mathsf{Board}$, where $\mathsf{Board} = \{\mathbf{B}^l_t : 0 \leq t, l \leq 8\}$ (indicating time and location indices). Additionally, at each timestep $t$, the board-state at $t$ is specified by valuations of all nine assignments of $l$ at $t$. These valuations are done considering $\mathcal{F} : \forall z \in \mathsf{B}^l_t,\ z \in \{X, O, \text{EMPTY}\}$. $\mathcal{F}$ delineates the causal mechanisms of tic-tac-toe, namely that (1) $X$ always makes the first move, (2) $X$ and $O$ alternate turns, and (3) the game continues until either (a) a player is able to win by placing three in a row (including horizontally, vertically, and diagonally), or (b) $\forall z \in \mathsf{B},\ z \neq \text{EMPTY}$. Furthermore, $\mathcal{P}$ encodes information about $\rho$ – that is, how likely the players choose the highest-utility move. The highest-utility move given a board-state can be calculated using the minimax algorithm, as depicted in Figure 1. The minimax algorithm assumes two players – a player that attempts to maximize their possibility of winning, and a player that attempts to minimize the possibility of the former player winning. Each possible terminal board state is given a utility score, which we define to be $Winner \times (EmptySpace + 1)$, where $Winner$ is $-1$ if $O$ wins, $0$ if it is tie, and $+1$ if $X$ wins, while $EmptySpace$ is the number of empty spaces left on the board at the time of the terminal board state. The latter half of the expression ensures that games that are won earlier are favored. For the sake of assuming an imperfect and probabilistic player, if there is more than one possible transition from $t$ to $t+1$, the agent takes move $argmax(Utility_t)$ or $argmin(Utility_t)$ depending on if it is $X$ or $O$'s play with probability $\rho + \frac{1-\rho}{n}$ where $n$ is the number of empty spaces, and for all other moves take those with probability $\frac{1-\rho}{n}$. Altering the probabilities of possible choices to reflect a uniform distribution might reflect a less skilled player, or comparatively $\rho = 1.0$ might reflect the plays of a professional player.

2.2. ALTERNATIVES. Firstly, previous work (Frankfurt 1969, Pereboom 2000) argues that *the number of alternative actions available to the causee* can distinguish between causal relationships

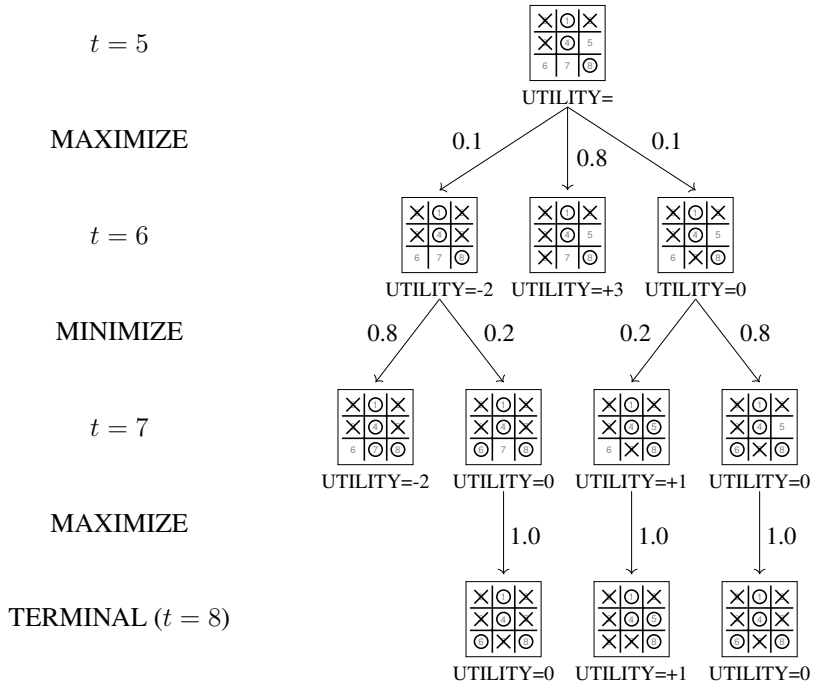


Figure 1: Iterating through a partial game using the minimax algorithm, where maximizer$= X$ and minimizer$= O$. Assuming $\rho = 0.8$, $P(\text{X wins}) = 0.82$ at $t = 5$.

in which the causer is (or is not) culpable for the action taken by the causee. This is also related to Lewis (1973)'s argument where actual causation is determined by looking at nearby possible worlds. Furthermore, and related to our upcoming discussion of intention, previous work (Halpern & Kleiman-Weiner 2018) has argued that an action taken by an agent when the agent could not do otherwise can never be intentional. This feature is also of interest for differentiating the semantics of causal verbs, since it provides the contrast in (8).

(8) a. The child was made to get into the car, although she could've chosen to do otherwise.
b. ?The child was forced to get into the car, although she could've chosen to do otherwise.

In tic-tac-toe, a higher number of empty squares signifies a higher degree of freedom and a lesser degree of influence of the causer, while a lower number of empty squares indicates fewer alternatives, suggesting a stronger influence. In this way, the number of potential moves in a tic-tac-toe board-state can be a rudimentary yet illustrative measure to concretize the continuum of causal influence denoted. So, our first measure ALT, which we expect to factor into the predictions of participant judgements of when *cause*, *make*, and *force* are acceptable, quantifies the number of alternative actions available to the causee.

Our upcoming experiment makes use of tic-tac-toe games as stimuli. In three-state tic-tac-toe sequences as in Fig. 2a, ALT is measured as the number of alternative actions the agent could have taken, excluding the action that was actually taken. So, $\text{ALT}(Y_1) = 5$.

2.3. INTENTION. Secondly, the concept of *intention* has been argued to be related to alternatives (Widerker & McKenna 2003) and also relevant for distinguishing causal situations (Copley 2018).

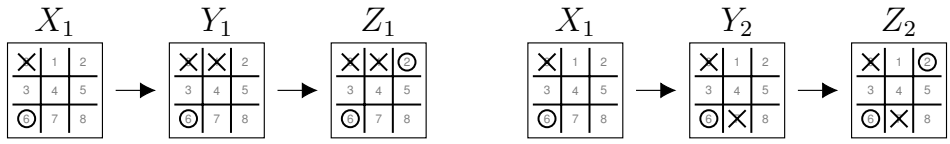


Figure 2: Examples of three-state tic-tac-toe sequences

Consider the following sentences:

(9) a. John caused the children to dance, but he didn't intend for the children to dance.
b. John made the children dance, but he didn't intend for the children to dance.
c. John forced the children to dance, but he didn't intend for the children to dance.

The intuition is that it is easiest to imagine a situation where (9-a) is true, then perhaps a bit more difficult to imagine a situation where (9-b) is true, and most difficult to imagine a situation where (9-c) is true. We can also see that these sentences are more acceptable with a hedge such as *accidentally* modifying the main verbs as in (10).

(10) a. John accidentally caused the children to dance. He didn't intend for the children to dance.
b. John accidentally made the children dance. He didn't intend for the children to dance.
c. John accidentally forced the children to dance. He didn't intend for the children to dance.

Building on this intuition, our second measure (INT) is a simplified version of the 'degree of intention' proposed by Halpern & Kleiman-Weiner (2018), which is defined within the framework of structural causal models. First, assume a causal model $\mathcal{M}$, a partial setting of the variables in $\mathcal{M}$, $\vec{u}$, an action $\vec{a}$, a goal $\vec{g}$, and a utility function $\mathbf{u}(w_{\mathcal{M}}, A \leftarrow \vec{a}, \vec{u})$. Let $\mathbf{u}'(w_{\mathcal{M}}, A \leftarrow \vec{a}, \vec{u}) = e^{\mathbf{u}(w_{\mathcal{M}}, A \leftarrow \vec{a}, \vec{u})}$, so that an agent's expected utility is strictly positive. Formally, our definition of INT is:

$$\mathsf{INT}(\mathcal{M}, \vec{a}, \vec{g}, \mathbf{u}') = \frac{\Pr((\mathcal{M}, \vec{u}) \models (A = \vec{a} \wedge G = \vec{g}))\mathbf{u}'(w_{\mathcal{M}}, A \leftarrow \vec{a}, \vec{u})}{\sum_{(\mathcal{M}, \vec{u}) \in \theta : (\mathcal{M}, \vec{u}) \models (A = \vec{a}' \wedge G = \vec{g})} \Pr(\mathcal{M}, \vec{u})\mathbf{u}'(w_{\mathcal{M}}, A \leftarrow \vec{a}', \vec{u})}$$

In prose, INT is the probability that an action performed in a state will result in the desired outcome, normalized by the probability of all alternative actions that would have resulted in the same outcome. First, we scale the utility values by the probability of the desired outcome relative to that action. Then, we normalize over all possible utility valuations (for actions for which the goal-state remains a possibility) to consider for comparative cases.

This definition captures our intuition that with respect to our tic-tac-toe examples in fig. 2a and fig. 2b, in the statement *Player O placed at location 2*, Player O is more intentional in taking this action in $Z_1$ than in $Z_2$. Specifically, any alternative to *Player O placed at location 2* in fig. 2a, e.g. *Player O placed at location 5*, would make it highly probable that *Player X* wins at the next time-step, thereby largely decreasing the probability of reaching the goal-state of *Player O*. The same is not true for fig. 2b. More generally, our definition takes into consideration that the degree

of intention is higher when the chosen action is more critical to achieving the desired outcome compared to alternatives. Since our goal is to use this measure as a predictor measured across tic-tac-toe sequences, where one "goal" (i.e., winning) is as morally good as another, we do not take into consideration side cases involving morality (Knobe 2003) that were central to Halpern & Kleiman-Weiner (2018)'s definition.

2.4. SUFFICIENCY. Thirdly, the notion of *causal sufficiency* has been well-represented in previous literature on causal verb selection – Glass (2023) argues that *cause* entails local sufficiency, while Nadathur & Lauer (2020) and Lauer & Nadathur (2018) argue that *make* conveys causal sufficiency. As suggested by force-theoretic work on causative verbs (Copley & Harley 2015, Talmy 1988, Wolff 2007), this distinguishes between causing and enabling verbs. Consider the following examples.

(11) a. The pirate made the prisoner walk down the plank.
b. The pirate let the prisoner walk down the plank.

We can say with certainty that *the prisoner walks down the plank* in (11-a), while it is less clear whether this result is guaranteed in (11-b). Thus, our third model (SUF) is Pearl (2019)'s "probability of sufficiency", which is defined as the probability that an event would be sufficient to produce an outcome. Descriptively, SUF denotes the capacity of $C$ to produce the outcome $E$ in situations where the agent of $C$ did some action other than the one encoded in $C$. Intuitively, *Player X placing at location 1* in $Y_1$ is more sufficient in bringing about *Player O placing at location 2*, than *Player X placing at location 7* in $Y_2$ is for bringing about the same. This is because in sequences where settings $Y_1$ and $Y_2$ don't result in *Player O placing at location 2* at the next time-step, it is more likely that $Y_1$ will *eventually* lead to *Player O placing at location 2* to block $X$'s clear three-in-a-row than $Y_2$, which does not present that danger to *Player O*.

Within the framework of SCMs, Pearl (2019, 2009) proposes the Probability of Sufficiency (SUF) mainly to explain why in the case when the presence of oxygen and a lit match are necessary for the occurrence of a fire, it is more felicitous to say that *The lit match caused the fire* than *The oxygen caused the fire*. As Pearl (2019) writes, the judgement is so because the presence of a lit match is more likely to be sufficient for the fire than the presence of oxygen. Assume that we have a causal model $\mathcal{M}$, a causer action $\vec{x} \notin \vec{u}$, where $\vec{u}$ is a partial setting of the variables in $\mathcal{M}$, and a causee action $\vec{y}$. Furthermore, $\vec{u}$ should have variable assignments for $X$ and $Y$ (where $X \neq \vec{x}$). Then, SUF is defined as:

$$\mathsf{SUF}(\mathcal{M}, \vec{x}, \vec{y}, \vec{u}) = \Pr(w_{\mathcal{M},\vec{u},Y=\vec{y}} \mid w_{\mathcal{M},\vec{u},X\leftarrow\vec{x}}),$$

which is how likely it is for $Y$ to become $Y = \vec{y}$ if $X$ were to counterfactually change from $X \neq \vec{x}$ to $X = \vec{x}$. Thus, SUF quantifies the ability of $X = \vec{x}$ to produce the outcome $Y = \vec{y}$ in situations where $Y \neq \vec{y}$. Performing this calculation requires a three step process: abduction (updating the prior probabilities in light of $\mathcal{M}$ and $\vec{u}$), action (intervening on $X$), and prediction in which we calculate the probability of $Y = \vec{y}$ given the updated variable values. Note that although the definition of SUF does not explicitly take into account the utility of an action, our implementation of probability across the tic-tac-toe game-tree does.

Going back to the examples from fig. 2, observe that "Player X *forced* Player O to place at location 2" is more felicitous in fig. 2a than fig. 2b. Recall that we predict a higher degree of

sufficiency to be somehow correlated with a higher degree of acceptability of *force*. Assume the notation that $\mathbf{B}_0 \ldots \mathbf{B}_8$ refers to each gridcell of a tic-tac-toe state and can take on values from $\{X, O, \textit{Empty}\}$. To support our prediction about the acceptability of *force*, we want to compare the abilities of (1) $\mathbf{B}_1 = X$ in producing $\mathbf{B}_2 = O$ and (2) $\mathbf{B}_7 = X$ in producing $\mathbf{B}_2 = O$ from fig. 2.

With regards to (1), recall that our context $\vec{u}$ cannot include $\mathbf{B}_1 = X$. So, assume the context is $Y_2$ from fig. 2b, which is $Y_2 = (\mathbf{B}_0 = X) \wedge (\mathbf{B}_6 = O) \wedge (\mathbf{B}_7 = X)$. However, after conditioning to have $\mathbf{B}_2 \leftarrow X$, we end up with variable settings equivalent to $Y_1$. Then, $\mathsf{SUF}(\mathcal{M}_{\text{ttt}}, \mathbf{B}_1 = X, \mathbf{B}_2 = O, Y_2) = 1.0$ in the case of $\rho = 1.0$ (where $\rho$ is the probability that the players choose the highest-utility move). It is clear that $\mathbf{B}_2 = O$ is the highest-utility move in this case because otherwise, Player $O$ would lose.

Next, we compare this with (2), the probability that $\mathbf{B}_7 = X$ produces $\mathbf{B}_2 = O$. We make the context of this $Y_1$. However, after conditioning to have $\mathbf{B}_7 = X$, we have $Y_2$. According to the minimax algorithm (see details in section 2.1), the next best move for $\textbf{Player}_O$ is now either location 2 or 8. Assuming the same $\rho$, the probability of sufficiency of $Y_2$ to produce $\mathbf{B}_2 = O$ is thus $0.5$. Since the probability of sufficiency for Player $X$'s move for producing $Z_2$ is greater than the probability of sufficiency for the same move to produce $Z_1$, this aligns with our intuition that the expression "Player X *forced* Player O to place at location 2" is more felicitous in fig. 2a than fig. 2b.

There is, however, an exception to this prediction. Assume instead that $\rho = 0.0$, meaning that the players choose their next move at random (e.g. assume that the players are infants). Then $\mathsf{SUF}(\mathcal{M}_{\text{ttt}}, \mathbf{B}_1 = X, \mathbf{B}_2 = O, Y_2) = \mathsf{SUF}(\mathcal{M}_{\text{ttt}}, \mathbf{B}_7 = X, \mathbf{B}_2 = O, Y_2)$ and the prediction is that neither expression should be preferred.

**3. Current experiment.** In this experiment, we test the three measures described above by creating a dataset of tic-tac-toe games encoding a range of ALT, INT, and SUF values, and evaluating their ability to predict the judgments. The data, analysis scripts, and experimental materials can be accessed here.

3.1. STIMULI. In order to generate the 30 stimuli, we first generated 21 full games of tic-tac-toe using a simulated player-and-opponent ran to fulfill a 5x5 design. The first dimension was how many turns it took for the game to complete, measured by the number of empty spaces $\delta$ remaining at the end of the game with $\delta \in \{1, 2, 3, 4, 5\}$. The second dimension measured how optimal $\rho$ each player was in selecting the highest-utility play as calculated via the minimax algorithm, with $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$. This results in 21 games (and not 25) because it is unlikely that a game continues until all squares of a tic-tac-toe board are filled unless $\rho = 1$.

Each possible 3-frame sequence from these 21 games were collected in order to create a set of 128 3-step frames. These 128 possible stimuli were then annotated with ALT, SUF, and INT values. Since SUF and INT values skewed towards the lower end of the scale, and since we are primarily interested in SUF and INT instances towards the higher end of the scale, we define a process to select stimuli where SUF and INT values differ the most, as well as cover the entire range of SUF and INT values. We first calculate the absolute difference between the SUF and INT values for each stimuli. Then, we define two sets of ten bins between 0 and 1 and assign each SUF and INT value into one of these bins. For each bin, we select 3 stimuli with the highest differences in SUF and INT values. This results in 30 stimuli. In order to artificially grow this set to 60, we

rotate each game board $90°$ clockwise in order to increase the diversity of the target set.

3.2. PARTICIPANTS. 109 native English speakers from the US and UK were recruited from Prolific. 19 participants were excluded from our analysis for failing at least one of two attentions check that asked whether a specific location was placed with a marker, given an image. Out of the remaining 90 participants, *age*: Mean = 36.1, SD = 14.53; *gender*: 50 female, 38 male, 2 non-binary. Each participant was provided with an introduction to the study and had to pass a simple comprehension question about tic-tac-toe to continue. Failing the comprehension check brought the participants back to the introductory instructions, after which they could re-attempt the comprehension question. Of those that passed the attention checks and comprehension questions, participants took on average 7 minutes and 18 seconds (SD = 3 minutes 42 seconds) to complete the task and were compensated 1.2 GBP.

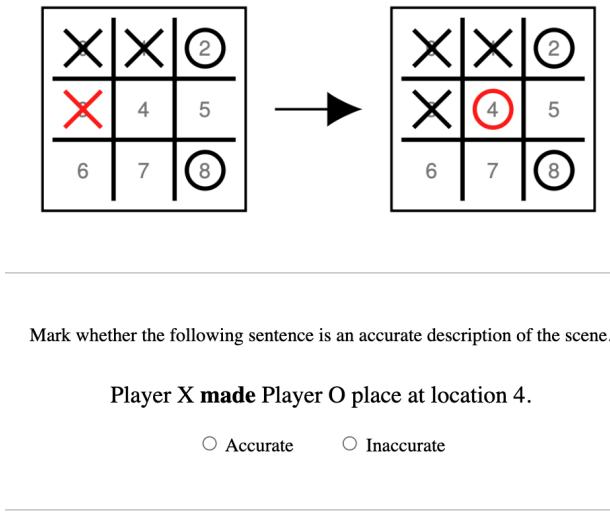


Figure 3: Example of experiment question

3.3. PROCEDURE. Participants were first shown a simple explanation of the rules for tic-tac-toe, and then presented with a comprehension question which asked the participant to select the "most likely" next move for a player that must choose a specific location in order to avoid losing. After answering this correctly, participants were presented with 20 pages that had one question each, where one page included both a tic-tac-toe stimulus and a sentence using one of the three causal verbs. The participants were asked to select whether the sentence using *cause*, *make*, or *force* was "Accurate" or "Inaccurate" in describing the stimulus. An example is shown in fig. 3.

Of the 20 questions, two were attention checks. The attention checks were designed to appear like the target questions, except participants were asked whether a player **placed** at a certain location.

**4. Results & Analysis.** Firstly, we find that holding the set of stimuli constant, participants were less likely to determine *made* than *caused* as accurate in describing a scenario, and less likely to determine *forced* than *made* as accurate (fig. 5). However, it was not the case that each stronger predicate's use was a subset of its weaker relatives, and so it is not clear whether these verbs lie

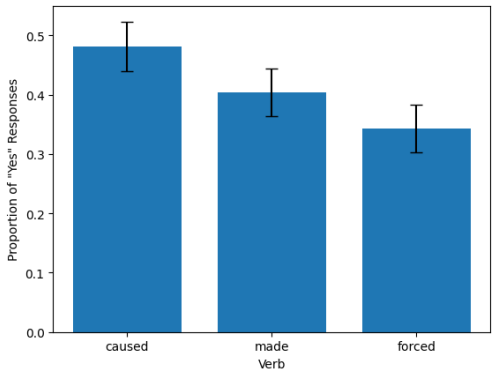


Figure 4: Proportion of "Yes" with 95% CIs

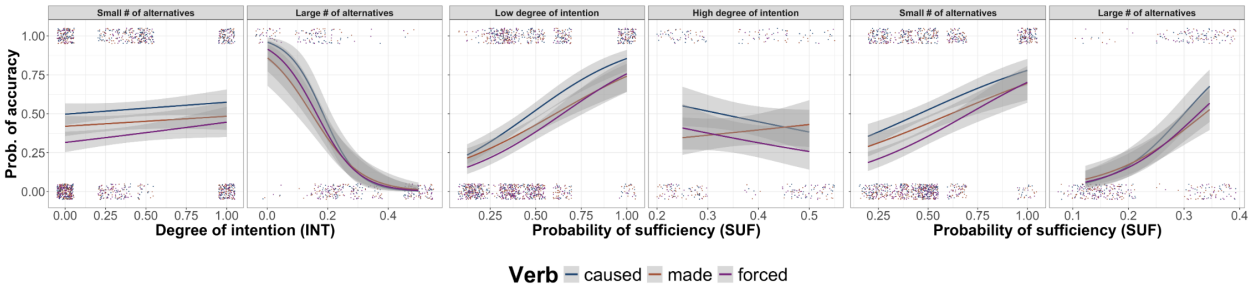


Figure 5: Probability of a "yes" rating as a function of each pair of ALT, SUF, and INT

in an asymmetric entailment relation (de Marneffe et al. 2010). Moving on, we first note that the anticorrelation between SUF and ALT is extremely strong ($-0.81$), which causes collinearity. This is somewhat expected, since the smaller number of alternatives that the causee has, the more sufficient the causer's action is for bringing about the result. To ensure stable coefficient estimates, we residualize SUF by ALT. This means that we take the vertical distance from the line to each of those points as using that distance as the predictor, rather than including the probability of sufficiency itself. The resulting predictor (SUFresidALT) is interpreted as capturing all the information that SUF provides that is not shared with ALT. We fit multiple Bayesian regressions with a Bernoulli family, using participant judgements as the outcome variable. In our first model (I; full model results in table 2), we include a full three-way interaction between the value of verb, INT, ALT, and SUFresidALT, as well as random effects for verb and participant. The random effects accounts for variability at the verb and participant levels. In our second model (II), we include all two-way interactions but exclude the three-way interaction between the continuous variables, but otherwise keep the random effects present in (I). In our third model (III), we do not include interactions between verb and the continuous variables, but keep the three-way interaction among the continuous variables (SUFresidALT, INT, SUF) as well as the categorical verb predictor and the random effects present in (I). Our last model (IV) includes only two-way interactions among the continuous variables and excludes the three-way interaction, but keeps the categorical verb predictor and the random effects present in (I).

We then compare these four models. We first find that the difference in ELPD (expected log

predictive density) between model III and model IV is $-20.9$, with a SE of $6.6$. Since III has a higher ELPD, it is the better model. The large negative difference suggests that dropping the three-way interaction among the continuous variables (as in IV) significantly reduces the model's predictive accuracy. This means that the interactions between `SUFresidALT`, `INT`, and `SUF` provides important information for optimizing predictions of when participants describe uses of *cause*, *make*, and *force* as accurate. Next, the difference in ELPD between I and II is $-21.8$, with a SE of $7.1$. I has a higher ELPD, indicating that it is the better model. Again, removing the three-way interaction among the continuous variables (in II) significantly reduces the model's predictive performance. Finally, we compare models I and III. The difference in ELPD between III and I is $-3.1$, with a SE of $4.8$. The difference is relatively small and the SE is larger than the absolute difference, meaning this difference is not significant. Therefore, both models perform similarly, with no clear advantage of including the interactions between `verb` and the continuous variables as in the full model (I).

Firstly, every standard deviation increase in `INT` causes the log-odds of a "yes" for *cause* to go up by $\sim 0.5$. This means that a stimuli with a higher degree of `INT` is more likely to have *cause*, *make*, and *force* rated as accurate. Also, every standard deviation increase in residualized `SUF` predictor causes the log-odds of a "yes" for *cause* to go up by $\sim 1.2$, which is more than twice the size of the effect of `INT`. So, a higher degree of residualized `SUF` increases the probability of accuracy much more than increasing `INT` does. Next, every standard deviation increase in `ALT` causes the log-odds of a "yes" for *cause* to go down by $\sim 0.8$. It is expected that this effect is in the opposite direction than the other two predictors, since `ALT` is anticorrelated with `SUF`. We also note that when `ALT` matches in sign with residualized `SUF` or `INT`, the log-odds of a "yes" for *cause* goes up $\sim 0.5$ in the produce of standard deviations.

Interestingly, there is a comparatively large and reliable interaction between the *made* level of `verb` with residualized `SUF` and `INT`, unlike the other levels of `verb`. The same observation holds for the verb *cause*, and the interaction between `SUFresidALT` and `ALT`, as well as `INT` and `ALT` (see the highlighted rows in table 2). These interactions suggest that *made* has some additional semantic component that is also a function of `SUF` and `INT`, that is not present in the other verbs. The same possibility holds for *cause* and its notable interactions. In order to explore this in more detail, we fit three additional models for each verb. After subsetting the data by `verb`, model V predicts participants' judgments of stimuli that use the verb *caused* by including three continuous predictors (as fixed effects), along with their interactions. The model also includes random intercepts for `participant` to account for variability in individual responses. Model VI and VII are similar, except VI only pertains to stimuli that use *made*, and VII only pertains to stimuli that use *forced* (see all complete model results in tabs. 3–5).

As we expect, all three models reliably use our three continuous measures as predictors of participant judgements of each verb. More interestingly, each verb has a unique combination of interactions that are reliable (see tab. 1). In model V, the interactions of continuous predictors that are reliable for *cause* are firstly, the interaction between `SUFresidALT` and `ALT`; secondly, the interaction between `INT` and `ALT`; and thirdly, the three-way interaction between `SUFresidALT`, `INT`, and `ALT`. As for *made* in model VI, surprisingly, the interactions that are reliable are not reflected in the original full model (I). The reliable predictors for judgements of *made* include the

| | *caused* | *made* | *forced* |
|---|---|---|---|
| SUFresidALT:INT | - | + | - |
| SUFresidALT:ALT | + | + | + |
| INT:ALT | + | + | + |
| SUFresidALT:INT:ALT | - | - | - |

Table 1: Estimates of interactions' intercepts by `verb`. Light grey indicates an unreliable effect.

interaction between `SUFresidALT` and `ALT`, as well as `SUFresidALT`, `INT`, and `ALT`. Finally, in model VII, the interaction between `SUFresidALT`, `INT`, AND `ALT` are reliable.

**5. Discussion.** To begin, our results support the prediction that intention, sufficiency, and possible alternative actions factor into the semantics of the causal senses of *cause* and *force*, which is demonstrated by the reliable intercepts for these level of `verb` in model I. Furthermore, each continuous measure shows as reliable in this model. It is less clear whether this holds for *made* since its credible interval includes 0 in the full model (I). However, we observe that the smaller *made*-specific model (VI) reliably uses all three of `SUFresidALT`, `INT`, and `ALT` as predictors. The uncertainty in model I may originate from a variety of factors – for example, even interactions that are reliable for *made*, such as its interaction with `SUFresidALT` and `INT`, have wide credible intervals ($0.02$ to $0.87$), which may contribute to higher overall uncertainty. Although the results for *made* are unclear, it seems unlikely that the concepts do not at all contribute to the semantics of *made*. Regarding alternatives, e.g., consider one participant of our task's post-survey comment:

> "If there were multiple options, [i.e.] more than one blank spot where the player could select, [...] I made an assumption that "forced" or "made" were inaccurate."

Evidently, participants take into consideration the number of alternatives that a causee has when judging the accuracy of *make*.

Taking a wider view, that each verb takes a unique combination of interactions supports the argument that these components convey distinct information to our models of participant judgements. Furthermore, the semantics of each verb takes into consideration a unique blend of each concept. It is also interesting to observe that *cause*, *make*, and *force* have a decreasing number of reliable interactions that are used as predictors. Speculatively, these results seem to convey that what we perceive as increasing "force" in these verbs is actually a conglomerate of multiple factors.

To conclude, our work has provided experimental support for our hypothesis that causal verbs such as *cause*, *make*, and *force* exhibit distinct and nuanced semantics, which are shaped by a combination of factors including sufficiency, intention, and alternatives. Through our experimental analysis, we demonstrated that no single predictor, such as sufficiency alone, fully determines the appropriateness of these verbs. Instead, the interactions between these factors play a unique role in determining how participants judge the accuracy of each of these verbs in describing various causal scenarios.

## References


Baglini, Rebekah & Elitzur A Bar-Asher Siegal. 2021. Modelling linguistic causation. *manuscript, Aarhus University and Hebrew University of Jerusalem* .

Cao, Angela, Atticus Geiger, Elisa Kreiss, Thomas Icard & Tobias Gerstenberg. 2023. A semantics for causing, enabling, and preventing verbs using structural causal models. In *Proceedings of the 45th annual conference of the cognitive science society*, 2947–2954.

Cao, Angela, Gregor Williamson & Jinho Choi. 2022. A cognitive approach to annotating causal constructions in a cross-genre corpus. In *Proceedings of the 16th linguistic annotation workshop (law) at lrec*, 151–159. Online: European Language Resources Association. http://lrec-conf.org/proceedings/lrec2022/workshops/LAWXVI/pdf/2022.lawxvi-1.18.pdf.

Childers, Zachary. 2016. *Cause and affect: Evaluative and emotive parameters of meaning among the periphrastic causative verbs in english*. Austin, TX: University of Texas, Austin Phd dissertation.

Copley, Bridget. 2018. Dispositional causation. *Glossa: a journal of general linguistics* 3. 10.5334/gjgl.507.

Copley, Bridget & Heidi Harley. 2015. A force-theoretic framework for event structure. *Linguistics and Philosophy* 38(2). 103–158. 10.1007/s10988-015-9168-x.

Davies, Mark. 2008–. The corpus of contemporary american english (coca). Available online at https://www.english-corpora.org/coca/.

Frankfurt, Harry G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66(23). 829–839. http://www.jstor.org/stable/2023833.

Glass, Lelia. 2023. Using the Anna Karenina Principle to explain why *cause* favors negative-sentiment complements. *Semantics and Pragmatics* 16(6). 1–48. 10.3765/sp.16.6.

Halpern, Joseph Y. 2016. *Actual Causality*. The MIT Press. https://doi.org/10.7551/mitpress/10809.001.0001.

Halpern, Joseph Y. & Max Kleiman-Weiner. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility.

Hammond, Lewis, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate & Michael Wooldridge. 2023. Reasoning about causality in games. *Artificial Intelligence* 320. 103919.

Klettke, Bianca & Philip Wolff. 2003. Differences in how english and german speakers talk and reason about cause. In *Proceedings of the annual meeting of the cognitive science society*, vol. 25, 675–680.

Knobe, Joshua. 2003. Intentional action and side effects in ordinary language. *Analysis* 63. 190–193.

Lauer, Sven & Prerna Nadathur. 2018. Sufficiency causatives. Unpublished manuscript.

Lewis, David. 1973. Causation. *Journal of Philosophy* 70(17). 556–567. 10.2307/2025310.

de Marneffe, Marie-Catherine, Christopher D. Manning & Christopher Potts. 2010. "was it good? it was provocative." learning the meaning of scalar adjectives. In Jan Hajič, Sandra Carberry, Stephen Clark & Joakim Nivre (eds.), *Proceedings of the 48th annual meeting of the association for computational linguistics*, 167–176. Uppsala, Sweden: Association for Computational Linguistics. https://aclanthology.org/P10-1018.

Nadathur, Prerna & Sven Lauer. 2020. Causal necessity, causal sufficiency, and the implications

of causative verbs. *Glossa: a journal of general linguistics* 5. 49–105.

Nadathur, Prerna & Elitzur A Bar-Asher Siegal. 2022. Modeling progress: causal models, event types, and the imperfective paradox. In *West coast conference in formal linguistics (wccfl)*, vol. 40, .

Pearl, Judea. 2009. *Causality: Models, reasoning and inference*. Cambridge University Press 2nd edn.

Pearl, Judea. 2019. Sufficient causes: On oxygen, matches, and fires. *Journal of Causal Inference* 7(2). 20190026. https://doi.org/10.1515/jci-2019-0026.

Pereboom, Derk. 2000. Alternative possibilities and causal histories. *Philosophical Perspectives* 14. 119–137. http://www.jstor.org/stable/2676125.

Schulz, Katrin. 2011. If you'd wiggled a, then b would've changed: Causality and counterfactual conditionals. *Synthese* 179(2). 239–251. 10.1007/s11229-010-9780-9.

Shibatani, M. 1976. *The grammar of causative constructions* Syntax and Semantics Online, ISBN: 9789004425774. BRILL. https://books.google.com/books?id=ft2KzgEACAAJ.

Talmy, Leonard. 1988. Force dynamics in language and cognition. *Cognitive Science* 12(1). 49–100. https://doi.org/10.1016/0364-0213(88)90008-0. https://www.sciencedirect.com/science/article/pii/0364021388900080.

Widerker, David & Michael McKenna (eds.). 2003. *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Ashgate.

Williamson, Gregor, Angela Cao, Yingying Chen, Yuxin Ji, Liyan Xu & Jinho D. Choi. 2023. Exploring a multi-layered cross-genre corpus of document-level semantic relations. *Information* 14(8). 431. 10.3390/info14080431.

Wolff, Phillip. 2007. Representing causation. *Journal of experimental psychology. General* 136. 82–111. 10.1037/0096-3445.136.1.82.

Wolff, Phillip, Bianca Klettke, Tracy Ventura & Geunyoung Song. 2005. Expressing causation in english and other languages. In Woo-kyoung Ahn, Robert L. Goldstone, Bradley C. Love, Arthur B. Markman & Phillip Wolff (eds.), *Categorization inside and outside the laboratory: Essays in honor of douglas l. medin*, 29–48. American Psychological Association. 10.1037/11156-003.

## Appendix

| Parameter | Estimate | Est. Error | l-95% CI | u-95% CI |
|---|---|---|---|---|
| Intercept | -0.43 | 0.15 | -0.74 | -0.14 |
| verbforced | -0.67 | 0.21 | -1.08 | -0.26 |
| verbmade | -0.25 | 0.19 | -0.61 | 0.11 |
| SUFresidALT | 1.19 | 0.16 | 0.89 | 1.50 |
| INT | 0.54 | 0.13 | 0.28 | 0.81 |
| ALT | -0.82 | 0.14 | -1.11 | -0.55 |
| SUFresidALT:INT | -0.30 | 0.16 | -0.61 | 0.02 |
| SUFresidALT:ALT | 0.44 | 0.16 | 0.13 | 0.76 |
| INT:ALT | 0.50 | 0.18 | 0.16 | 0.86 |
| verbforced:SUFresidALT | 0.09 | 0.23 | -0.36 | 0.55 |
| verbmade:SUFresidALT | -0.32 | 0.22 | -0.74 | 0.10 |
| verbforced:INT | 0.11 | 0.20 | -0.28 | 0.51 |
| verbmade:INT | -0.11 | 0.19 | -0.47 | 0.25 |
| verbforced:ALT | 0.10 | 0.21 | -0.34 | 0.52 |
| verbmade:ALT | 0.05 | 0.19 | -0.33 | 0.44 |
| SUFresidALT:INT:ALT | -0.72 | 0.19 | -1.07 | -0.35 |
| verbforced:SUFresidALT:INT | 0.32 | 0.23 | -0.15 | 0.77 |
| verbmade:SUFresidALT:INT | 0.45 | 0.22 | 0.02 | 0.87 |
| verbforced:SUFresidALT:ALT | -0.34 | 0.24 | -0.80 | 0.13 |
| verbmade:SUFresidALT:ALT | -0.03 | 0.22 | -0.46 | 0.40 |
| verbforced:INT:ALT | -0.40 | 0.30 | -1.00 | 0.19 |
| verbmade:INT:ALT | -0.39 | 0.25 | -0.87 | 0.11 |
| verbforced:SUFresidALT:INT:ALT | -0.42 | 0.29 | -0.99 | 0.15 |
| verbmade:SUFresidALT:INT:ALT | 0.26 | 0.26 | -0.24 | 0.76 |

Table 2: Parameter estimates for model I.

| Parameter | Estimate | Est. Error | l-95% CI | u-95% CI |
|---|---|---|---|---|
| Intercept | -0.46 | 0.15 | -0.76 | -0.17 |
| SUFresidALT | 1.21 | 0.17 | 0.90 | 1.54 |
| INT | 0.54 | 0.14 | 0.27 | 0.82 |
| ALT | -0.83 | 0.14 | -1.13 | -0.56 |
| SUFresidALT:INT | -0.28 | 0.16 | -0.60 | 0.04 |
| SUFresidALT:ALT | 0.46 | 0.17 | 0.14 | 0.79 |
| INT:ALT | 0.52 | 0.18 | 0.16 | 0.89 |
| SUFresidALT:INT:ALT | -0.71 | 0.19 | -1.07 | -0.35 |

Table 3: Parameter estimates for model V.

| Parameter | Estimate | Est. Error | l-95% CI | u-95% CI |
|---|---|---|---|---|
| Intercept | -0.65 | 0.14 | -0.93 | -0.39 |
| SUFresidALT | 0.87 | 0.15 | 0.58 | 1.17 |
| INT | 0.44 | 0.13 | 0.19 | 0.69 |
| ALT | -0.74 | 0.13 | -1.00 | -0.50 |
| SUFresidALT:INT | 0.16 | 0.15 | -0.13 | 0.45 |
| SUFresidALT:ALT | 0.41 | 0.15 | 0.12 | 0.70 |
| INT:ALT | 0.09 | 0.17 | -0.23 | 0.43 |
| SUFresidALT:INT:ALT | -0.44 | 0.17 | -0.78 | -0.10 |

Table 4: Parameter estimates for model VI.

| Parameter | Estimate | Est. Error | l-95% CI | u-95% CI |
|---|---|---|---|---|
| Intercept | -1.13 | 0.17 | -1.48 | -0.80 |
| SUFresidALT | 1.26 | 0.18 | 0.92 | 1.61 |
| INT | 0.63 | 0.15 | 0.35 | 0.92 |
| ALT | -0.70 | 0.17 | -1.04 | -0.38 |
| SUFresidALT:INT | -0.01 | 0.17 | -0.34 | 0.32 |
| SUFresidALT:ALT | 0.08 | 0.18 | -0.27 | 0.42 |
| INT:ALT | 0.12 | 0.24 | -0.35 | 0.59 |
| SUFresidALT:INT:ALT | -1.15 | 0.23 | -1.61 | -0.71 |

Table 5: Parameter estimates for model VII.