Supplementary Materials to "Talking About What Would Happen Versus What Happened: Tracking Congressional Speeches during COVID-19"

Section A. Counterfactual Detection Rules

Section A elaborates on the procedure of counterfactual detection through two tables and one figure. Parsing and tagging results (see Table S1), the visualized dependency tree (see Figure S1), and the extended set of counterfactual forms and syntactic rules (see Table S2) are provided with supplementary text.

Our counterfactual detection rules utilize three types of lexical and syntactic information, based on version 1.2.1 of the R "spacyr" package (Benoit & Matsuo, 2020).[1] After preprocessing (i.e., removing redundant white spaces, special symbols, and stopwords), each sentence in the Congressional speech corpus was split into individual words or punctuation, a process known as "tokenization." For each token, we identified its "lemma," "POS (parts-of-speech) tag," and "dependency relation." First, "lemma" is a base token form, and "lemmatization" includes converting plural nouns to singular nouns (e.g., "districts" to "district" in Table S1), finding the root verb of a past-tense verb (e.g., transforming "was" or "were" into "be"), and others.

Second, "POS tagging" is the process in which grammatical information, such as verbs or nouns, are tagged into tokens. POS tags can be either Universal or Detailed. To

[1] The "spacyr" package is the R wrapper to the Python "spaCy" package. For details, please visit the "spacyr" package webpage: https://cran.r-project.org/web/packages/spacyr/index.html

illustrate, “was” is simply a verb (i.e., “VERB”) in the Universal tag set,[2] but it can also be tagged as Detailed as “VBD” or past-tense verb. Our study utilized the detailed POS tags of the Penn Treebank, as provided in the R “spacyr” package.

Third, the relations between tokens are obtainable through “dependency parsing.” Each token serves a role in each sentence, for example, “to modify” other tokens or “to be modified” by them. In other words, a sentence has a hierarchical structure in which tokens are interdependent to other tokens. Thus, we represented each sentence as a CoNLL-style dependency tree since its dependency structure can reflect the unidirectional relationship between the antecedent and consequent in a counterfactual expression. Please refer to the results of the dependency parsing (see the “Relation” column in Table S2) and the visualized dependency tree (see Figure S1). The example sentence in Table 1 (“If we do not move this bill, the death will be in our districts.”) is tokenized into 16 tokens, and each token is labelled with a lemma, a POS tag, and a dependency relation.

Our counterfactual detection rules are primarily based on the single root verb and its leaf verbs.[3] In Table S1, the root verb is “Be,” and “Move” is the only verb depending on the root verb. Therefore, we suspect the “Be” and “Move” clauses to be the consequent and the antecedent, respectively. To identify whether the consequent or the antecedent is in past or present tense, or whether it is a modal verb (e.g., should, could), we additionally extract auxiliary particles (i.e., “aux” or “neg” in the “Relation” column), depending on either of the two verbs. Please note that we are focusing on the verbs only and not other POS (e.g., root and leaf nouns) as we need to identify a hypothetical situation in counterfactuals.

---

[2] https://universaldependencies.org/u/pos/all.html

[3] As an additional note, we also used the root verb to identify time-focusing of the sentence. For example, the consequent in Table S1 is “will be,” so the sentence will be labelled as “future-focusing.” Likewise, if the root verb is past tense (i.e., “VBD”) or base form (i.e., “VB”) without “will,” the time-focusing would be “past-” or “present-focusing” in each case.

Table S1

An Example of Tokenization, Lemmatization, POS Tagging, and Dependency Parsing

ID (#)	Token	Lemma	POS Tag[a]	Dependency	
				Head Token ID (#)	Relation[b]
1	If	If	IN	5	mark
2	We	-PRON-	PRP	5	nsubj
3	Do	Do	VBP	5	aux
4	Not	Not	RB	5	neg
5	Move	Move	VB	12	advcl
6	This	This	DT	7	det
7	Bill	Bill	NN	5	dobj
8	,	,	,	12	punct
9	The	The	DT	10	det
10	Death	Death	NN	12	nsubj
11	Will	Will	MD	12	aux
12	**Be**	**Be**	**VB**	**12**	**ROOT**
13	In	In	IN	12	prep
14	Our	-PRON-	PRP$	15	poss
15	districts	District	NNS	13	pobj
16	.	.	.	12	punct

Note: Using the spacy_parse() function in the R "spacyr" package (version 1.2.1), the annotation results of the first sentence in Table 1 ("If we do not move this bill, the death will be in our districts.") are displayed. For dependency parsing, "Relation" denotes the relationship between the token and its head token, matched with the "Head Token ID (#)."

Based on the dependencies, two hierarchical relations are shaded: (1) **"Be (#12, ROOT)"** and its dependencies in grey color, (2) "Move (#5, advcl)" and its dependencies in orange color. Note that "Move" is the only verb depending on the root verb, and all the shaded tokens are either verbs (i.e., "VB" in the "POS Tag" column) or auxiliary particles (i.e., "aux" or "neg" in the "Relation" column). Using these verbal hierarchies, the antecedent and consequent candidates are underlined. See Figure S1 for the visualized results of Table S1.

[a]POS (part-of-speech) tags from the Penn Treebank tag set. For shaded areas, VBP = Verb, 3rd person singular present; RB = Adverb; VB = Verb, base form; MD = Modal.

[b]Universal dependency relations in the Stanford Dependencies. For shaded areas, aux = auxiliary; neg = negative; advcl = adverbial clause modifier; ROOT = root. More details are available in the R "spacyr" package manual.

Table S2

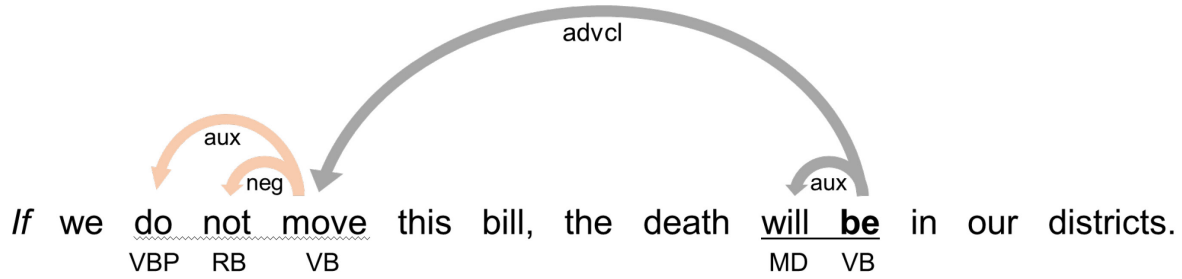
Three Forms of Counterfactual Expression and Syntactic Rules

Form	Example	Keyword	Antecedent	Consequent
Conditional Conjunction	*If* we do not move this bill, the death will **be** in our districts. (Mr. GARAMENDI) *Unless* we act soon, we will **see** mass layoffs, devastating tax increases, and a breakdown in public safety and essential services. (Mr. MENENDEZ)	"If" or "Unless"	Base verb, past-tense verb, or modal verb	Modal verb
Verb Inversion	*Had* he (= Mr. HOYER) been serious about that, there would have **been** a discussion before this bill ever came to the floor. (Mr. PALMER) *Were* I (= Judge Ginsburg) to rehearse here what I would say and how I would reason on such questions, I would **act** injudiciously. (quotes Mr. LANKFORD)	"Had" or "Were"	Past-tense verb	Modal verb
Wish/Should	I ***wish*** this bill was not necessary, but unfortunately, it is now more imperative than ever. (Mr. LOFGREN) We shouldn't be asking that question. That *should* have **been** a question for the Afghans. (Mr. GOHMERT)	"wish" or "should"	Past-tense (modal) verb	N/A

Note: Extending Table 1 in the main text, the **root** verb, consequent, antecedent and *keyword* are highlighted in each of the six sentences. Note that the "Verb Inversion" form allows only upper cases of "Had" or "Were" at the beginning of a sentence, whereas the "Conditional Conjunction" form includes both upper and lower cases of "If" or "Unless" (c.f., excluding "as if"). For the "Wish/Should" form, the consequent is omitted yet implied as a better situation. Especially for the "Should" verb, our syntactic rules necessitate the presence of a VBN (Verb, past particle) in the antecedent (e.g., should have "been").

Figure S1

An Example of CoNLL-style Dependency Tree



Note: Based on the annotation results in Table S1, the dependency structure of the antecedent and consequent is presented. Colors of arrow are matched to either of the two hierarchies (i.e., "be" in grey; "move" in orange), along with "Relation" tags underneath the arrows and POS tags below the line.

Using the annotation results, we can extract counterfactual expressions by matching the results to the predefined set of three counterfactual forms. In main text, we have introduced the counterfactual detection rules using the three example sentences in Table 1. In Table S2, the extended set of six counterfactual sentences is presented. As mentioned in main text, the syntactic rules for each form are as follows (please refer to relevant examples in parentheses).

Regarding the "Conditional Conjunction" form, the consequent consists of a modal verb (e.g., "will be" or "will see") while the antecedent consists of a base verb (e.g., "move" or "act"), a past-tense verb, or a modal verb, preceded by a conditional conjunction, such as "If" or "Unless." For the "Verb Inversion" form, the sentence begins with a verb inversion, like "Had he …" or "Were I …"), followed by a past-tense antecedent and a modal consequent. Accordingly, the "Verb Inversion" form allows only upper cases of "Had" or "Were" at the beginning of a sentence, whereas the "Conditional Conjunction" form includes both upper and lower cases of "If" or "Unless" (c.f., excluding "as if"). In the "Wish/Should" form, the clause in the past tense, following "wish" (a past-tense verb; e.g., "was") or "should" (a past-tense modal verb; e.g., "should have been"), is the antecedent, and the consequent is

omitted yet implied as a better situation. Especially for the “Should” verb, our syntactic rules necessitate the presence of a VBN (Verb, past particle) in the antecedent (e.g., should have “been”).

Section B. MZIP and MP Model Description

Section B provides a full description of the multi-level zero-Inflated Poisson (MZIP) and the multi-level Poisson (MP) models adopted in the study. In the following paragraphs, we demonstrated the rationale for using an MP approach with an offset variable. The need and importance of a zero-inflated model were pointed out, especially for counterfactual expressions, but not for time-focusing.

First, it is reasonable to expect that our dependent variables—the number of counterfactual expressions or verbs whose tenses are past, present, or future—are more likely to appear in longer speeches consisting of many sentences. Regarding the first issue, we treated the total number of sentences in a speech as an offset (Long, 1996), modeling the incidence ratio of an individual member using counterfactual expressions or specific time-orientation in a speech.

Second, Congressional speech data are nested with multiple hierarchies. Individual speeches may be correlated if they were made by the same member of Congress, or by those from the common state. To statistically address those cluster effects across members and states, random effects are modeled. In brief, we fitted an MZIP model to estimate the number of counterfactual expressions and time-focusing in Congressional speech.

Finally, as described above, counterfactual expressions are not frequently used in Congressional speech, meaning that there are excessive zeros in the number of counterfactual expressions (i.e., about 1% of total sentences were counterfactual sentences). In other words, conventional regression models for count variables, such as simple Poisson or negative binomial regressions, fail to predict the occurrence of counterfactual expressions adequately. To address this, the zero-inflated Poisson (ZIP) regression model (Cameron &

Trivedi, 2013) was adopted. Statistically, the ZIP regression assumes a mixed distribution of zero and positive counts, meaning that it simultaneously estimates whether an event (i.e., counterfactual expression) occurs, as well as the number of times it occurs (i.e., frequency of counterfactual expressions). The ZIP regression, comprising the logistic and Poisson model, predicts zero and non-zero count responses, respectively. Unlike counterfactual expressions, time-focusing variables do not suffer from excessive zeros, so we did not implement zero-inflated models.

The model formula of the MZIP model is as follows: Under the standard ZIP model, the i^{th} observation Y_i follows the mixture distribution composed of two parts, where ϕ_i is the proportion of zero responses from a logistic model, and λ_i is the rate parameter of the Poisson model.

$$P(Y_i = y) = \begin{cases} \phi_i + (1-\phi_i)e^{-\lambda_i}, & y = 0 \\ (1-\phi_i)\dfrac{\lambda_i^y e^{-\lambda_i}}{y!}, & y \geq 1 \end{cases}$$

Let Y_{ij} denote the i^{th} observation of the j^{th} subject. Then, we can consider the subject random effect (u_j) as a part of the Poisson model. Using the log link function, the rate parameter λ_i is modeled as shown. This can be viewed as an extension of the generalized linear mixed model (GLMM) with a mixture of distributions.

$$log(\lambda_{ij}) = {x_{ij}}^T \beta + u_j$$

For time-focusing, by contrast, the dependent measures are count variables clustered across members and states, but presumably not zero-inflated. Over 98% of the total sentences were successfully classified as past (18.3%), present (68.2%), or future (11.6%) focused. Thus, when modeling time-focusing, we adopted MP regression models without implementing a zero-inflated model.

References

Benoit, K., & Matsuo, A. (2020). spacyr (*R package version 1.2.1)* [Computer software]. Retrieved from https://spacyr.quanteda.io

Cameron, A. C. & Trivedi, P. K. (2013). *Regression analysis of count data*. Cambridge, United Kingdom: Cambridge University Press.

Long, J. S. (1996). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA, U.S.A.: Sage Publications.