Using Graded Factual Difference-Making to query people's internal causal variables through causal judgments

Anonymous

January 2026

1 Introduction

The problem of determining what to formulate as a causal variable has been called "the problem of variable choice" (Goddu & Gopnik, 2024; Woodward, 2016) or "learning causal variables" (Schölkopf et al., 2021). A potential solution is to derive predictions from normative accounts of causation and compare them against people's causal judgments, which some have distinguished into two kinds (Griffiths & Tenenbaum, 2005; Quillien & Lucas, 2024): The first kind, *categorical causal judgments*, involves separating causes from non-causes. The second kind, *causal selection* (or *strength*) *judgments*, involves ordering the different causes in the order from most causal to least causal.

Andreas and Günther (2025) present Factual Difference-Making (FDM) as a normative account of actual causation. In contrast to the existing accounts (Halpern & Pearl, 2005), FDM relies on the syntax of the structural equations. This enables it to distinguish between logically equivalent but syntactically different structural equations. However, FDM is an account of categorical causal judgments. Kominsky and Phillips (2025) present an extension to FDM to account for causal selection judgments. (We call this extension Graded FDM.) Similar to the Counterfactual Effect Size model (Quillien & Lucas, 2024) and the Necessity and Sufficiency model (Icard et al., 2017), Graded FDM can grade different causes as being more or less causal than others. CESM and NSM make no predictions about categorical causal judgments, while Graded FDM aims to be an account of both categorical causal judgments as well as causal selection judgments. The current work computationally implements Graded FDM and, through two preregistered experiments, tests whether Graded FDM is not only normative but also descriptive of laymen causal judgments. In doing this, we check whether the causal models inside people's minds contain certain causal variables and not others.

2 Predictions

The particular situation we consider corresponds to Experiment 3 of Quillien and Lucas (2024). This involved participants playing several rounds of a game. In each round, participants drew balls from a number of boxes by pressing a button. If the participants succeeded in obtaining at least one purple ball and one orange ball, they won that round (Figure 1). This corresponds to the logical structure given by the structural equation $Win = (A \vee B) \wedge C$, in which A, B, C corresponds to obtaining a colored ball from the respective boxes.

$Win = (A \vee B) \wedge C$ can be rewritten according to either of the two following sets of logically equivalent but syntactically different structural equations:

$$D = A \vee B$$
$$Win = D \wedge C$$

and

$$D = A \wedge C$$
$$E = B \wedge C$$
$$Win = D \vee E$$

CESM as well as NSM do not distinguish between the two sets of structural equations. However, Graded FDM predicts that, for the first set of equations, participants' judgments should be consistent with experiment 3 of Quillien and Lucas (2024), that is, $A < B < C$. In contrast, for the second set of equations, participants' causal selection judgments should be $A < B > C$. We test this prediction in experiment 1.

3 Experiment 1: Do participants' causal judgments differ for syntactically different but logically equivalent structural equations?

Methods

Our experiment 1 closely matches experiment 3 of Quillien and Lucas (2024). Our main manipulation comprised varying the colors of colored balls in the boxes across two conditions of the within subjects factor *ABcolor*:

- Identical: The colors of the colored balls in boxes A and B were the same, in particular, purple. The color of the colored balls in box C was orange. We expected this to induce causal judgments based on the first set of structural equations above with $D = A \vee B$ being an internal variable.
- Distinct: The colors of the colored balls in boxes was brown, blue and pink in A, B and C respectively. We expected this to induce causal judgments based on the second set of structural equations above with $D = A \wedge C$ and $E = B \wedge C$ as internal variables.

Results

$N = 200$ participants were recruited from Prolific. We noted Bayes factor in support for *ABcolor* as a predictor to be 211. However, the pattern of results did not match our predictions. 95% HDI for the difference in causal judgments for each cause overlapped across the two conditions of *ABcolor*[1] (Figure 2, left).

[1] We also manipulated probability of C to test other expectations based on internal variables, but we skip discussing it for brevity.

4 Experiment 2: Do the results of Experiment 1 generalize to (i) a different response method and (ii) more intuitive stimuli?

Methods

Experiment 2 extended 1 in two ways:

1. We obtained causal selection judgments separate from categorical causal judgments.
2. We used vignettes describing relatable events, based on previous findings (Fiddick et al., 2000; Romoli et al., 2022; Sperber et al., 1995) that participants' intuitions often improve with increased intuitiveness of the task scenario.

Syntactic manipulation was made linguistically, and *ABcolor* was now a between subjects factor. First, participants selected the causes of an event described in the vignette amongst one or more of the five candidate causes they are presented with. Subsequently, they arranged the causes they selected from most causal to least causal. They could rank two causes as being equally causal. The normality of events was manipulated through the descriptions in the vignette and participants also provided normality rankings after the causal rankings.

Results

We recruited N=80 participants from Prolific. Bayes factor in favor of the model incorporating *ABcolor* was noted to be 0.16 (Figure 2, right). Restricting participants to the subset who provided normality rankings according to our expectations, the Bayes factor in favor of the model incorporating *ABcolor* comes out as 5.36. However, the 95% HDIs for the causes differ from the expectations obtained by Graded FDM.

5 Discussion

Taken together, our results suggest models of actual causation (Icard et al., 2017; Quillien & Lucas, 2024) should go beyond variable valuations and take syntax into account which may create hidden variables that influence causal judgments. Even though we find an effect of syntax, our predictions differ from those made by Graded FDM (Kominsky & Phillips, 2025). Another model that is sensitive to syntax may explain these results. However, our results do not make any claims about the psychological plausibility of FDM proposed by Andreas and Günther (2025) in general.

References

Andreas, H., & Günther, M. (2025). Factual difference-making. *Australasian Philosophical Review*, 1–31. https://doi.org/10.1080/24740500.2025.2477315

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the wason selection task. *Cognition*, *77*(1), 1–79. https://dx.doi.org/10.1016/S0010-0277(00)00085-8

Goddu, M. K., & Gopnik, A. (2024). The development of human causal learning and reasoning. *Nature Reviews Psychology*, *3*(5), 319–339. https://doi.org/10.1038/s44159-024-00300-5

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. https://dx.doi.org/10.1016/j.cogpsych.2005.05.004

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887. https://doi.org/10.1093/bjps/axi147

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. http://doi.org/10.1016/j.cognition.2017.01.010

Kominsky, J. F., & Phillips, J. (2025). Actual norm violations. *Australasian Philosophical Review*, *9*(2), 198–204. https://doi.org/10.1080/24740500.2025.2592580

Quillien, T., & Lucas, C. G. (2024). Counterfactuals and the logic of causal selection. *Psychological Review*, *131*(5), 1208–1234. https://dx.doi.org/10.1037/rev0000428

Romoli, J., Santorio, P., & Wittenberg, E. (2022). Alternatives in counterfactuals: What is right and what is not. *Journal of Semantics*, *39*(2), 213–260. https://dx.doi.org/10.1093/jos/ffab023

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, *57*(1), 31–95. https://dx.doi.org/10.1016/0010-0277(95)00666-M

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072. https://doi.org/10.1007/s11229-015-0810-5

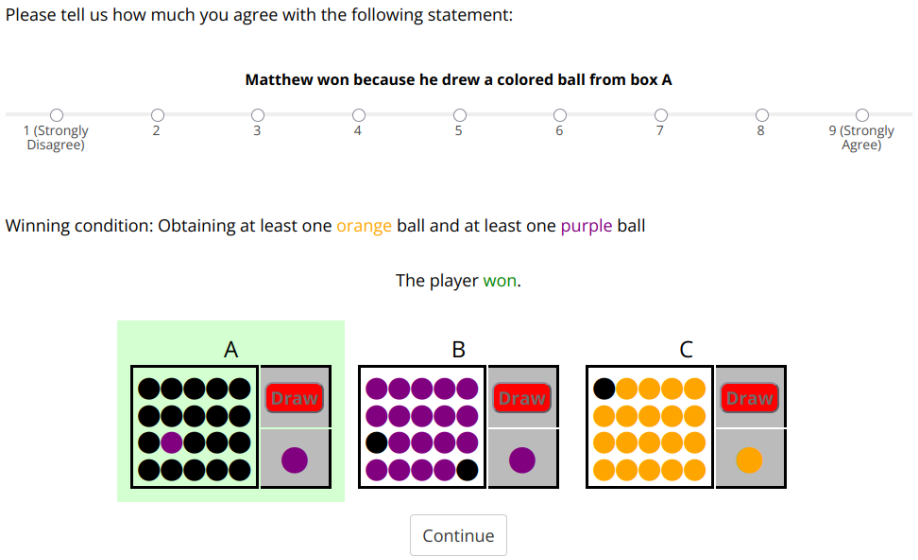

Figure 1: A test trial asking participant their causal judgments towards the 'Win' event in experiment 1.

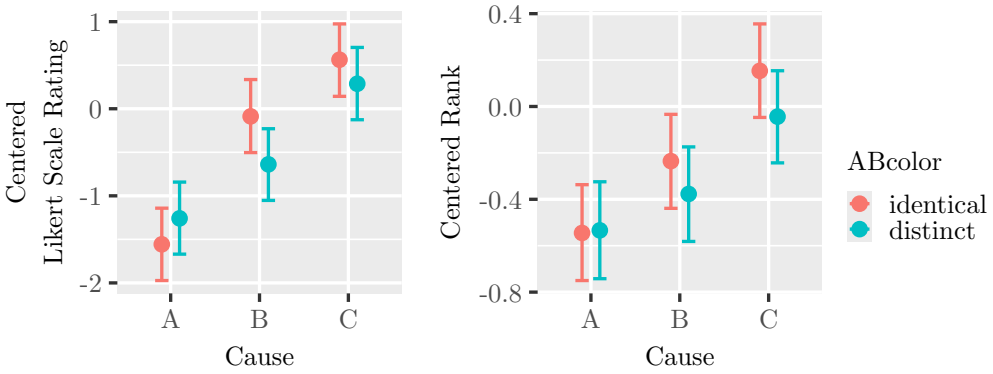

Figure 2: 95% HDIs of centered likert scale ratings (experiment 1, left) and centered ranks (experiment 2, right) obtained from the data across *ABcolor* identical and distinct conditions for the three causes. The likert scale rating for each cause is converted to a centered likert scale rating by subtracting each participant's mean ratings across all conditions. The centered ranks are obtained similarly.