Multiple causes: an experiment in jspsych

[redacted]

1985 words

Online Experiments for Language Scientists

University of Edinburgh

Multiple causes: an experiment in jspsych

Introduction

The experiment has three between-participant conditions. They can be changed between by adding ?condition=1 , ?condition=2 or ?condition=3 after the .html on the end of the url [redacted]

For convenience you can also click to access the three conditions:

Click here for the `pictorial` condition.

Click here for the `numerical` condition.

Click here for the `verbal` condition.

Motivation

When multiple causes contribute to an event, people tend to regard some factors as more "causal" than others. Even when all the factors must be present to cause the effect, we still naturally tend to find the rarer cause more important. For example, the forest fire is caused by the one-off spark, not the ubiquitous oxygen in the air; the election is won by victory in a key swing state, not by the multitude of safely won constituencies. Many studies to date have compared two causes at a time (Icard et al., 2017; Morris et al., 2018; Quillien, 2020b). A popular scenario for discussing questions involving probability is the *random draw*, often involving drawing balls from an urn or box (MacKay & Mac Kay, 2003). For example, Morris et al. (2018) found that when a player needs a coloured ball from each box to win, and they do win, the ball from the rarer bucket is judged more causal. The intuition is that most cases where the rare coloured ball is drawn correlate with winning, because in most there is already a coloured ball drawn from the frequent bucket. Although cognitive modelling of the mental processes involved is outside the scope of the current project, it is important to say that Quillien's model (Quillien, 2020a, 2020b) relies on *counterfactuals*, where people imagine *what may have been*.

The obvious next step is to compare *three* causes and see how people rate causes that vary in their frequency of occurrence. In an as-yet unpublished study, Quillien (2021) had participants play a virtual game where they selected balls from three urns randomly (i.e. the experiment had a “three-cause structure”. Specifically, Quillien (2021) described to participants a simple game (based on Morris et al. (2018)) where the player makes a random draw from three urns, each of which has 20 balls : N coloured balls and 20-N black balls. The player wins the game if they get at least two coloured balls. Quillien had n=290 participants play 10 rounds of the game themselves, and then showed them the outcome of a round that another player (“Joe”) played. In that round, Joe drew a coloured ball from each of the three urns and therefore won the game. Participants were then asked, for each urn, how much they agreed that getting a coloured ball from that urn caused Joe to win the game. Results in a repeated-measures anova showed a significant effect of probability: judgements were significantly highest were for the **intermediate** probability urn. This is perhaps surprising if we followed the predictions of the earlier, two-cause experiments (Morris et al., 2018): because the rarer cause was judged most causal in those experiments, it may have suggested the rarest of the three causes would be judged most causal this time.

This surprising result makes Quillien (2021) results ripe for replication. The current experiment first replicates his design with the smallest change possible (my participants will not play the game themselves). I do this because, although it is conventional to train participants, there is a chance that with repeated exposure to the frequencies, his participants were merely learning associations between the different frequencies and winning, rather than actually thinking about the situation. If we are investigating counterfactuals using Quillien’s setup, it seems the next logical step to test his setup with this one change. Having an existing experiment to replicate in jspsych is a good test because otherwise it could be tempting to cut short at any point and rationalise it as being my aim all along. Beyond that, the next important thing in my next two conditions is to test whether the effect holds across conditions using increasingly verbal stimuli, for reasons detailed in the next section.

Linguistic aspects

Language and vision are arguably our two main modalities of thought, and consequently there is a long history in psychology of debating their separateness versus sameness in different cognitive systems. Potter and Faulconer (1975) claim concepts are stored generally rather than in a specifically lexical system; Kane et al. (2004) found that working memory tasks are domain general, whereas short-term memory tasks are more domain specific (although verbal and visual short term memory may use the same neural systems, Majerus et al. (2010).

And yet, pictures do seem to help with reasoning about concrete problems. Bauer and Johnson-Laird (1993) showed people puzzles expressed in either words or pictures, and found people got many more answers correct (30 percentage points) and answered faster when they used diagrams to answer than verbal descriptions. However, see Oestermeier and Hesse (2000) for a non-empirical discussion of the interplay between words and pictures in causal arguments and of the difficulty of expressing the same thing in words and pictures.

Counterfactual simulation models are well supported in vision, but seem to have been used less in the area of language. Gerstenberg et al. (2021) used realistically moving, animated, "pinball"-type stimuli and monitored participants with eye tracking methods as they asked questions like, "Would Ball A have gone through the gap if Ball B had not knocked it off course". They conclusively showed people "simulating" the trajectory a ball would have been likely to take if it had not been obstructed, i.e. direct evidence in support of counterfactuals.

It is less clear, however, whether similar mechanisms are in play when people reason about causes using language. The design of the urn game lends itself to a "halfway house" where the frequency or contents of the urns can be communicated in words rather than pictures, or even progressively more minimal instantiations of the same basic setup, using "frequency" words.

Design

Scenario and set-up of the current experiment

My current experiment contains three conditions. Condition 1 reproduces the second part of Quillien's experiment ("Joe's game"; see Figure 1). My major change is to omit the first part of his experiment where participants play the game themselves, as justified in the Motivation section above. Condition 2 is the *next* logical step in a progressively more verbal experiment: it encodes the expectations and probabilities of the coloured versus black balls in words rather than pictures (e.g. "Box A has 19 black balls and 1 coloured"). Condition 3 in turn completely bypasses concrete quantities, instead encoding the frequencies only verbally (e.g. "Box A has almost all black balls"). The three conditions are varied between participants to test the robustness of (Quillien, 2020b)'s results. No difference is actually expected between conditions; the most interesting finding would be if the results hold across the three conditions. In all conditions, if the intermediate cause is judged most causal then it replicates the finding of Quillien (2021).

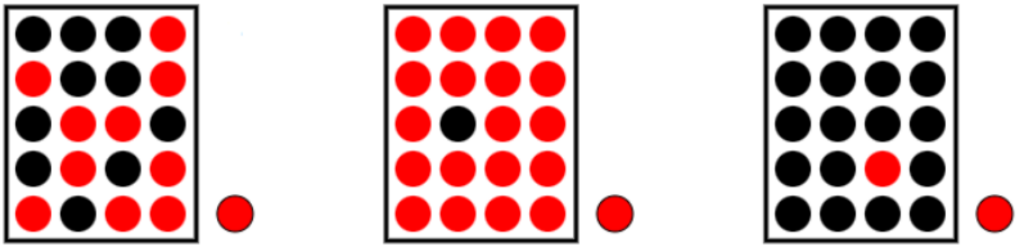

Figure 1

Stimuli from Quillien 2021 showing the intermediate box (50:50 red:black; left), the frequent box (0.95:0.05; middle) and (0.05:0.95; right) used in the preamble and pictorial condition of the current experiment.

You can play Quillien's three-cause structure yourself here. When prompted for a code to continue you can enter any character.

All participants see the same preamble which contains both verbal descriptions and pictures. This was a difficult decision to make because there is a strong argument to give participants a preamble in accordance with their condition (only pictorial, verbal, etc). This way is marginally better however as it keeps as close as possible to the design of Quillien (2021).

After that, participants are shunted to one of the three conditions, where they are shown the three stimuli (rare, intermediate and frequent) in their corresponding format (pictorial, numerical or verbal) in a random order. Each trial contains a buffer saying, “Joe plays the game... He randomly selects a ball from Box [A/B/C]”.

Box names A, B and C are not randomised. As the order of the alphabet is fixed, it would confuse people unnecessarily to counterbalance the letters themselves. I decided best to randomly present only the contents of the boxes.

Importantly, the order of the randomised stimuli is stored in a list and tagged each time it is referred to, so that although the experimenter has no control over what proportions of red and black balls are in Boxes A, B and C, the experiment itself remembers, and shows the participants the correct items at the correct time.

Miscellaneous design decisions

Quillien (2021) used a nine-point Likert scale but I decided on an 11-point scale as this bypasses the constraints of the ordinal scale and allows arithmetic operations e.g. taking a mean (Wu & Leung, 2017).

Text in the instructions is centred rather than left aligned after Quillien (2021).

Appraisal

Despite being rather simple, this experiment is a fully functioning scenario that uses probability and an easily imaginable setup to probe people’s causal judgements.

The most sophisticated part of the coding is in referencing the position of the shuffled stimuli for reminding the participant what was in each box at each stage.

Technical limitations

To preempt reviewer comments, the experiment does have an area where randomisation could be used but it is not. Firstly, there is only one player (Joe). Strictly speaking there should probably be several players with a mix of genders, ethnicities, names etc, to mitigate any associations people may have with the name Joe. However, this aspect is of limited theoretical relevance; if the design holds throughout piloting then I can introduce randomisation here for the real thing.

A further tweak that will need to be implemented before actually running it is to add a comprehension check to ensure that everyone who progresses to the experiment knows how the game works. Similarly, participants will be rewarded with a bonus for getting answers right to keep them motivated. These aspects have not been implemented for the piloting as they are standard practice and less important from either a theoretical or programming point of view.

I have also not dealt here with nuances of how and where the data is saved. That and both the preamble and debrief, as well as consent, timings, ethics, payment, etc, will need to be fleshed out to run the real thing, which I intend to do in early 2022.

This experiment is an adequate implementation of the agreed-upon design. But is it a good test of the *theory?* The next section discusses an insight I had while designing this experiment, which has implications for future work.

Implications for theory

I am testing whether, in a three-cause design, the cause with intermediate frequency is judged most causal. However, it seems the most important aspect of the design is that only two coloured balls are necessary for Joe to win the game. Arguably, because this setup tests “intermediate” across three causes, it fails to distinguish between “intermediate” and “second-highest” because, when there are three items, these are the same thing. The rule could be that “in any case where two conditions are needed, the

second-most-frequent cause is seen as causal".

In other words, there is no inconsistency between the results of the two-ball experiments (Morris et al., 2018), (Quillien, 2020b) and the results of the three-ball experiment: in each case people were judging the "second most frequent" ball as causal. It is **not** that Morris et al. (2019)'s participants were judging the "rarest" ball as causal; they were judging the "second most frequent" ball as causal; it just happens that, in this case, those terms reference the same ball.

This could only be tested with five or more causes, to tease out whether it would be the truly intermediate one (3rd) or the second-closest to the frequent urn (4th). This could be tested in a case where the game can only be won when not two but three coloured balls are drawn.

This realisation does not exactly cast doubt on the whole enterprise, but it does remove the inconsistency between the previous work and Quillien (2021) and hence some of the surprise factor. This is not a problem; on balance it is better to fit consistently with previous work while advancing it, and it does give inspiration for a later series of experiments building on this design.

References

Bauer, M. I. & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological science*, *4*(6), 372–378.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*.

Icard, T. F., Kominsky, J. F. & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W. & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of experimental psychology: General*, *133*(2), 189.

MacKay, D. J. & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Majerus, S., D'Argembeau, A., Martinez Perez, T., Belayachi, S., Van der Linden, M., Collette, F., Salmon, E., Seurinck, R., Fias, W. & Maquet, P. (2010). The commonality of neural networks for verbal and visual short-term memory. *Journal of cognitive neuroscience*, *22*(11), 2570–2593.

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T. & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions.

Oestermeier, U. & Hesse, F. W. (2000). Verbal and visual causal arguments. *Cognition*, *75*(1), 65–104.

Potter, M. C. & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature*, *253*(5491), 437–438.

Quillien, T. (2020a). *Actual causation, normality, and the mind as intuitive statistician.* https://xphiblog.com/actual-causation-normality-and-the-mind-as-intuitive-statistician/

Quillien, T. (2020b). When do we think that x caused y? *Cognition*, *205*, 104410.

Quillien, T. (2021). *Report: Two studies on causation with a three-cause structure.*

Wu, H. & Leung, S.-O. (2017). Can likert scales be treated as interval scales?—a simulation study. *Journal of Social Service Research*, *43*(4), 527–532.

[redacted]

GRADEMARK REPORT

FINAL GRADE

75/100

GENERAL COMMENTS

Instructor

Clear review of previous experiments on multiple causes. However, it felt unclear to me whether your new conditions were really testing something different from the Quillien experiment, and specifically whether you're really getting at something linguistic here since (as you point out) there is still quite a lot of visual information in the conditions that are intended to be more verbal. Nonetheless, a replication with added linguistic framing is a sensible rationale and it's clear that you've thought carefully about how to achieve that.

The experiment looks good and clearly departs from the lab code. You've successfully implemented three separate conditions which is great - just remember that you'd need to think about random condition assignment if you were running this for real. There's quite a lot going on visually (sometimes red balls, sometimes orange; sometimes a simple birds-eye view of the box, sometimes a more complex 3D rendition) but luckily the critical task is quite simple so I think your participants would be able to follow easily enough. The Likert scales worked although I did wonder whether a slider might have been easier from a participant perspective (there's a jsPsych plugin for this!). Instructions were nice and clear.

Your report is well written and justifies your design decisions reasonably. It's also clear that you've given thought to the logistics of running this for real even if you haven't implemented all of those details yet. Your insight about a potential explanation for the results of previous work is an interesting one.

Strengths: Clear report and strong lit review; good working experiment with limitations clearly acknowledged.

Weaknesses: Nothing major in the implementation; some skepticism about the linguistic aspect of the research question.

PAGE 1

PAGE 2

PAGE 3

Comment 1

I know what you mean, in that you've replicated the rest of the setup and retained the critical trial, but at face value this sounds like quite a big change! May have been worth exploring the potential implications of this decision in more detail.

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10