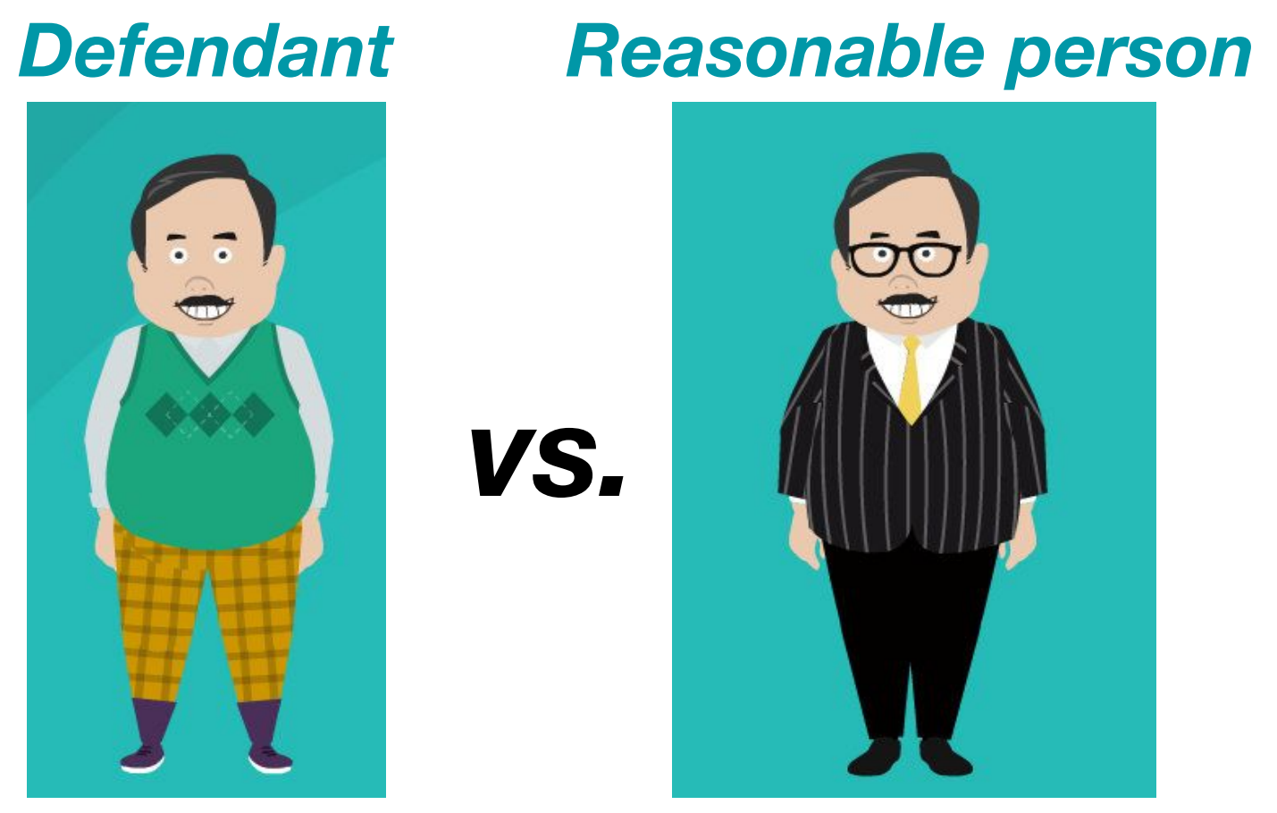
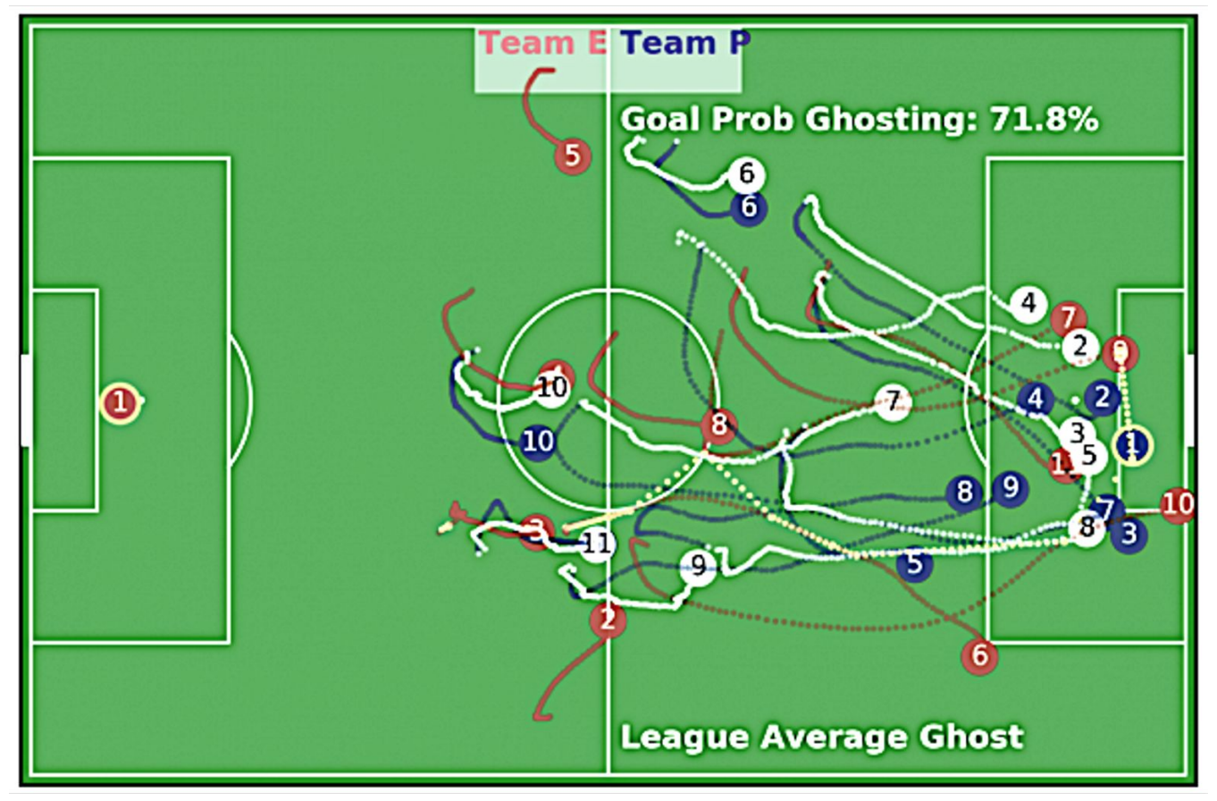
The role of counterfactual reasoning in responsibility judgments

Sarah Wu (sarahawu@stanford.edu) & Tobias Gerstenberg
Department of Psychology, Stanford University

Introduction



"Reasonable person" legal standard



"Ghosting" in AI sports analytics

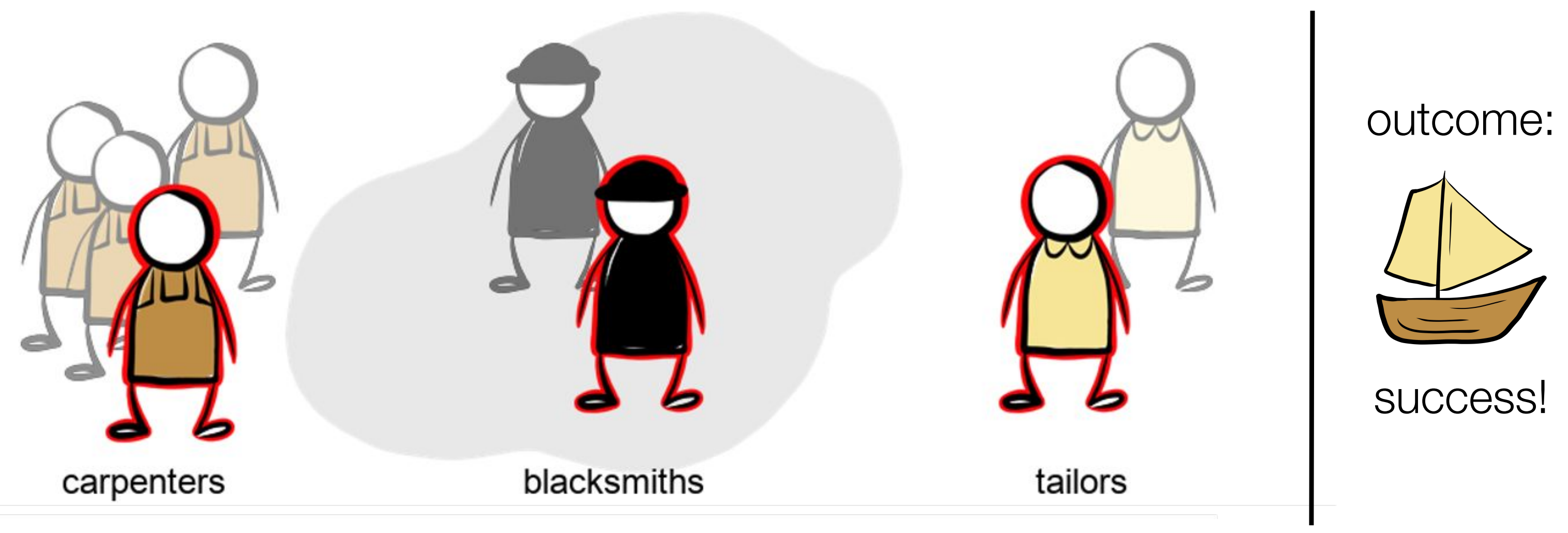
Responsibility ⟷ Causality ⟷ Counterfactuals

- Extending counterfactual theories of causation[1-3]
- Responsibility as difference-making: judging by comparing *what happened* with *what would have happened* in relevant counterfactual situations
- Counterfactual potency = *if-likelihood* x *then-likelihood*[4]

Experiment 1

Setup: OSF

- Three agents contribute equally to a positive outcome
- Agents may be busy, but others can take their place
- Manipulate number of possible replacements



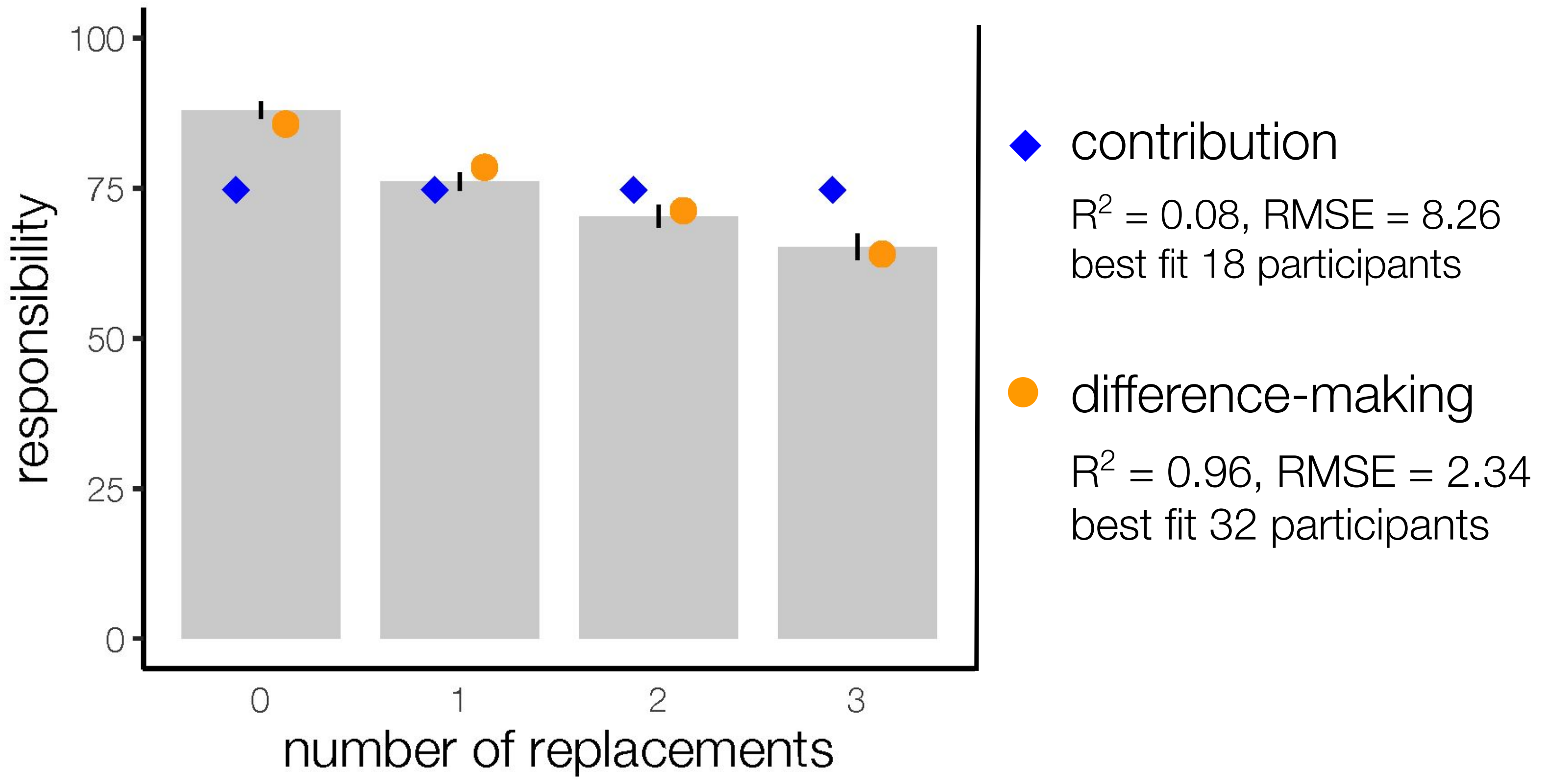
How responsible are each of the three highlighted agents for the success?

Hypotheses:

- ◆ **Contribution model:** Responsibility is about individual contribution
  - Prediction: Uniform judgments across all trials
- ● **Difference-making model:** Responsibility is about counterfactual difference-making
  - Prediction: number of replacements ↑, responsibility ↓

Experiment 1 (cont.)

Results: Responsibility decreases with the number of replacements, even when the outcome and individual contributions are the same. (n = 50, error bars = 95% CI)

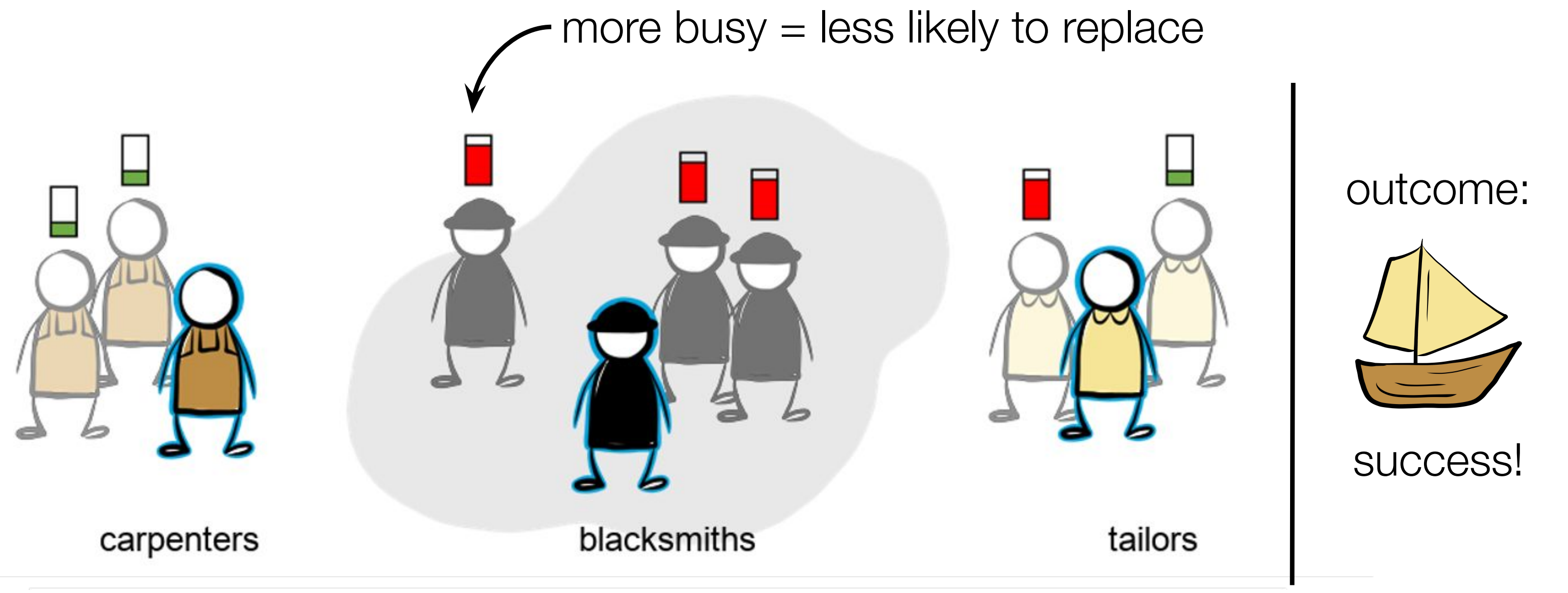


What if participants are simply mapping the number of replacements without computing any counterfactuals?

Experiment 2

Setup: OSF

- Manipulate number and "busyness" of replacements

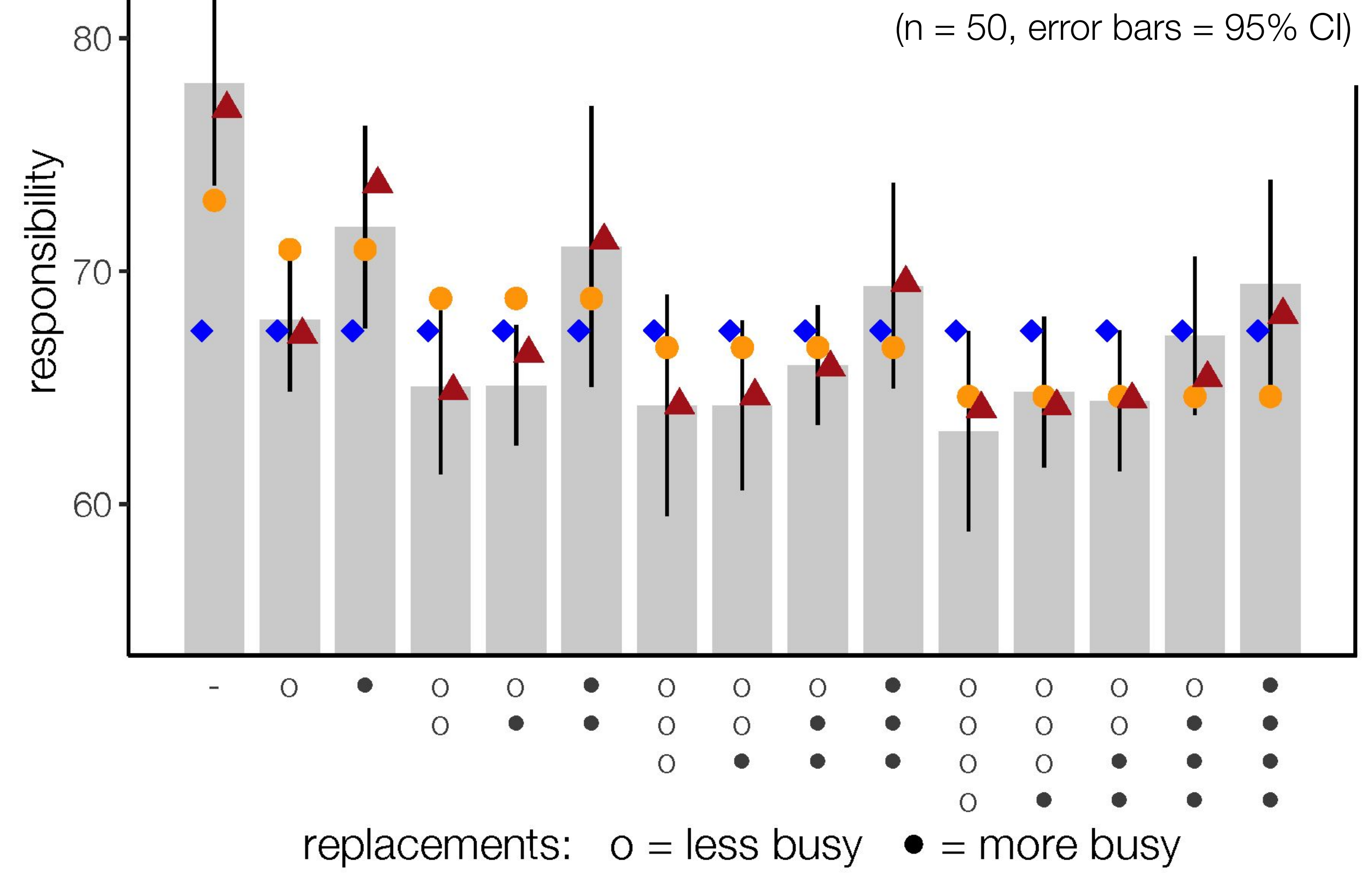


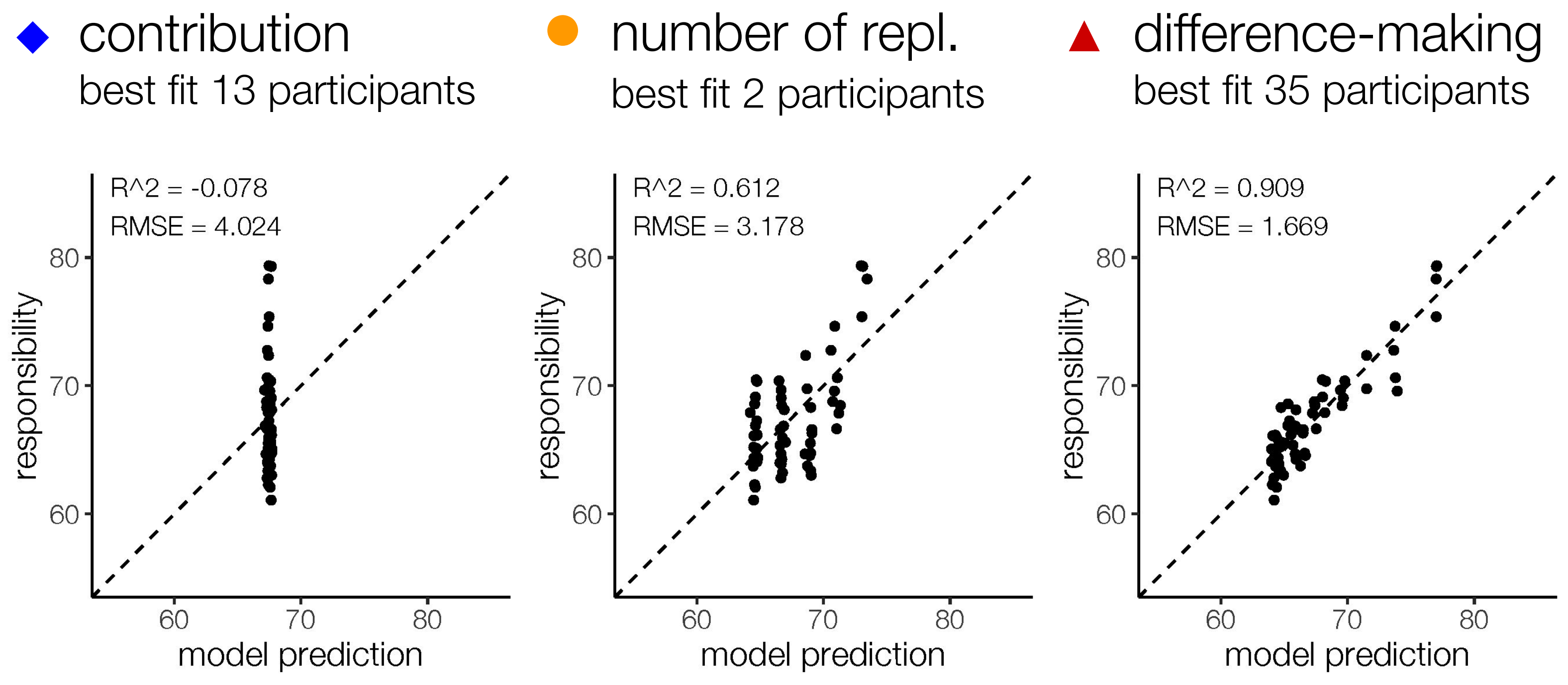
Hypotheses:

- ◆ **Contribution model** (same as Experiment 1)
- ● **Number of replacements model:** Mapping number of replacements without counterfactuals
  - Prediction: number of replacements ↑, responsibility ↓, but no difference with varying busyness
- ▲ **Difference-making model:** Responsibility is about counterfactual difference-making
  - Prediction: probability of finding available replacement (based on number and busyness) ↑, responsibility ↓

Experiment 2 (cont.)

Results: Responsibility decreases with the probability of replacement, not simply the number.

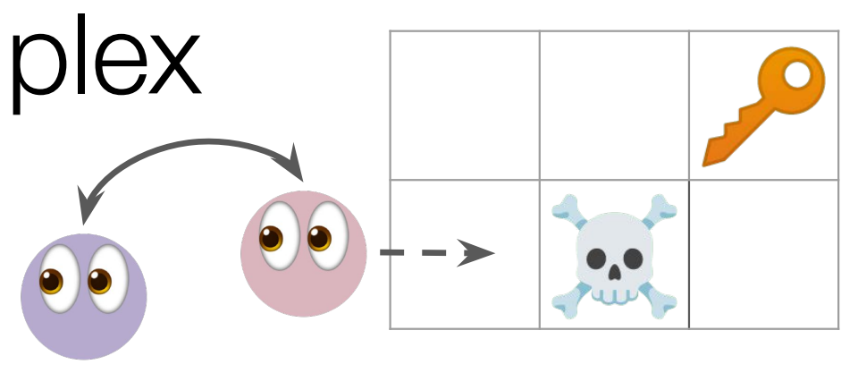




Discussion

- In judging responsibility, people consider counterfactual scenarios and assign responsibility to the extent that counterfactual outcomes would have been different
  - Here: "what if the highlighted agent had been busy?"
- Responsibility ratings well predicted by *then-likelihood*

Future directions: exploring more complex and explicitly simulated counterfactuals



References

1. Lewis (1973). *J Philos.* 2. Halpern & Pearl (2005). *Br J Philos Sci.* 3. Gerstenberg et al. (2021). *Psychol Rev.* 4. Petrocelli et al. (2011). *J Pers Soc Psychol.*