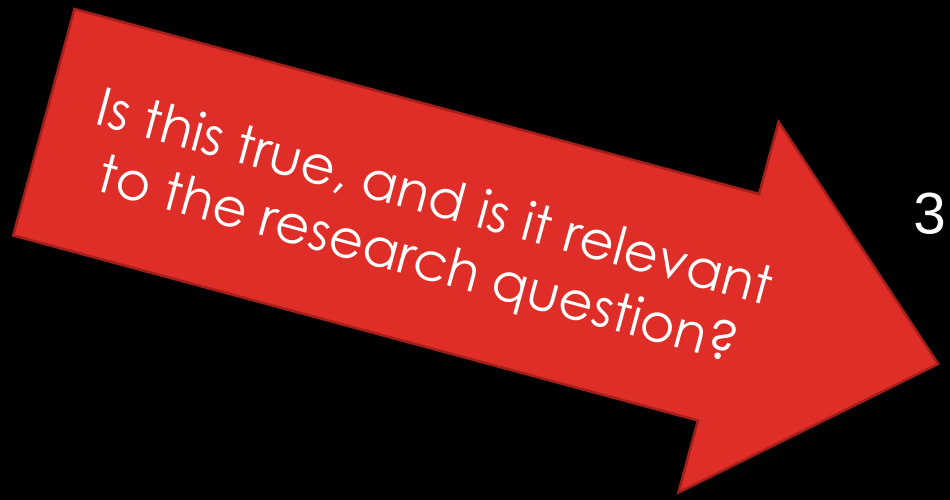
Deconstructive Replication: How to Tell if a Thought Experiment-Population Pairing is Fit for Purpose

Dan Weijers & Peter Unger

AXΦII: 2nd Annual Australasian Experimental Philosophy Conference, 29 July 2017, University of Waikato

CHALLENGING THOUGHT EXPERIMENTS

- Are thought experiments useful or misleading?



1. An experience machine life has more happiness than a normal life
2. If happiness is all that really matters in life, then the vast majority of people would choose an experience machine life over a normal life
3. The vast majority of reasonable people would not choose an experience machine life over a normal life
4. Therefore, happiness is not all that really matters in life

INSPIRATION

- While explaining my “replicate with a tweak” process to my wife:
- “Sounds like you are just manipulating the scenario until you get the result you want”
- Evidence for a bias in another domain is not enough

- What we need is a way to assess a scenario-population pairing’s fitness for purpose in relation to a particular research question
- Different scenario-population pairings can then be compared to see which is better for answering the research question

DECONSTRUCTIVE REPLICATION

- A rigorous method for assessing the scenario-population pairing's fitness for purpose in relation to the relevant research question

Useful research agendas for x-phi!

- See how fit-for-purpose famous thought experiments are
- See whether they can be improved
- See whether certain populations are the problem, or whether the scenario is the problem
  - Implications for expertise

THE BASIC METHOD

- Test the original scenario on a population
- Ask why they chose that
- Assess the justifications (& relevant biases lit.) to see what scenario or population modifications are suggested
- Make those changes, Preferably 1-by-1, test in the same way, compare results

There is this amazing experience machine designed by super-duper neuropsychologists…

1. Would you live the rest of your life in such a machine?
   - Yes
   - No
2. Briefly justify your choice for 1.

THE BASIC METHOD

- Test the original scenario on a population
- Ask why they chose that
- Assess the justifications (& relevant biases lit.) to see what **scenario or population** modifications are suggested
- Make those changes, Preferably 1-by-1, test in the same way, compare results

There is this amazing experience machine designed by super-duper neuropsychologists… **NO MORE IN AND OUT EVERY TWO YEARS**

1. Would you live the rest of your life in such a machine?
   - Yes
   - No
2. Briefly justify your choice for 1.

THE METHOD NEEDS TO…

…test for these things:

- Overall fitness for purpose
- Scenario fitness for purpose
- Population fitness for purpose
- Relative fitness for purpose statistical tests for each of these

…provide suggestions for:

- What about the scenario might be causing the trouble
- What about the population might be causing the trouble

THE QUALITATIVE ANALYSIS PART OF DECONSTRUCTIVE REPLICATION

Method

There are 9 main categories of answer. Which main category an answer fits in depends on 3 things:

1. The research question
2. The exact wording of the scenario and questions
3. The respondents' answers to the main and "please justify" questions

Then group by meaning within categories

Categories

1. Malicious response
2. Opposite justification
3. Imaginative resistance
4. Overactive imagination
5. Reasonable resistance
6. Useful response
7. No justification
8. Reasonable rejection
9. Demonstrates misunderstanding

DESCRIPTION OF THE CATEGORIES

Categories

1. Malicious response – e.g. “F*@k this s#!t”
2. Opposite justification – i.e. justification perfectly matches alternate choice
3. Imaginative resistance – e.g. “The exp. mach. wouldn’t make *me* feel good”
4. Overactive imagination – e.g. “I’d need electrodes plugged into my brain!”
5. Reasonable resistance – e.g. “I don’t want to force my family to plug in too”
6. Useful response – i.e. justification matches response & not 1-5,6-9
7. No justification – e.g. “ ”
8. Reasonable rejection – e.g. “You have forced me into a false dichotomy”
9. Demonstrates misunderstanding – e.g. “There’s no way I’m getting into an unhappiness machine designed by super-duper sadists!

DECONSTRUCTIVE REPLICATION: QUANT. OF QUAL. ANALYSIS

Categories

1. Malicious response
2. Opposite justification
3. Imaginative resistance
4. Overactive imagination
5. Reasonable resistance
6. Useful response
7. No justification
8. Reasonable rejection
9. Demonstrates misunderstanding

Uses

- % of Cat. 5+8+(0.5x(3+4+7)) answers = **a scenario's unfitness for purpose**
- Compare "unfitness for purpose" scores of different scenarios on the same population to see whether the scenario is more fit for purpose
  - E.g. Fisher's Exact test (1-tailed?)
- Investigate order effects
  - E.g. X+5 life before X increases CS view 21%
  - But why?
  - E.g. creased imaginative resistance (regarding last 5 years being unenjoyable)

DECONSTRUCTIVE REPLICATION: QUANT. OF QUAL. ANALYSIS

Categories

1. Malicious response
2. Opposite justification
3. Imaginative resistance
4. Overactive imagination
5. Reasonable resistance
6. Useful response
7. No justification
8. Reasonable rejection
9. Demonstrates misunderstanding

Uses

- % of Cat. 1+2+9+(0.5x(3+4+7)) answers = **a population's unfitness for purpose**
- Compare "unfitness for purpose" scores a different population given the same scenario to see whether the population is more fit for purpose
  - E.g. Fisher's Exact test (1-tailed?)
- Split your sample by a demographic factor to see whether some groups are less fit for purpose
  - E.g. showing that youths are less fit for purpose
  - And have evidence for why
    - E.g. old people can't be happy

DECONSTRUCTIVE REPLICATION: QUANT. OF QUAL. ANALYSIS

Categories

1. Malicious response
2. Opposite justification
3. Imaginative resistance
4. O[illegible]ctive imagination
5. R[illegible]ance
6. [illegible]
7. [illegible]
8. Reasonable [illegible]
9. Demonstrates misunder[illegible]

Interesting question: should we use clean or dirty TEs for teaching?

Uses

- % of Cat. 1+2+3+4+5+7+8+9 answers = **a scenario-population pairing's <u>un</u>fitness for purpose**
  - E.g. Nozick's experience machine-1st years = 39%+ unfitness (Cat. 3,4,5; Weijers, 2014)
- Compare this with other scenario-population pairing's fit for purpose (raw) score for a statistical test of relative fitness for purpose
  - E.g. Fisher's Exact test (1-tailed?)
  - Grounds for replacing famous thought experiment (at least with some populations)
  - E.g. My "Self" experience machine-1st years = 23%+ unfitness (Cat. 3,4,5; Weijers, 2014)

DECONSTRUCTIVE REPLICATION IN ACTION

- 2B (see handout)

2B	Pre anal-ysis #	Pre-a %	Post anal-ysis #	Post-a %	Clean differ-ence %	Clean p-value
\	23	56%	17	81%	25%	0.0911
/	8	20%	0	0%	-20%	0.0429
|	10	24%	4	19%	-5%	0.7548
Tot	41	100%	21	100%		**0.0620**

- Pairing fit-for-purpose = 51%
- Pairing unfit-for-purpose = 49%
- Scenario unfit-for-purpose = 27%
- Population unfit-for-purpose = 22%
- Cat.7 = 15% (no response)
- Cat.9 = 12% (demos misunderstanding)
  - Thinks one subject is still alive and will miss their friend (the other one!)
- Cat.5 = 10% (reasonable resistance)
  - Says the long life especially is not desirable because of being single with no kids

DECONSTRUCTIVE REPLICATION IN ACTION

- 2B3 (see handout)

2B3	Pre anal-ysis #	Pre-a %	Post anal-ysis #	Post-a %	Clean differ-ence %	Clean p-value
\	15	52%	15	63%	11%	0.5787
/	1	3%	0	0%	0%	1.0000
|	13	45%	9	38%	-7%	0.7799
Tot	29	100%	24	100%		**0.6715**

- Pairing fit-for-purpose = 83%
- Pairing unfit-for-purpose = 17%
- Scenario unfit-for-purpose = 9%
- Population unfit-for-purpose = 9%
- Cat.7 = 14% (no response)
- Cat.9 = 0% (demos misunderstanding)
- Cat.5 = 0% (reasonable resistance)

DECONSTRUCTIVE REPLICATION IN ACTION

- **2B**
- Pairing fit-for-purpose = 51%
- Pairing unfit-for-purpose = 49%
- Scenario unfit-for-purpose = 27%
- Population unfit-for-purpose = 22%
- Cat.7 = 15% (no response)
- Cat.9 = 12% (demos misunderstanding)
- Cat.5 = 10% (reasonable resistance)

- **2B3 (looks cleaner by eye)**
- Pairing fit-for-purpose = 83%
- Pairing unfit-for-purpose = 17%
- Scenario unfit-for-purpose = 9%
- Population unfit-for-purpose = 9%
- Cat.7 = 14% (no response)
- Cat.9 = 0% (demos misunderstanding)
- Cat.5 = 0% (reasonable resistance)

DECONSTRUCTIVE REPLICATION IN ACTION

- Fisher's exact 2x3 2-tailed

2B	Pre anal-ysis #	Pre-a %	Post anal-ysis #	Post-a %	Clean differ-ence %	Clean p-value
\	23	56%	17	81%	25%	0.0911
/	8	20%	0	0%	-20%	0.0429
|	10	24%	4	19%	-5%	0.7548
Tot	41	100%	21	100%		**0.0620**

2B3	Pre anal-ysis #	Pre-a %	Post anal-ysis #	Post-a %	Clean differ-ence %	Clean p-value
\	15	52%	15	63%	11%	0.5787
/	1	3%	0	0%	0%	1.0000
|	13	45%	9	38%	-7%	0.7799
Tot	29	100%	24	100%		**0.6715**

So, 2B3 is cleaner.

DECONSTRUCTIVE REPLICATION: QUANT. OF QUAL. ANALYSIS

Categories

1. Malicious response
2. Opposite justification
3. Imaginative resistance
4. Overactive imagination
5. Reasonable resistance
6. Useful response
7. No justification
8. Reasonable rejection
9. Demonstrates misunderstanding

Uses

- % of Cat. 6 (useful response) answers = **a scenario-population pairing's fitness for purpose**
- Compare this with other scenario-population pairing's fit for purpose (raw) score for a statistical test of relative fitness for purpose
  - E.g. Fisher's Exact test (1-tailed?)

ISSUE #1: SHOULD CAT.7 (NO RESPONSE) COUNT AS USEFUL?

- No
- We can't qualitatively check them, so they are a risk
- Quantitative analysis **?might?** reveal that the Cat.7 responses are much more like random answers than the useful (Cat. 6 answers)
- An "eye-ball analysis" of 34 studies shows that in 65% of the studies (22/34), the Cat. 7 scores were different to the Cat. 6 scores
- A FUTURE analysis of 34 studies **?might?** show that in over 50% of the studies, the Cat. 7 scores were not statistically indistinguishable from the random chance result

ISSUE #2: SHOULD ONLY CAT.6 (USEFUL) RESPONSES BE USED IN INITIAL QUANTITATIVE ANALYSIS?

- Depends on research qn – "reasonable people think…"
- But, mainly yes!
- We have established all other responses as problematic, so why include them?
  - Even if we got some wrong, the data should still be cleaner overall
- Unhelpful responses (Cat. 1-5,7-9) are often not distributed evenly between the options, so it can make a big difference to the results
- Run statistical test on the cleaning
  - E.g. 1-tailed 2*X Fisher's exact test
  - X=number of response options

- 2C2 (see handout)

	Pre anal-ysis #	Pre-a %	Post anal-ysis #	Post-a %	Clean differ-ence %	Clean p-value
A	13	30%	0	0%	-30%	0.0138
B	8	19%	0	0%	-19%	0.0975
C	22	51%	15	100%	**49%**	0.0004
Tot	43	100%	15	100%		**0.0019**

ISSUE #3: TAKING RESPONDENTS' JUSTIFICATIONS TOO SERIOUSLY?

- From Haidt etc., we know that these justifications might (at least) be bogus, so why trust them?
- This is partial trust – they are used as a guide, and the quantitative analysis of the next iteration of the scenario-population pairing indicates whether those justifications were trustworthy
- Qualitative cleaning can be seen as a kind of competence test
- So, just using the cleaned results is increasing the chances of the population being "reasonable" at least in regards to the particular scenario

ISSUE #4: WHAT COUNTS AS A USEFUL PROPORTION FOR PHILOSOPHY?

- It should depend on whether the cleaned results are used
- Dirty results: 70%+ = vast majority of reasonable people?
- Clean results: 80% = vast majority of reasonable people?

1. An experience machine life has more happiness than a normal life
2. If happiness is all that really matters in life, then the vast majority of people would choose an experience machine life over a normal life
3. **The vast majority of reasonable people** would not choose an experience machine life over a normal life
4. Therefore, happiness is not all that really matters in life

CONCLUSION

Things I'd like feedback on

- Name: "deconstructive replication"
- Phrase: "scenario-population pairing"
- "Vast majority" levels for clean and dirty results
- 1 vs 2-tailed fisher's exact test use
- Are the nine categories right? Need more or less?
- Say 2 independent qualitative coders is good, and 1 is OK, and doing it yourself is not OK
- Would you use the method? Why/why not?