

Using open data to teach reproducibility

SSC Annual meeting

Léo Belzile, HEC Montréal

Wednesday, May 28, 2025

Background information

I teach “Experimental Design and Statistical Methods” to PhD students in management sciences, coming from

- marketing,
- organization behaviour and human resources,
- user experience,
- information technology, etc.

Students have **heterogeneous background** in terms of both statistics knowledge and programming skills.

Reproducibility crisis

Experimental Economics Replication Project

In a systematic replication project of experimental studies published in high-impact economics journals 61% replicated.

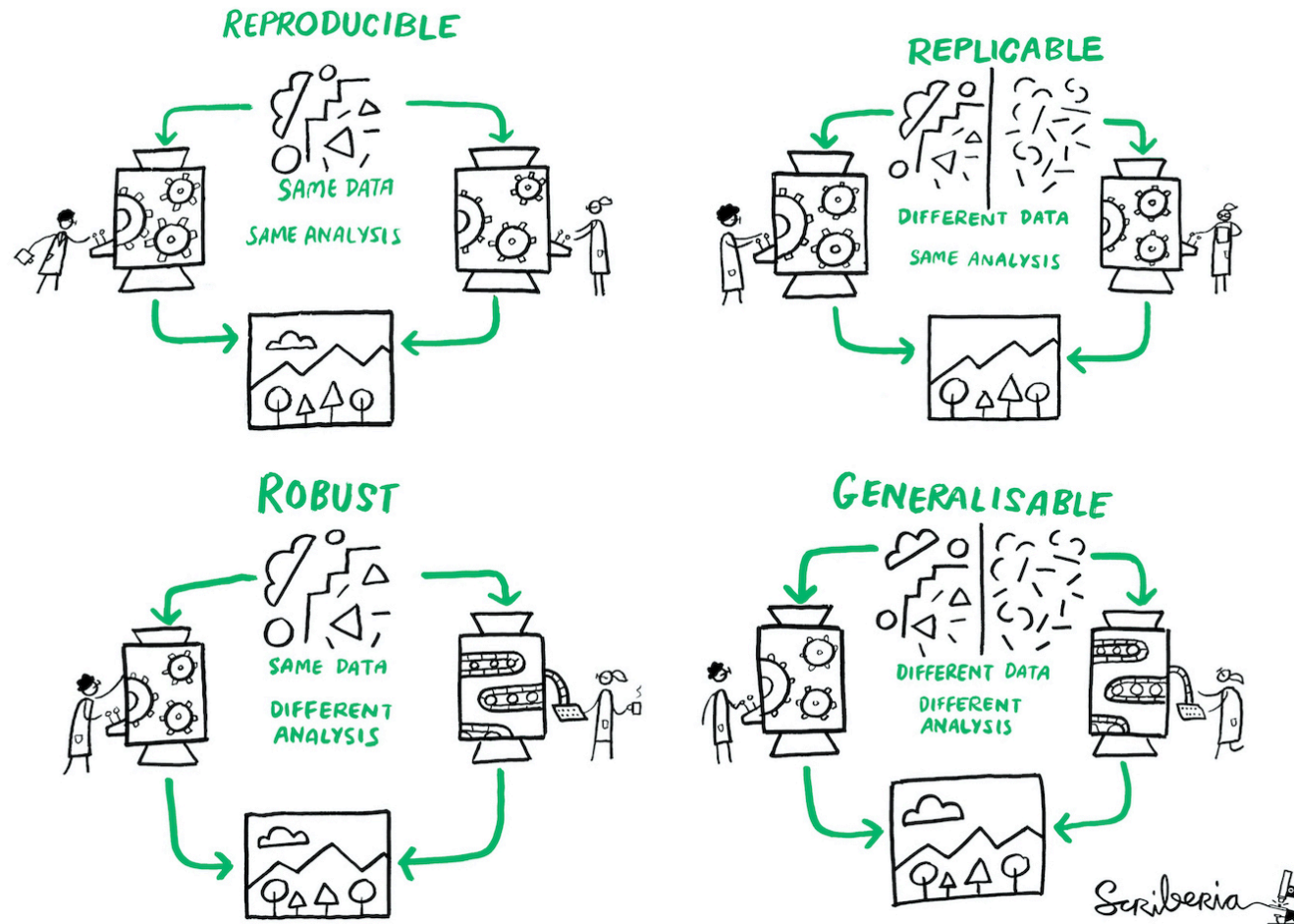
Abstract

The reproducibility of scientific findings has been called into question. To contribute data about reproducibility in economics, we replicate 18 studies published in the *American Economic Review* and the *Quarterly Journal of Economics* in 2011-2014. All replications follow predefined analysis plans publicly posted prior to the replications, and have a statistical power of at least 90% to detect the original effect size at the 5% significance level. We find a significant effect in the same direction as the original study for 11 replications (61%); on average the replicated effect size is 66% of the original. The reproducibility rate varies between 67% and 78% for four additional reproducibility indicators, including a prediction market measure of peer beliefs.

The garden of forking paths

- Change of paradigm and generational change: my students supervisor have not been trained to do work properly, we need to form the students to adapt to the evolving landscape.
- Good practices (preregistration, sharing code and data, transparency) help alleviate reproducibility problems.
- New easy-to-use tools out there to help researchers (e.g., ResearchBox).
- Instructors can leverage these to generate learning opportunities!

Defining reproducibility



Definition of different dimensions of reproducible research (from The Turing Way project, illustration by Scriberia).

Ingredients for reproducible research

Mandatory

- Raw or cleaned data
- Methodological details:
 - models
 - variables
 - statistics (coefficients, standard errors, degrees of freedom, p -values, effect size, confidence intervals).

Optional

- Preregistration report
- Questionnaires, etc.
- Code

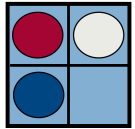
Details may be hidden in code (also via comments) and in the supplementary material.

Checking data agreement

It is helpful if papers report descriptive statistics (sample size, sociodemographic) to ensure we use the same data subsample.

There should be a clear description of exclusion rules or missing values for raw data.

How to find open data



RESEARCHBOX
Open Research Made Easy

- Data drawn from Open Science Foundation (OFS) or ResearchBox
- Use search with keywords (abstract, or code)

ResearchBox #282 - 'Virtual Communication Curbs Creativity'

Bingo Table

- Show file names
- Show file IDs
- Show timestamps

<input type="checkbox"/> Select All	<input type="checkbox"/> Pre-Registration	<input type="checkbox"/> Materials	<input type="checkbox"/> Data	<input type="checkbox"/> Code	<input type="checkbox"/> Other
<div>Alt Explanation:</div> <div><input type="checkbox"/> Conversation Coordination</div>			<div><input type="checkbox"/> csv - Convo Coordination</div> <div></div>	<div><input type="checkbox"/> R Code</div>	<div><input type="checkbox"/> Read-me - data for analyses</div>
<div>Alt Explanation:</div> <div><input type="checkbox"/> Group Processes</div>				<div><input type="checkbox"/> R Code - Dominance</div> <div><input type="checkbox"/> R Code - Fear of Eval</div> <div><input type="checkbox"/> R Code - Illusion of Prod</div> <div><input type="checkbox"/> R Code - Production Block</div> <div><input type="checkbox"/> R Code - Social Facil</div> <div>>>This cell has more files.</div>	<div><input type="checkbox"/> Read-me - data for analyses</div>
<div>Alt Explanation:</div> <div><input type="checkbox"/> Mimicry</div>			<div><input type="checkbox"/> csv - Facial Mimicry</div> <div><input type="checkbox"/> csv - Linguistic Mimicry</div> <div></div>	<div><input type="checkbox"/> R Code - Facial Mimicry</div> <div><input type="checkbox"/> R Code - Linguistic Mimicry</div>	
<div>Alt Explanation:</div> <div><input type="checkbox"/> Non Verbal Behavior</div>			<div><input type="checkbox"/> csv - Muted Video Coding</div> <div></div>	<div><input type="checkbox"/> R Code - Muted Video Coding</div>	

Criteria for selecting papers

Papers should cover a broad variety of fields

- interesting story or research question
- good experimental design
- course material overlap!
- enough details for reproducibility

For my course, diversity of topics and fields is important.

Sharing open-access data

All datasets are preprocessed and bundled via an R package with documentation (with online page).

- Data cleaning is an essential tool, but too big a hurdle for my audience for anything but simple wrangling
- Some papers provide cleaning script, but
 - the code is often messy,
 - different softwares are used and
 - package obsolescence (tidyverse) complicates things.

Activities based on published papers

- Presentation of case studies in class
- Discussion and criticism of methodological and results section
- Weekly assignments (reproducibility)
- Peer-review project in teams

Lesson plan

- Recap of previous session, and problem set review
- Story time: describe the study for ‘motivation’
- Substantive content (methods, concepts)
- Reproducing results (workbook with code and conclusions, versus the paper)
- Annotated papers with comments and criticism

Peer-reviewing task

- Students learn by example, not by osmosis
- Show them the good, the bad and the ugly.
- Compare code with the paper, if available:
 - what is omitted?
 - do results and descriptions match?

What we don't (often) teach

- How to do randomization (especially for repeated measures)
- How to choose the response variable
- How to clean the data
- How to make effective data visualization (dynamite plots)!
- How to defined sensible exclusion criteria
- How to report conclusions

We can learn or teach these from (counter)examples drawn from published papers.

1. Impact of videoconferencing on idea generation



In the laboratory experiment, we randomly assigned half of the pairs to work together in person and the other half to work together in separate, identical rooms using videoconferencing. The pairs in the virtual condition interacted with a real-time video of their partner's face displayed on a 15-inch retina-display screen with no self-view. The image was taken during the first batch of data collection in the laboratory. Consent was obtained to use these images for publication.

Brucks and Levav (2022)

In a laboratory study [...] we demonstrate that videoconferencing hampers idea generation because it focuses communicators on a screen, which prompts a narrower cognitive focus. Our results suggest that virtual interaction comes with a cognitive cost for creative idea generation.

(Subjective) measurement of the number of creative ideas, variety of models that can be fit, comparing different tests.

- Model: Linear mixed model, negative binomial regression

2. Suggesting amounts for donations

Moon and VanEpps (2023)

Across seven studies, we provide evidence that quantity requests, wherein people consider multiple choice options of how much to donate (e.g., \$5, \$10, or \$15), increase contributions compared to open-ended requests.

Our findings offer new conceptual insights into how quantity requests increase contributions as well as practical implications for charitable organizations to optimize contributions by leveraging the use of quantity requests.

Model: Tobit type II regression and Poisson regression (independence test)

FIGURE 1

DONATION REQUESTS IN THE REAL WORLD

A

Your impact: \$1 = 10 meals
Every dollar you give can provide at least 10 meals to families in need through the Feeding America network of food banks.

Gift Amount (Required)

☒ One-Time ☐ Monthly

Giving monthly is the most effective way to help feed kids and families year-round.

B

Direct Relief

Giving is good medicine.
Everything Direct Relief does, every person whose life is saved or improved, stems from an act of generosity that's purely voluntary. You don't have to donate. That's why it's so extraordinary when you do.

For other ways to donate, including mailing a check or making a stock donation, please [click here](#).

NOTE.—A quantity request used by Feeding America with donation choice options ranging from \$25 to \$1,000 (A) and an open-ended request used by Direct Relief (B).

3. Integrated decisions in online shopping

Duke and Amir (2023)

Customers must often decide on the quantity to purchase in addition to whether to purchase. The current research introduces and compares the quantity-sequential selling format, in which shoppers resolve the purchase and quantity decisions separately, with the quantity-integrated selling format, where shoppers simultaneously consider whether and how many to buy. Although retailers often use the sequential format, we demonstrate that the integrated format can increase purchase rates.

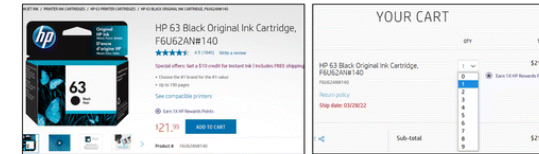


Figure 1. (Color online) Example Quantity-Sequential Selling Format, HP Printer Ink Notes. (Left) Screenshot of a product information page shown to customers on HP's website. (Right) Screenshot of the shopping cart page after adding this item to the cart. Web locations: <https://store.hp.com/us/en/pdp/hp-63-black-original-ink-cartridge> and <https://store.hp.com/us/en/AjaxOrderItemDisplayView>. Both screenshots were taken on March 25, 2022.



Figure 2. (Color online) Hypothetical Quantity-Integrated Selling Format, HP Printer Ink Note. To accommodate additional quantity options, the website could also consider a fourth button such as "Add 4 or more to cart"; alternatively, customers could simply click one of the presented "add" buttons multiple times.

Model: logistic regression.

4. Subjective perception of distance

Maglio and Polman (2014) postulate that “spatial orientation toward (vs. away from) a location will cause that location to feel closer.”

202 volunteers were recruited at the Bay Street subway station in Toronto, Ontario, Canada. All participants were asked to rate the subjective distance of another subway station on the line that they were traveling; the station was either coming up (e.g., the next stop) or just past (e.g., the previous stop).

Model: 2 x 4 between-subject ANOVA.

Refactoring and contrasts

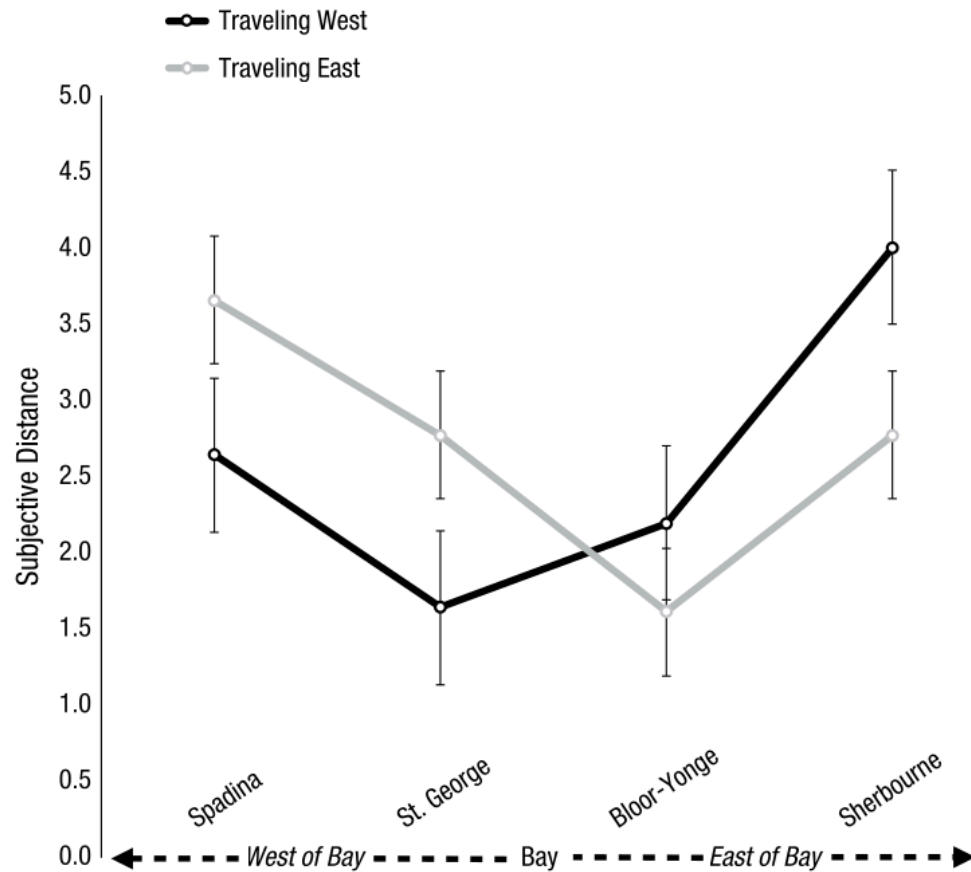
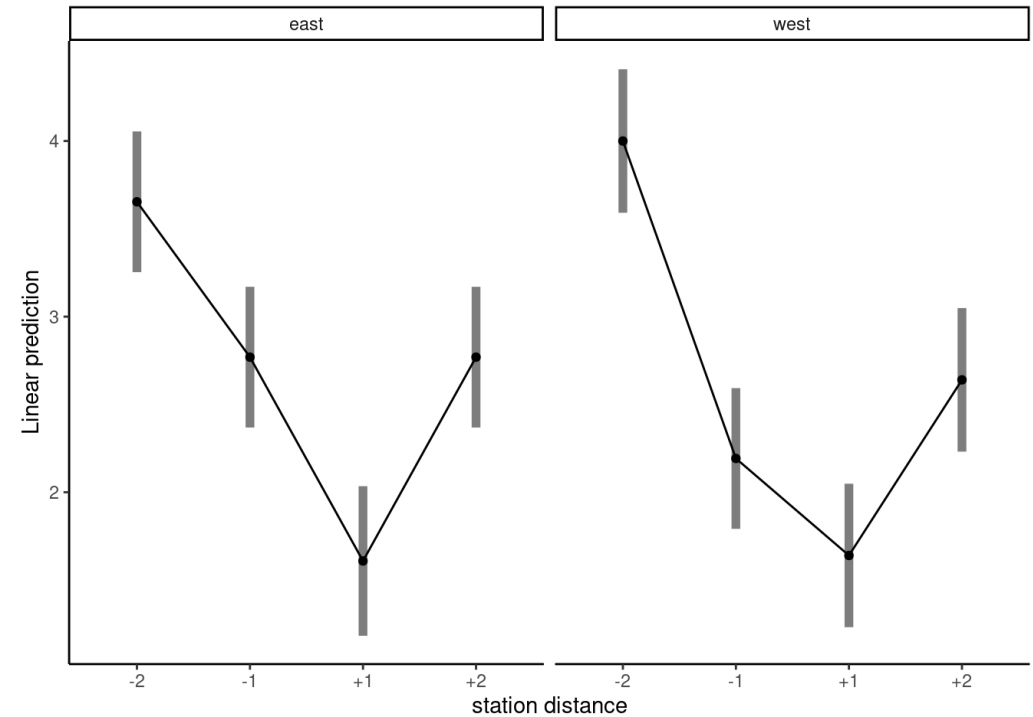


Fig. 1. Results from Study 1: subjective-distance rating as a function of the subway station being evaluated and the participant's orientation. All participants were physically located at the Bay Street station, at the midpoint between the St. George and Bloor-Yonge stations. Error bars indicate ± 1 SE.

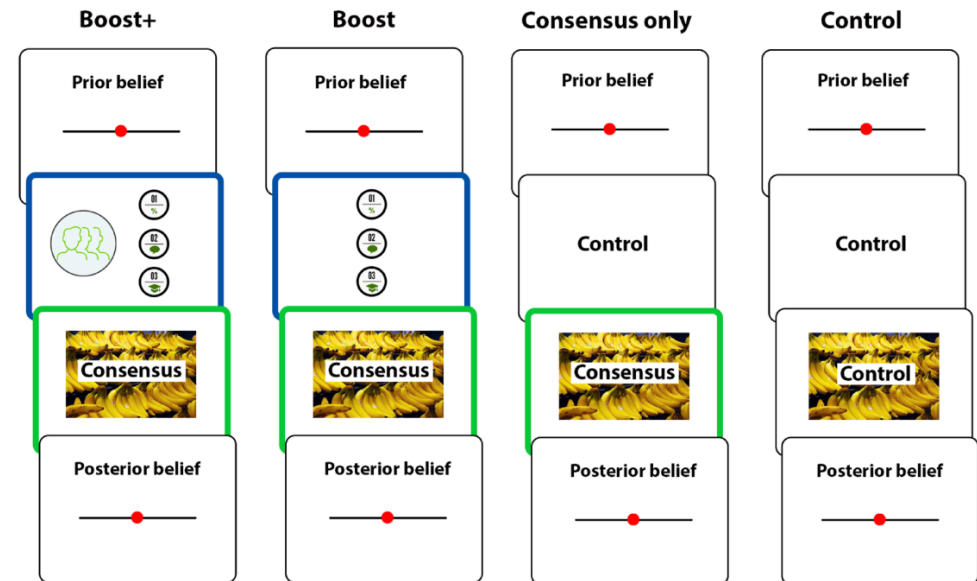


5. Boosting understanding of scientific consensus to correct false beliefs

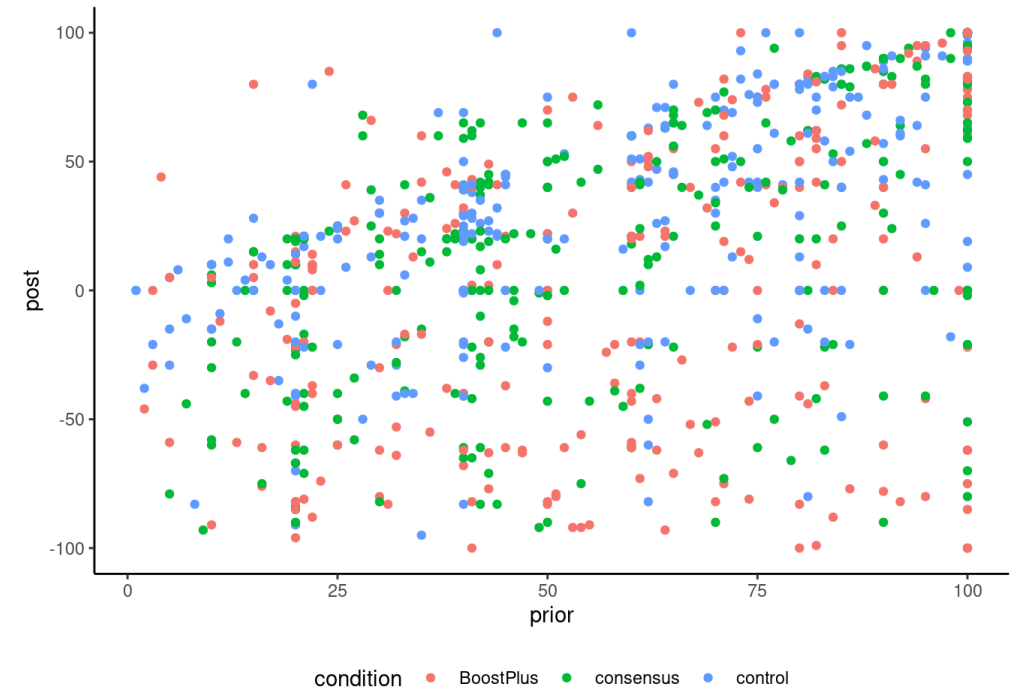
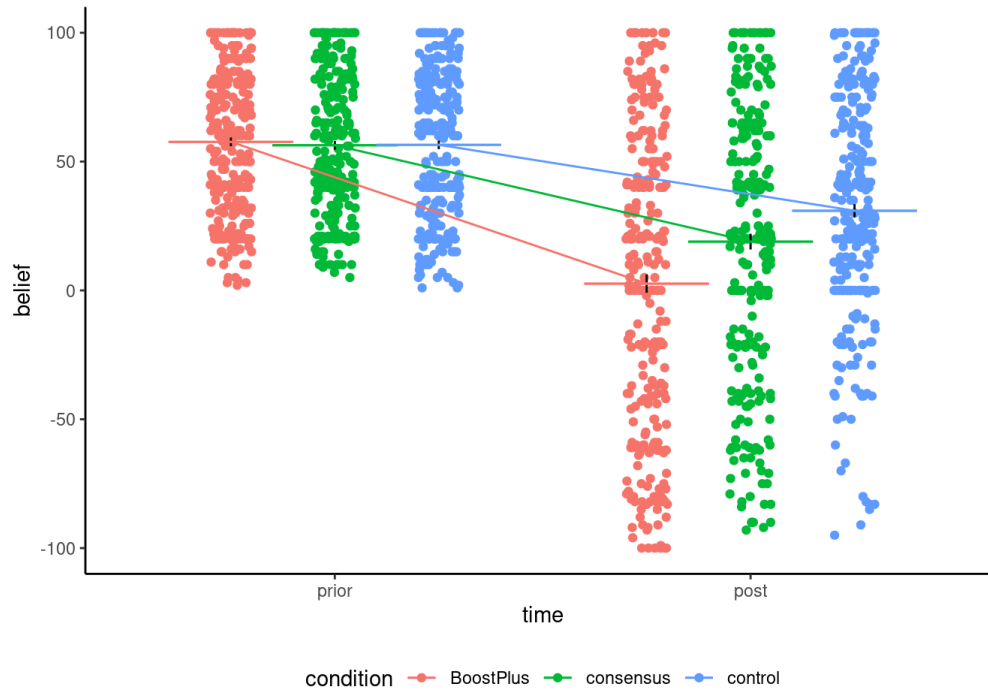
Stekelenburg et al. (2021)

In three experiments with more than 1,500 U.S. adults who held false beliefs, participants first learned the value of scientific consensus and how to identify it. Subsequently, they read a news article with information about a scientific consensus opposing their beliefs.

Model: (3 group between-subject ANCOVA)



Belief of harm of genetically engineered food



Preregistered, choice of topic (Study 1 on climate change produced no difference), second study underpowered, power calculation, scale of measurement and response, population heterogeneity

6. Social presence in social media products

Poirier et al. (2024)

Classification of Social Presence in Photos and Sample Photos (Study 1)




	Low social presence condition	Medium social presence condition	High social presence condition
Features	<ul style="list-style-type: none"> • No human present • No personal object • Limited warmth 	<ul style="list-style-type: none"> • No human present • Personal object • High warmth 	<ul style="list-style-type: none"> • Human present • Personal object • High warmth
Sample photos			

Fig. 1. Classification of Social Presence in Photos and Sample Photos (Study 1).

We conducted three experiments to explore how varying levels of perceived social presence (i.e., physical, implied, or absent human presence) in product photos affect consumer responses. The results indicate that social presence positively influences consumers' emotions and product photo diagnosticity, which, in turn, positively impacts purchase intentions.

Peer-review task

- Select papers that are suitable, but offer students a choice
 - level of difficulty
 - only methods and concepts covered in class
 - access to data!
- Narrow the choice if there are multiple experiments
- Ask them to focus only on statistical aspects

Data collection

- Discuss with students the reliability of data collected on Prolifics and Amazon M'Turk (external validity?)
- Randomization and stratification: the tools must relate to the software used for data collection (e.g., Qualtrics)

Choose your battles

- Methodology course, students must learn programming on the fly (via assignments and based on examples, I provide recordings with notebooks).
- I allow for both R and SPSS, but this complicates in-class presentation and management.

Maximally uninteresting

Often, results section formatting is terse and procedural. Students find them intimidating (akin to an unknown foreign language).

Students must nevertheless learn to criticize statistical methodology and spot potential problems.

Supporting our proposition that virtual pairs narrow their visual focus to their shared environment (that is, the screen), virtual pairs spent significantly more time looking directly at their partner ($M_{\text{virtual}} = 91.4$ s, s.d. = 58.3, $M_{\text{in-person}} = 51.7$ s, s.d. = 52.2, linear mixed-effect regression, $n = 270$ participants, $b = 39.70$, s.e. = 6.83, $t_{139} = 5.81$, $P < 0.001$, Cohen's $d = 0.71$, 95% CI = 0.47–0.96), spent significantly less time looking at the surrounding room ($M_{\text{virtual}} = 32.4$ s, s.d. = 34.8, $M_{\text{in-person}} = 61.0$ s, s.d. = 43.1, linear mixed-effect regression, $n = 270$ participants, $b = 28.75$, s.e. = 5.10, $t_{143} = 5.64$, $P < 0.001$, Cohen's $d = 0.74$, 95% CI = 0.49–0.99; Fig. 2) and remembered significantly fewer unexpected props in the surrounding room ($M_{\text{virtual}} = 1.53$, s.d. = 1.38, $M_{\text{in-person}} = 1.95$, s.d. = 1.38, Poisson mixed-effect regression, $n = 302$ participants, $b = 0.25$, s.e. = 0.10, $z = 2.47$, $P = 0.014$, Cohen's $d = 0.30$, 95% CI = 0.08–0.53) than in-person pairs. There was no evidence that the time spent looking at the task differed by modality ($M_{\text{virtual}} = 176.1$ s, s.d. = 64.0, $M_{\text{in-person}} = 186.8$ s, s.d. = 60.2, linear mixed-effect regression, $n = 270$ participants, $b = 10.65$, s.e. = 7.60, $t_{268} = 1.40$, $P = 0.162$, Cohen's $d = 0.17$, 95% CI = -0.07–0.41; see Supplementary Information C for model assumption tests).

Spot the problems

A single factor (Levels of social presence: No vs. Medium vs. High) **within-subjects design** was used. In total, 335 participants recruited on Amazon Mechanical Turk (MTurk.com) completed an online survey.

A repeated measure ANOVA with social presence (no, medium, and high) as the independent variable and positive emotional reaction as the dependent variable revealed a significant main effect ($F(1, 240) = 19.115, p < 0.001$).

Linear regressions with random intercepts to account for repeated measures were performed. Results show that both positive emotional reaction ($t(2135) = 14.722; \beta = 0.308, p < 0.0001$ [...]

The good, the bad and the ugly

Degrees of freedom do not match with reported sample size or methods.

Digging into the data reveals an incomplete unbalanced design

- between 3 and 10 measurements per participant, mode 6.
- out of 335 participants, 94 did not see all three categories!

Analysis in SPSS kept only complete cases!

The degrees of freedom reveal that the author used ordinary linear regression (no accounting for repeated measures).

Weekly exercises

- Textbook or simulated data works as well, but lack motivation and often the messy features of real data.
- Difficulty of examples and questions: follow Goldilock's principle
 - they needs to be easy for first exposition,
 - but hard enough to relate to problem set.

Lessons learned the hard way

- Some students try to just 'get the answer right', without learning the method.
- There can be built up frustration, especially for people with low programming proficiency (steep learning curve)
- Despite examples, unequal use of literal reporting (Quarto) (lots of unformatted output and copy-paste).

Free to use

- Link to slides: <https://lbelzile.github.io/SSC2025-reproducibility/>
- Course webpage and textbook
- R package with documentation

References

- Brucks, Melanie S., and Jonathan Levav. 2022. "Virtual Communication Curbs Creative Idea Generation." *Nature* 605 (7908): 108–12.
<https://doi.org/10.1038/s41586-022-04643-y>.
- Duke, Kristen E., and On Amir. 2023. "The Importance of Selling Formats: When Integrating Purchase and Quantity Decisions Increases Sales." *Marketing Science* 42 (1): 87–109. <https://doi.org/10.1287/mksc.2022.1364>.
- Maglio, Sam J., and Evan Polman. 2014. "Spatial Orientation Shrinks and Expands Psychological Distance." *Psychological Science* 25 (7): 1345–52.
<https://doi.org/10.1177/0956797614530571>.
- Moon, Alice, and Eric M VanEpps. 2023. "Giving Suggestions: Using Quantity Requests to Increase Donations." *Journal of Consumer Research* 50 (1): 190–210.
<https://doi.org/10.1093/jcr/ucac047>.
- Poirier, Sara-Maude, Sarah Cosby, Sylvain Sénécal, Constantinos K. Coursaris, Marc Fredette, and Pierre-Majorique Léger. 2024. "The Impact of Social Presence Cues in Social Media Product Photos on Consumers' Purchase Intentions." *Journal of Business Research* 185: 114932.
<https://doi.org/10.1016/j.jbusres.2024.114932>.

Stekelenburg, Aart van, Gabi Schaap, Harm Veling, and Moniek Buijzen. 2021.

“Boosting Understanding and Identification of Scientific Consensus Can Help to Correct False Beliefs.” *Psychological Science* 32 (10): 1549–65.

<https://doi.org/10.1177/09567976211007788>.