Mini-tutorial

Towards Designing Unbiased Replication Studies in Information Visualization

Poorna Talkad Sukumar and Ron Metoyer
University of Notre Dame

What is experimenter bias?

What is experimenter bias?

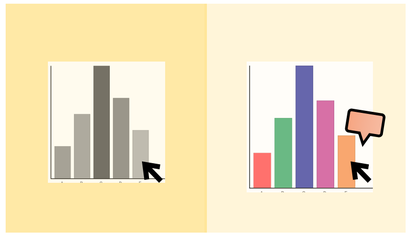
The tendency of researchers to conduct experiments in ways that bring about the expected outcome

What are some examples of experimenter bias?

What are some examples of experimenter bias?

Straw man comparisons or Win-lose setups

Outdated or weak baseline conditions are chosen

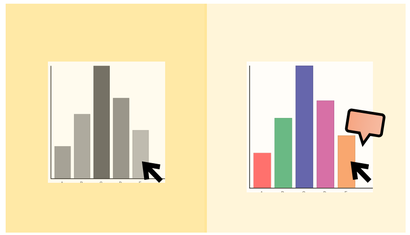


01

What are some examples of experimenter bias?

Straw man comparisons or Win-lose setups

Outdated or weak baseline conditions are chosen



01

Experimenter behaviors

Experimenter's body language and delivery of instructions influencing participant responses

02

What are some examples of experimenter bias?

Many more…

01	02	03	04	…
Sampling bias	Over-reliance on hypotheses	Not accounting for confounding factors	Tasks not representative of real-world activities	

Why is experimenter bias critical to replication studies?

Why is experimenter bias critical to replication studies?

What are the possible consequences?

Why is experimenter bias critical to replication studies?

What are the possible consequences?

Failed replications and replication crisis

01

Invalid replication and/or original results

02

Adversely affecting reputations of researchers and publication venues

03

What you will learn from this tutorial…

Improve

Some guidelines for avoiding biases when designing replication studies in InfoVis

01

Increase

How to conduct more replication studies with minimal effort,
e.g., by adding another condition to compare

02

How we arrived at the guidelines…

Paper collection

Searched and sampled 16 replication studies in InfoVis published at CHI and IEEE TVCG between 2008-2018

01

Coding

Studied the differences in experimental designs between the replication studies and studies they replicated

02

Table 1: An overview of the replication studies in information visualization sampled and used in our characterization.

Replication Study		Study being replicated			
Publication	Publication venue and year	Publication	Publication venue and year	Type of evaluation according to [35]	Characterization categories (Section 6) applicable to replication study
Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	CHI 2010	1. Graphical perception: Theory, experimentation, and application to the development of graphical methods [8] 2. Alpha, contrast and the perception of visual metadata [62]	1. J. Am. Statistical Assoc., 1984 2. Color Imaging Conf. 2009	User Performance	6.1, 6.2, 6.7
Perceptual Guidelines for Creating Rectangular Treemaps [32]	IEEE TVCG 2010	Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	CHI 2010	User Performance	6.2
The Impact of Social Information on Visual Judgments [26]	CHI 2011	1. Graphical perception: Theory, experimentation, and application to the development of graphical methods [8] 2. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	1. J. Am. Statistical Assoc., 1984 2. CHI 2010	User Performance	6.1, 6.4
Assessing the effect of visualizations on bayesian reasoning through crowdsourcing [42]	IEEE TVCG 2012	1. How to improve Bayesian reasoning without instruction: frequency formats [16] 2. Pictorial representations in statistical reasoning [4] and others	1. Psychological Review 1995 2. Applied Cognitive Psychology 2009	User Performance	6.1, 6.4, 6.5, 6.7
How Visualization Layout Relates to Locus of Control and Other Personality Factors [72]	IEEE TVCG 2012	1. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction [17] 2. Using personality factors to predict interface learning performance [18]	1. IEEE VAST 2010 2. HICSS 2010	User Performance, User Experience	6.1, 6.5, 6.7
Does an eye tracker tell the truth about visualizations?: findings while investigating visualizations for decision making [31]	IEEE TVCG 2012	A comparative study of three sorting techniques in performing cognitive tasks on a tabular representation [27]	IJHCI 2013	User Performance	6.7
Influencing Visual Judgment through Affective Priming [21]	CHI 2013	1. Graphical perception: Theory, experimentation, and application to the development of graphical methods [8] 2. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	1. J. Am. Statistical Assoc., 1984 2. CHI 2010	User Performance	6.1, 6.4, 6.6
Interactive visualizations on large and small displays: The interrelation of display size, information space, and scale [28]	IEEE TVCG 2013	Sizing up visualizations: effects of display size in focus+context, overview+detail, and zooming interfaces [55]	CHI 2011	User Performance, User Experience	6.1, 6.7
Ranking Visualizations of Correlation Using Webers Law [22]	IEEE TVCG 2014	The perception of correlation in scatterplots [53]	Computer Graphics Forum 2010	User Performance	6.1, 6.3, 6.7
Four Experiments on the Perception of Bar Charts [65]	IEEE TVCG 2014	Graphical perception: Theory, experimentation, and application to the development of graphical methods [8]	J. Am. Statistical Assoc., 1984	User Performance	6.1, 6.3, 6.7
Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability [47]	IEEE TVCG 2016	1. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing [42] and others	1. IEEE TVCG 2012	User Performance	6.1, 6.4, 6.5, 6.7
HindSight: encouraging exploration through direct encoding of personal interaction history [13]	IEEE TVCG 2017	Storytelling in information visualizations: Does it engage users to explore data? [3]	CHI 2015	User Performance, User Experience	6.1, 6.4
The attraction effect in information visualization [10]	IEEE TVCG 2017	1. Between a rock and a hard place: The failure of the attraction effect among unattractive alternatives [39] 2. Distinguishing among models of contextually induced preference reversals [69]	1. Journal of Consumer Psychology 2013 2. Journal of Experimental Psychology: Learning, Memory, and Cognition 1991	User Performance	6.1, 6.4, 6.5, 6.7
Correlation Judgment and Visualization Features: A Comparative Study [71]	IEEE TVCG 2018	1. The perception of correlation in scatterplots [53] 2. Ranking visualizations of correlation using Webers law [22]	1. Computer Graphics Forum 2010 2. IEEE TVCG 2014	User Performance	6.1, 6.3, 6.8
Blinded with Science or Informed by Charts? A Replication Study [11]	IEEE TVCG 2018	Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy [64]	Public Understanding of Science 2016	User Performance	6.1, 6.3, 6.7, 6.8
Modeling Color Difference for Visualization Design [63]	IEEE TVCG 2018	Enabling designers to foresee which colors users cannot see [52] and others	CHI 2016	User Performance	6.1, 6.5

Details in our BELIV paper

https://osf.io/q38pg

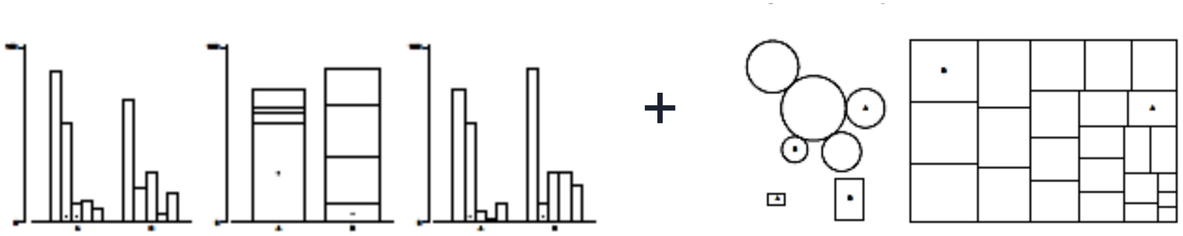
Improve

Guidelines for specific replication types

Replication type | Different Conditions

Replication type | Different Conditions

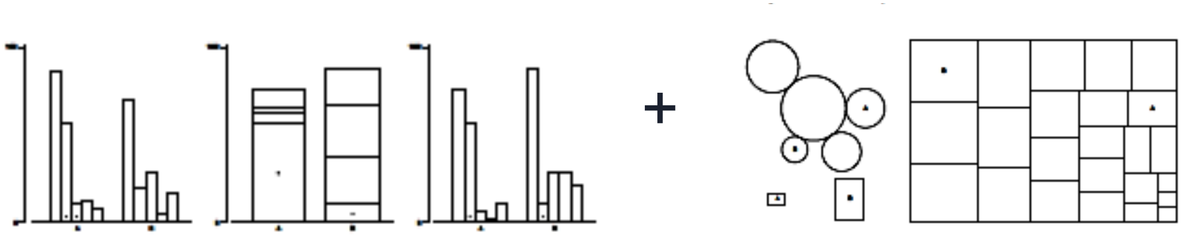
Strict replication + with different conditions



e.g., Heer and Bostock's graphical perception study, CHI 2010

Replication type | Different Conditions

Strict replication + with different conditions

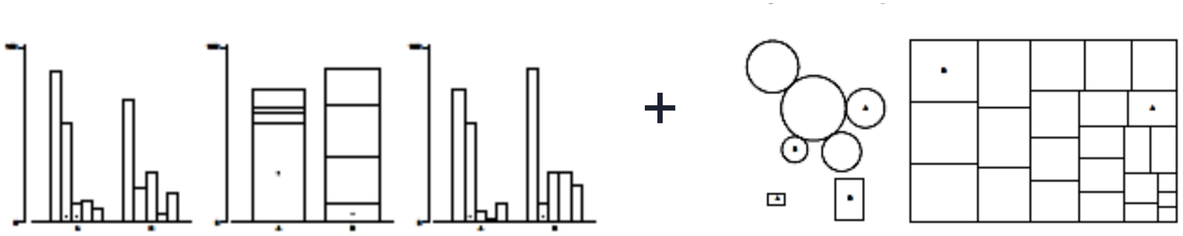


e.g., Heer and Bostock's graphical perception study, CHI 2010

GUIDELINES > Replicate original study closely before extending to new conditions to facilitate comparison of the study designs

Replication type | Different Conditions

Strict replication + with different conditions



e.g., Heer and Bostock's graphical perception study, CHI 2010

GUIDELINES

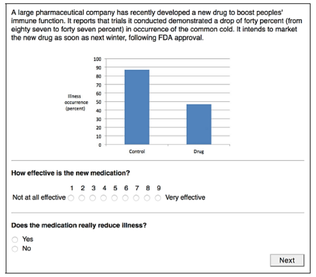
> Replicate original study closely before extending to new conditions to facilitate comparison of the study designs

> Ensure the new conditions chosen can be meaningfully validated using the same approach when extending prior findings

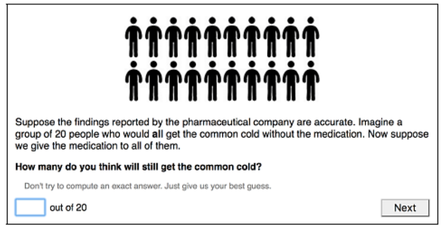
Replication type | Introspection

Replication type | Introspection

Further investigate prior findings + either by adding additional factors to study or by studying factors in isolation



+

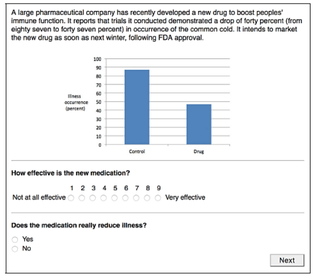


additional comprehension test

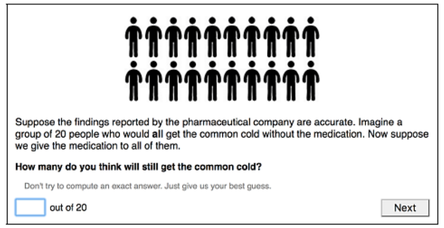
e.g., Dragicevic and Jansen's "Blinded with Science" study, IEEE TVCG 2017

Replication type | Introspection

Further investigate prior findings + either by adding additional factors to study or by studying factors in isolation



+



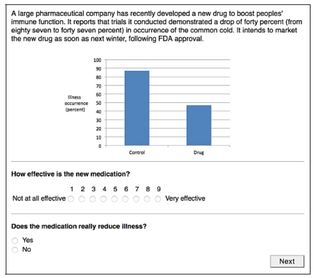
additional comprehension test

e.g., Dragicevic and Jansen's "Blinded with Science" study, IEEE TVCG 2017

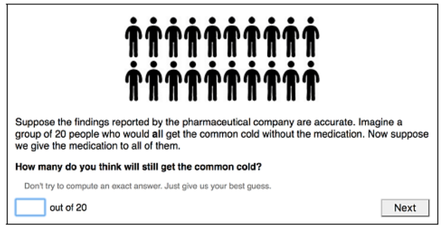
GUIDELINES > Apply high-level (as opposed to well-defined) hypotheses

Replication type | Introspection

Further investigate prior findings + either by adding additional factors to study or by studying factors in isolation



+



additional comprehension test

e.g., Dragicevic and Jansen's "Blinded with Science" study, IEEE TVCG 2017

GUIDELINES

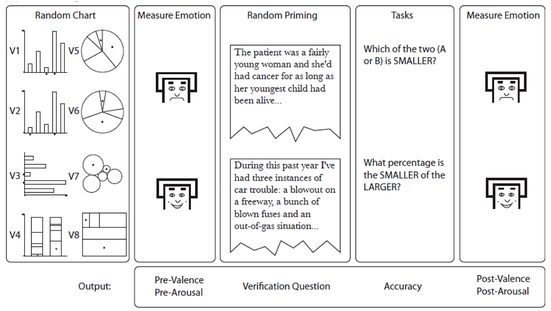
> Apply high-level (as opposed to well-defined) hypotheses

> Ensure that the specified factors of interest are adequately justified and their exploration in the context of the earlier study is meaningful and unbiased

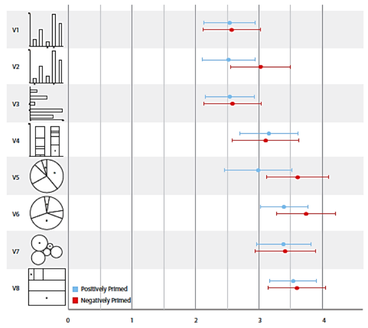
Replication type	Replicating Conditions

Replication type | Replicating Conditions

Reuse earlier study's conditions (and associated tasks) in a different context + with results *somewhat* comparable to earlier study



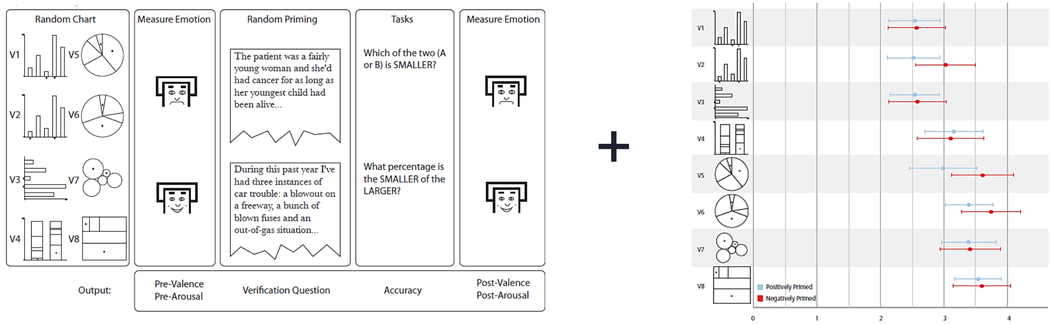
+



e.g., Harrison et al.'s "influence of affective priming on visual judgments" study, CHI 2013

Replication type | Replicating Conditions

Reuse earlier study's conditions (and associated tasks) in a different context + with results *somewhat* comparable to earlier study



e.g., Harrison et al.'s "influence of affective priming on visual judgments" study, CHI 2013

GUIDELINE > Ensure that the replicated conditions (and tasks) provide a meaningful context when embedding newly-proposed ideas

Replication type | Conceptual replications

Replication type | Conceptual replications

High-level comparison with independent study designs

e.g., *"Our study failed to replicate previous findings in that subjects' accuracy was remarkably lower and visualizations exhibited no measurable benefit [in facilitating Bayesian reasoning]"*

— *Micallef et al.*, IEEE TVCG 2012

Replication type | Conceptual replications

High-level comparison with independent study designs

e.g., *"Our study failed to replicate previous findings in that subjects' accuracy was remarkably lower and visualizations exhibited no measurable benefit [in facilitating Bayesian reasoning]"*

— *Micallef et al.*, IEEE TVCG 2012

GUIDELINE > Consider general experimenter biases in conceptual replications

Increase

Conducting more replication studies

Replicate instead of simply rebuilding!

Earlier study

vs.

Vis1
(control)

Vis2

01

Replicate instead of simply rebuilding!

Earlier study

Vis1 (control) vs. Vis2

01

New study

Vis2 (control) vs. Vis3 vs. Vis1

02

Replicate instead of simply rebuilding!

Earlier study

Vis1 (control) vs. Vis2

01

New + Replication study

Vis2 (control) vs. Vis3 vs. Vis1

02

Improve

General Guidelines

Crowdsourcing

14 out of the 16 replication papers collected used crowdsourcing

Crowdsourcing can help in mitigating biases due to

Sampling (larger and more diverse populations)	Experimenter behavior (no experimenter-participant interactions)	Biases introduced by experiment settings
01	02	03

GUIDELINE > Consider crowdsourcing for replication studies

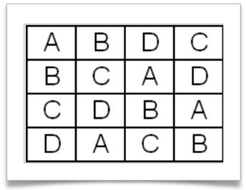
Sampling

Larger samples were generally used in the replication studies (compared to the original studies), especially those using crowdsourcing

Failures to replicate and replication crises often attributed to smaller sample-sizes in replication studies

GUIDELINE > Use larger samples in replication studies

Within-subjects/between-subjects



A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

We found 2 instances where a within-subjects study was changed to between-subjects in the replication

Certain types of studies mandate the use of either within-subjects or between-subjects design, e.g., Harrison et al.'s affective priming study, CHI 2013

GUIDELINES

> Provide justification when changing a within-subjects study to between subjects and vice versa.

> Ensure the study is adequately powered when changing a within-subjects study to between subjects.

Disparity in study descriptions

Some papers reported they had difficulty replicating certain aspects due to the lack of details in the original papers

There were also instances where more details were included in the original studies

GUIDELINE

> Include detailed study-design descriptions in all publications, replications or otherwise, to enable more faithful replications and for reviewers to ascertain that valid, unbiased experimental designs were employed

TOPICS COVERED

Replication types:

- Different Conditions
- Introspection
- Replicating Conditions
- Conceptual Replication

General:

- Crowdsourcing
- Sampling
- Within-subjects and between-subjects
- Disparity in study descriptions

Thank you

UNIVERSITY OF NOTRE DAME

Links to the paper and presentation can be found here:
http://sites.nd.edu/poorna-talkadsukumar/

Poorna Talkad Sukumar
ptalkads@nd.edu

Ronald Metoyer
rmetoyer@nd.edu