Increasing the Transparency of Research Papers with Explorable Multiverse Analyses

Pierre Dragicevic

Inria

Yvonne Jansen

Sorbonne Université

Abhraneel Sarma

University of Michigan

Matthew Kay

University of Michigan

Fanny Chevalier

University of Toronto

MESSAGE

Promote and *support*
transparent statistical reporting

More trustworthy

More interpretable

Easier to verify

Easier to replicate

MESSAGE

Promote and *support* transparent statistical reporting

More trustworthy

More interpretable

Easier to verify

Easier to replicate

RESULTS

Like the original paper we use an estimation approach, meaning that we report and interpret all results based on (unstandardized) effect sizes and their interval estimates [4]. We explain how to translate the results into statistical significance language to provide a point of reference, but we warn the reader against the pitfalls of dichotomous interpretations [5].

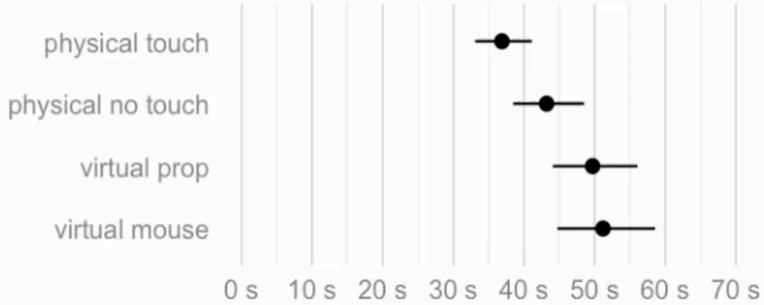


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the population mean 95% of the time across replications, under some assumptions [8]. In practice, if we assume we have very little prior knowledge about population means, each interval can be informally interpreted as a range of plausible values for the population mean, with the midpoint being about 7 times more likely than the endpoints [9].

Figure 3 shows the (geometric) mean completion time for each condition. At first sight, *physical touch* appears to be faster than the other conditions. However, since condition is a within-subject factor, it is preferable to examine within-subject differences [9], shown in Figure 4.

AP Photo/Wilfredo Lee

Do hurricanes with more feminine names cause more deaths?

Female hurricanes are deadlier than male hurricanes

Kiju Jung[a,1], Sharon Shavitt[a,b,1], Madhu Viswanathan[a,c], and Joseph M. Hilbe[d]

[a]Department of Business Administration and [b]Department of Psychology, Institute of Communications Research, and Survey Research Laboratory, and [c]Women and Gender in Global Perspectives, University of Illinois at Urbana–Champaign, Champaign, IL 61820; and [d]Department of Statistics, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, AZ 85287-3701

Edited* by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 14, 2014 (received for review February 13, 2014)

Do people judge hurricane risks in the context of gender-based expectations? We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes. Laboratory experiments indicate that this is because hurricane names lead to gender-based expectations about severity and this, in turn, guides respondents' preparedness to take protective action. This finding indicates an unfortunate and unintended consequence of the gendered naming of hurricanes, with important implications for policymakers, media practitioners, and the general public concerning hurricane communication and preparedness.

gender stereotypes | implicit bias | risk perception | natural hazard communication | bounded rationality

Estimates suggest that hurricanes kill more than 200 people in the United States annually, and severe hurricanes can cause fatalities in the thousands (1). As the global climate changes, the frequency and severity of such storms is expected to increase (2). However, motivating hurricane preparedness remains a major

violence and destruction (23, 24). We extend these findings to hypothesize that the anticipated severity of a hurricane with a masculine name (Victor) will be greater than that of a hurricane with a feminine name (Victoria). This expectation, in turn, will affect the protective actions that people take. As a result, a hurricane with a feminine vs. masculine name will lead to less protective action and more fatalities.

Archival Study

To test this hypothesis, we used archival data on actual fatalities caused by hurricanes in the United States (1950–2012). Ninety-four Atlantic hurricanes made landfall in the United States during this period (25). Nine independent coders who were blind to the hypothesis rated the masculinity vs. femininity of historical hurricane names on two items (1 = very masculine, 11 = very feminine, and 1 = very man-like, 11 = very woman-like), which were averaged to compute a masculinity-femininity index (MFI). A series of negative binomial regression analyses (26, 27) were performed to investigate effects of perceived masculinity-femininity of hurricane names (MFI), minimum pressure, normalized

Female hurricanes are deadlier than male hurricanes

Kiju Jung[a,1], Sharon Shavitt[a,b,1], Madhu Viswanathan[a,c], and Joseph M. Hilbe[d]

[a]Department of Business Administration and [b]Department of Psychology, Institute of Communications Research, and Survey Research Laboratory, and [c]Women and Gender in Global Perspectives, University of Illinois at Urbana–Champaign, Champaign, IL 61820; and [d]Department of Statistics, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, AZ 85287-3701

Edited* by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 14, 2014 (received for review February 13, 2014)

Do people judge hurricane risks in the context of gender-based expectations? We use more than six decades of death rates from [illegible] to show that severe hurricanes with feminine names [illegible] significantly more deaths than severe hurricanes with masculine names. Laboratory experiments indicate that this is because hurricane names lead to gender-based expectations about severity and this, in turn, guides respondents' preparedness to take protective action. This finding indicates an unfortunate and unintended consequence of the gendered naming of hurricanes, with important implications for policymakers, media practitioners, and the general public concerning hurricane communication and preparedness.

gender stereotypes | implicit bias | risk perception | natural hazard communication | bounded rationality

Estimates suggest that hurricanes kill more than 200 people in the United States annually, and severe hurricanes can cause fatalities in the thousands (1). As the global climate changes, the frequency and severity of such storms is expected to increase (2). However, motivating hurricane preparedness remains a major

violence and destruction (23, 24). We extend these findings to hypothesize that the anticipated severity of a hurricane with [illegible] will affect the protective actions that people take. As a result, a hurricane with a feminine vs. masculine name will lead to less [illegible] and thus more fatalities.

Archival Study

To test this hypothesis, we used archival data on actual fatalities caused by hurricanes in the United States (1950–2012). Ninety-four Atlantic hurricanes made landfall in the United States during this period (25). Nine independent coders who were blind to the hypothesis rated the masculinity vs. femininity of historical hurricane names on two items (1 = very masculine, 11 = very feminine, and 1 = very man-like, 11 = very woman-like), which were averaged to compute a masculinity-femininity index (MFI). A series of negative binomial regression analyses (26, 27) were performed to investigate effects of perceived masculinity-femininity of hurricane names (MFI), minimum pressure, normalized

"If you torture the data long enough, it will confess."

– Ronald Coase, 1988 (Nobel Laureate)

Female hurricanes are not deadlier than male hurricanes

Jung et al. (1) assert that hurricanes that made landfall in the United States killed more people when they had female names rather than male names. The article has stirred much controversy. Criticisms range from the inclusion of hurricanes from the era before they were given male names (2) lective interpretation and the ov their results from the archival of their hypothesis (3), to the e of their six behavioral experim populations in at-risk situation

The criticism of this lette one: the results of their archi function of the selective incl sors. Using the same data, m variables, I show in Table 1 are not robust to the inclu two-way interaction they o analysis. Model 1 reprodu main results. Models 2–4 s that female- and male-n were equally deadly canno the interaction effect of metric pressure and its toll is included. A more letter should have stated

Models 2–4 show th lower barometric pressu tolls and that hurricanes

Table 1. Results from

tolls had smaller death tolls when the hurricanes were strong (lower pressure), but higher death tolls when the hurricanes were weak (higher pressure). The latter result is driven by the pre-1978 sample (model 5). In the post-1978 sample, the interaction effect insignificant and the damage toll

safety infrastruct other characteri after 1978. Eve of safety infras and an over vergence of might expl the death

Weather and Climate Extremes 12 (2016) 80–84

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

Hurricane names: A bunch of hot air?

Gary Smith *

Department of Economics, Pomona College, United States

ARTICLE INFO

Article history:

LETTER

Are female hurricanes really deadlier than male hurricanes?

Jung et al. (1) claim to show that "feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes" (p. 1). This conclusion is mainly obtained by analyzing data on fatalities caused by hurricanes in the United States (1950–2012). By reanalyzing the same data, we show that the conclusion is based on biased presentation and invalid statistics.

The reasoning in ref. 1 is fundamentally based on the regression models reported in their table S2, in particular, model 4. However, due to the interaction terms combined with extreme values and weak significance, the analysis is based on a very fragile model; e.g., the model predicts almost 20,000 deaths for hurricane Sandy, which actually caused

Now, we explain our claim that the results are presented in a biased way. By holding the minimum pressure at its mean in prediction of counts of deaths, the authors only report the influence of MFI and normalized damage (figure 1 in ref. 1). This ignores the influence of the second interaction term MFI minimum pressure, which shows an opposite influence (see the estimated parameters on p. 5, first paragraph). By considering the counts of deaths under constant normalized damage, the results are contrary: male-named hurricanes with a low minimum pressure (strong hurricanes) are associated with more deaths than female ones (Fig. 1).

In the light of an alternating male-female differences between male- or female-named hurricanes for deaths, minimum pressure, category, and damages.

To conclude, the analyses given in ref. 1 are examples of the fact that prediction models using interaction terms have to be handled and interpreted carefully; in particular, using insignificant variables is not expedient and may lead to statistical artifacts.

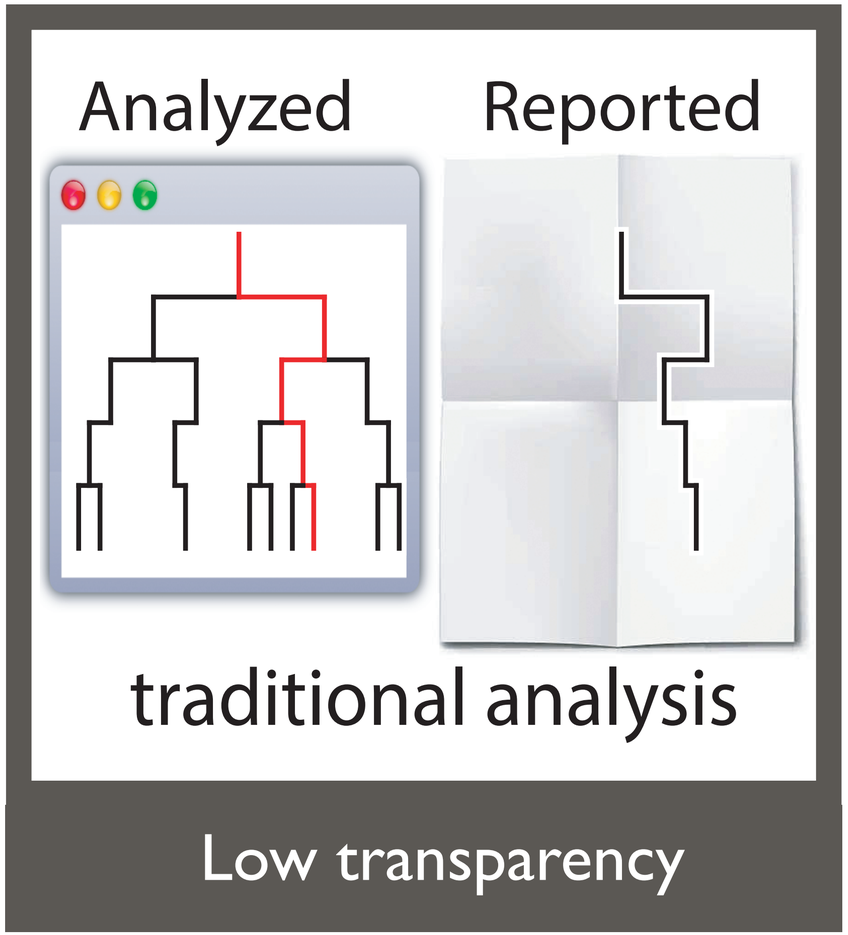
To summarize, the data do not contain evidence that feminine-named hurricanes cause more deaths than masculine-named hurricanes.

Björn Christensen [a] ***and Sören Christensen*** [b,1]

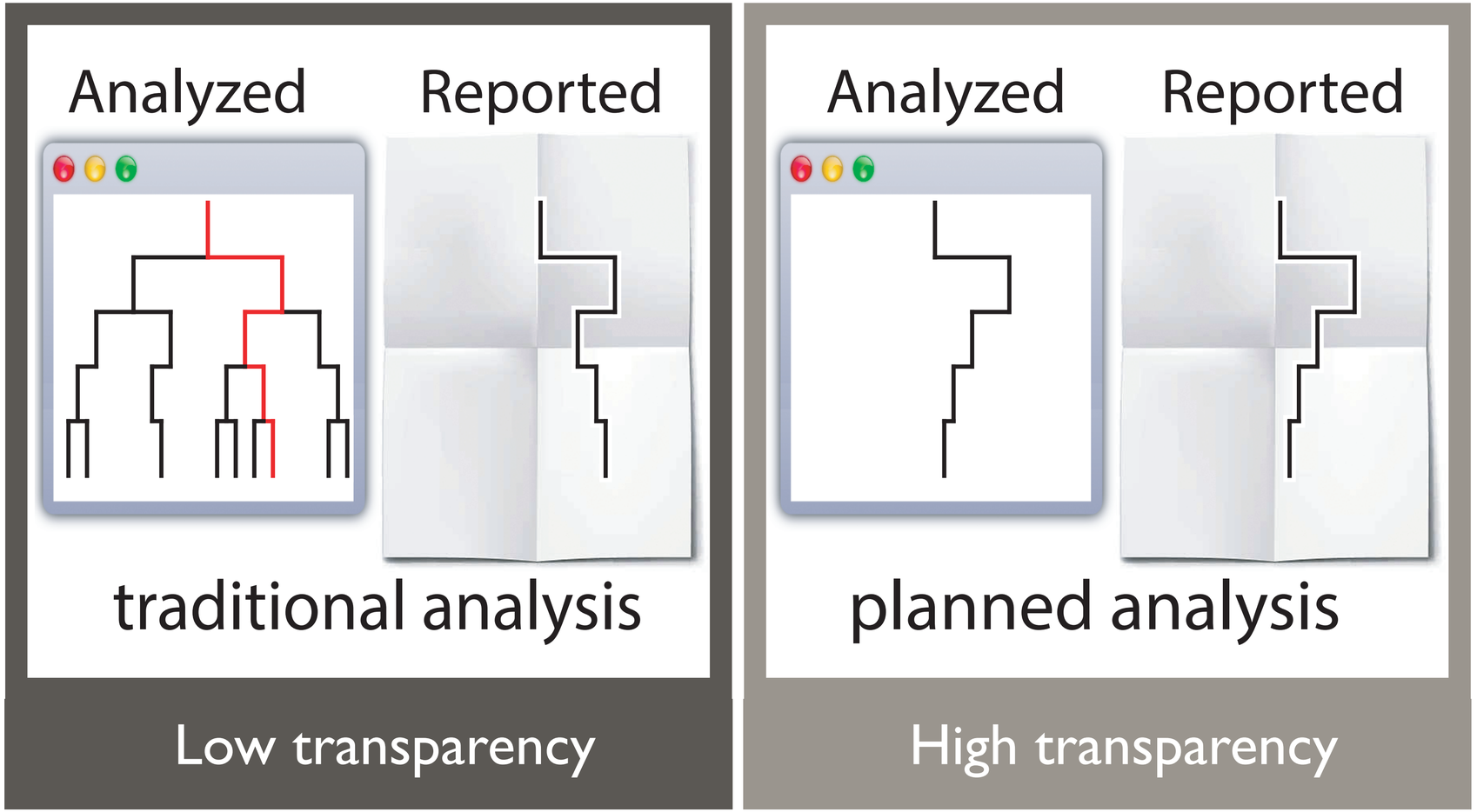
-named hurricanes are deadlier because people do not take them ser-
is based on a questionable statistical analysis of a narrowly defined data
not robust in that it is not confirmed by a straightforward analysis of
ata.

on (less than 39 mph), tropical storm (39–73 mph), hur-
nore than 73 mph), and major hurricane (more than
Tropical storms and hurricanes are generally given
Hurricane Sandy, but tropical depressions are not.
al. (2014) examine a narrowly defined dataset: U.S.
m Atlantic hurricanes that made landfall in the United
n a strong, surprising conclusion is drawn from re-
it can be instructive to see whether the conclusion is
espect to the myriad decisions used to restrict the
re are several issues:
clude tropical storms? In 1994 Tropical Storm Al-
lfall near Destin, Florida, with maximum sustained
h and caused historic flooding in Alabama and
lted in at least 30 deaths and caused $1 bil-
4 dollars). Alberto was classi
a hurricane, be
mph th

STATISTICAL ANALYSIS & REPORTING



STATISTICAL ANALYSIS & REPORTING



Do soccer referees give more red cards to dark-skinned players than light-skinned ones?

SAME DATA, MANY POSSIBLE ANALYSES

Pierre's plan

Skin colour:
3-levels: dark, medium, light

Co-variate use:
Control for positions (defensive players may commit more fouls)

Transformations:
Average skin-tone ratings (mean)

Exclusions:
Player-referee with no red card

Analytic approach:
Multilevel regression

Matt's plan

Skin colour:
5 levels: (very) dark, medium, (very) light

Co-variate use:
Control for referees' skin colour

Transformations:
Max (darkest) skin-tone rating

Exclusions:
Games with no red card

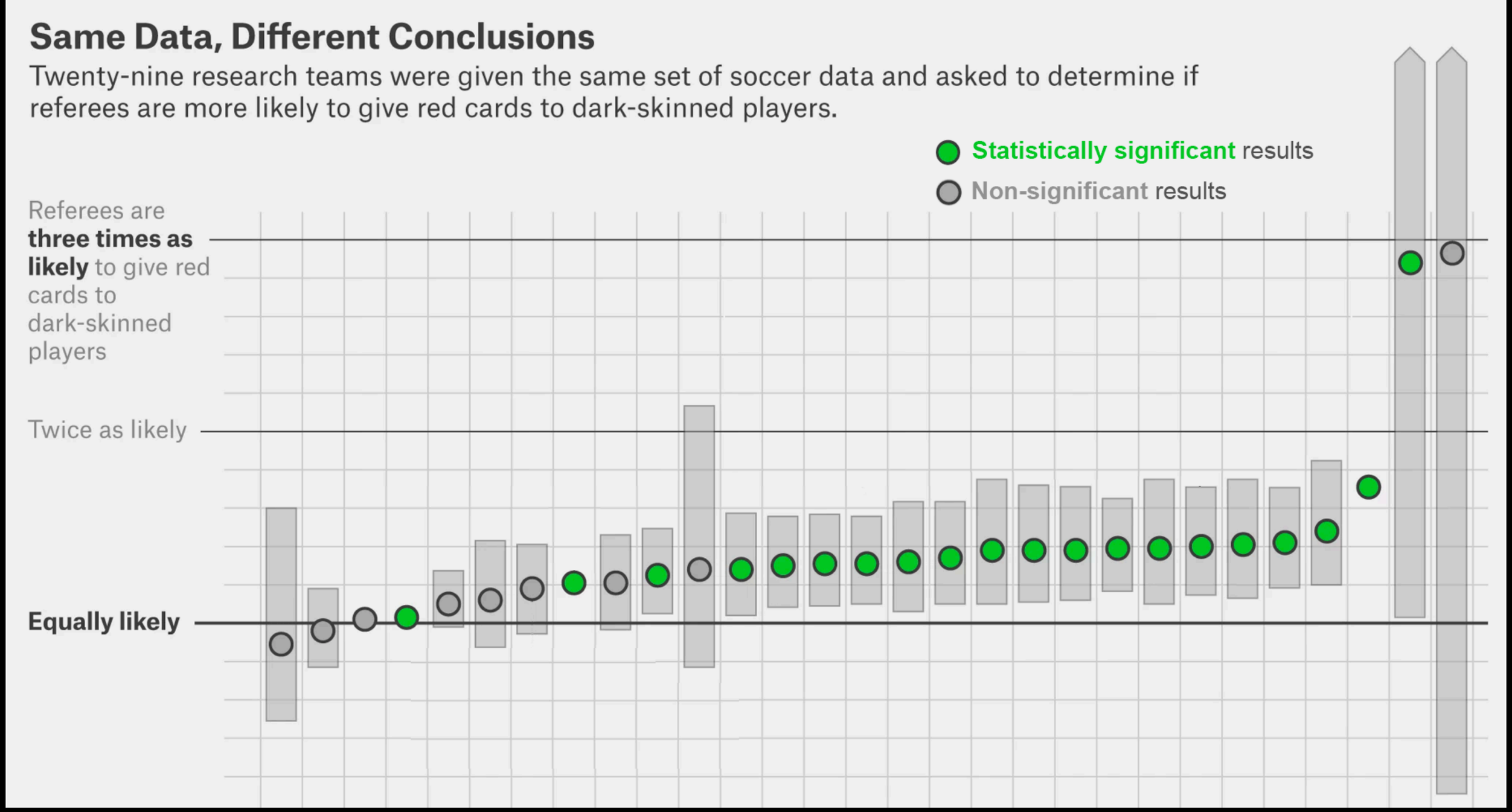
Analytic approach:
Bayesian logistic regression

Same Data, Different Conclusions

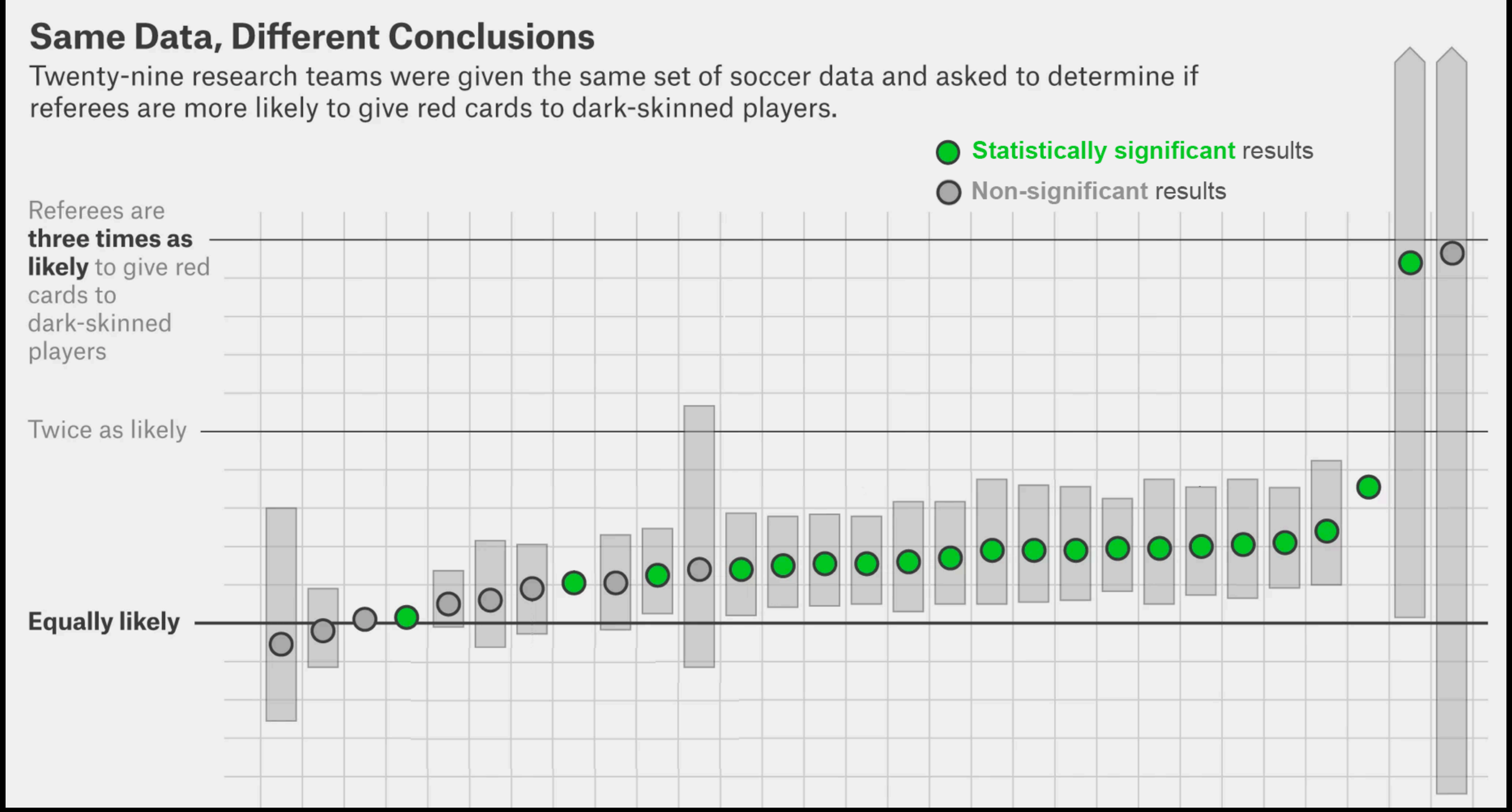
Referees are **three times as likely** to give red cards to dark-skinned players

Twice as likely

Equally likely

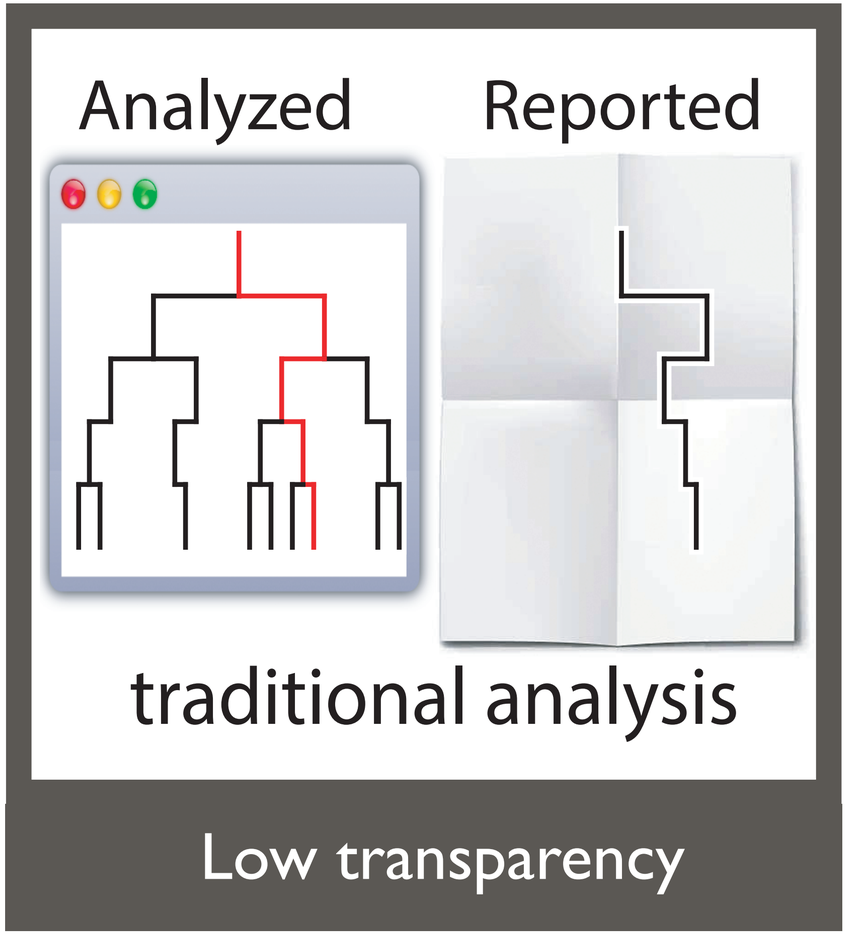


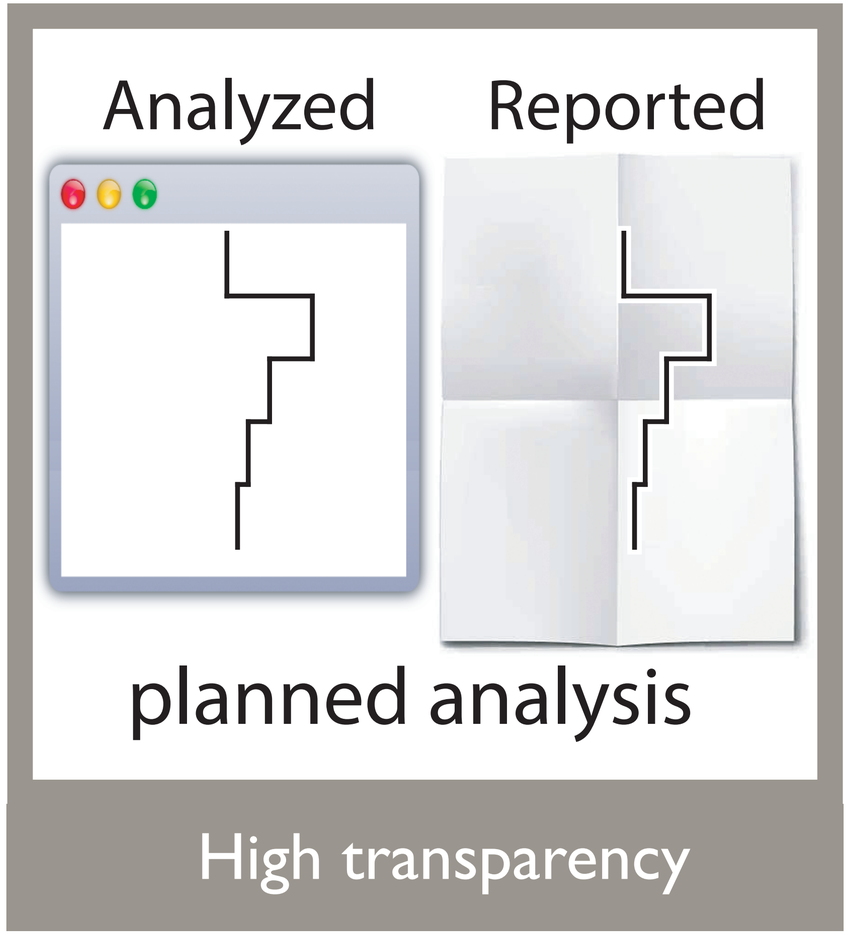
Brian A. Nosek et al. (2015) Many analysts, one dataset: Making transparent how variations in analytical choices affect results

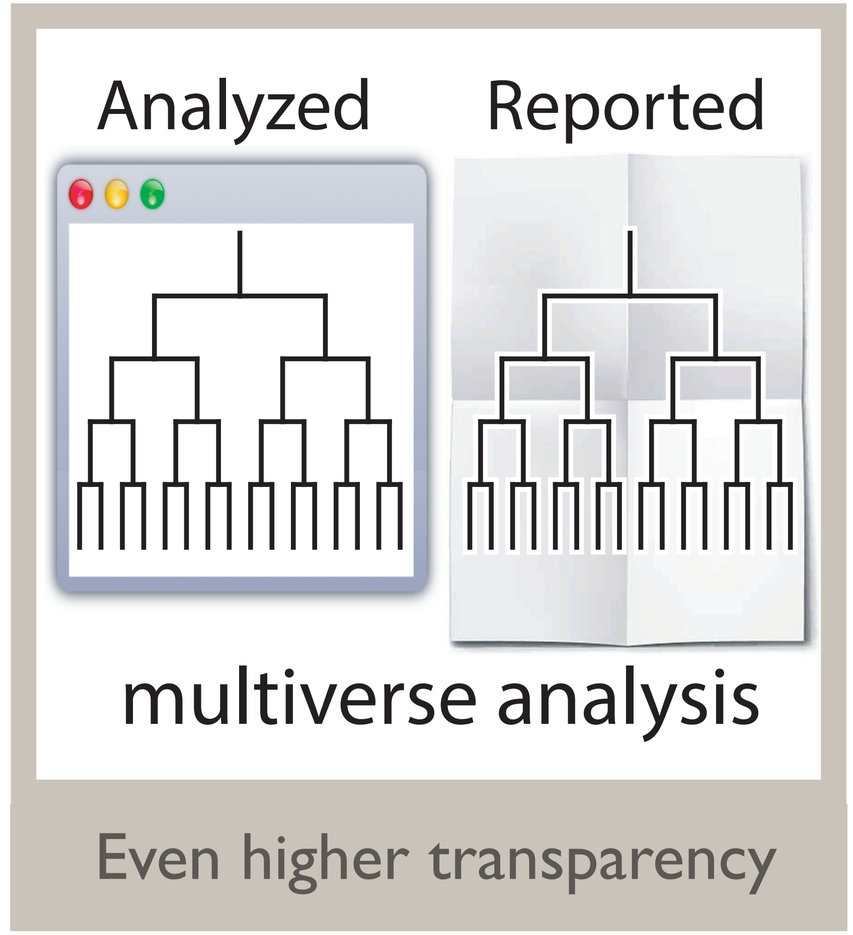


Brian A. Nosek et al. (2015) Many analysts, one dataset: Making transparent how variations in analytical choices affect results

STATISTICAL ANALYSIS & REPORTING

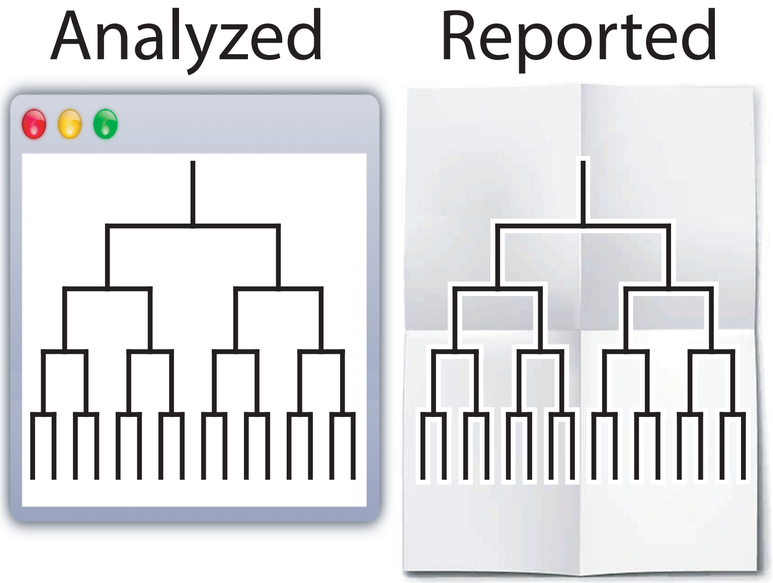






MULTIVERSE ANALYSIS

Performing and *reporting* all analyses
corresponding to a
large set of reasonable analysis scenarios.

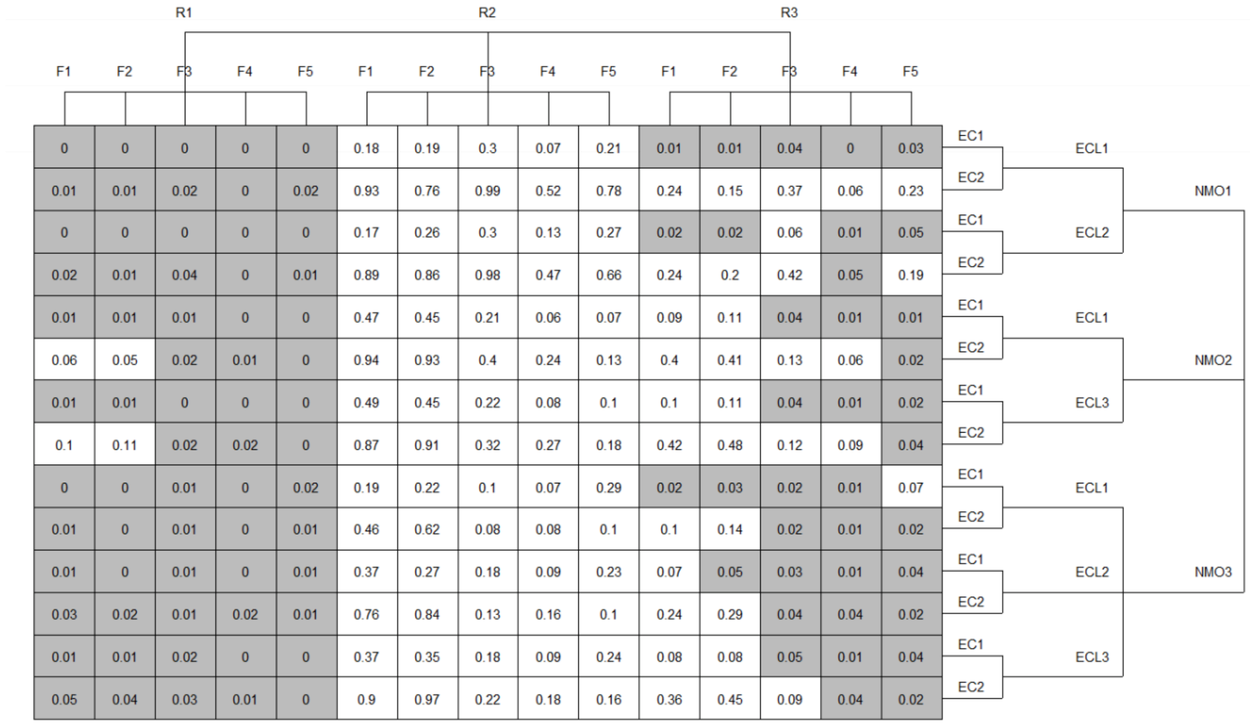


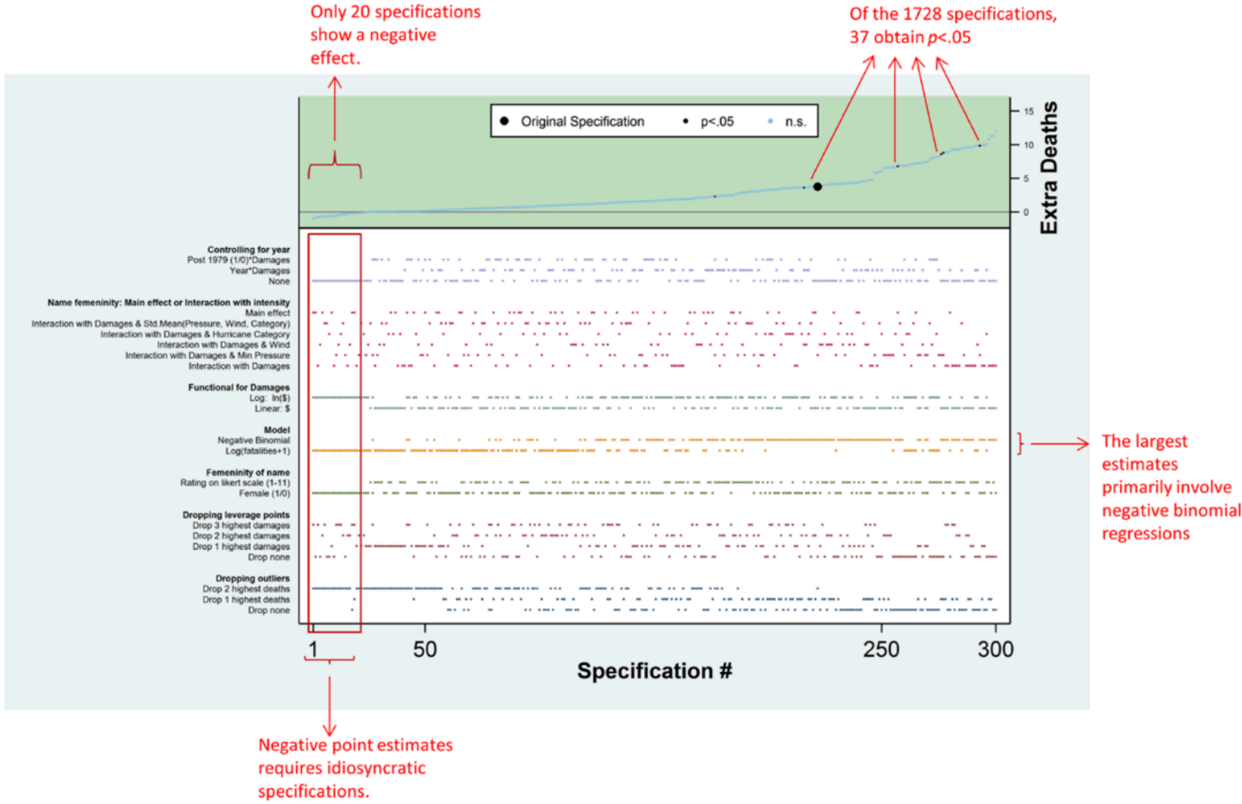
Steegen et al. (2016) Increasing Transparency Through a Multiverse Analysis

Simonsohn et al. (2015) Specification Curves: Descriptive and Inferential Statistics on All Reasonable Specifications

How to report all of these analyses?

VISUAL SUMMARIES





Steegen et al. (2016)
Increasing Transparency Through a Multiverse Analysis

Simonsohn et al. (2015) Specification Curves: Descriptive and Inferential Statistics on All Reasonable Specifications

Can we do better?

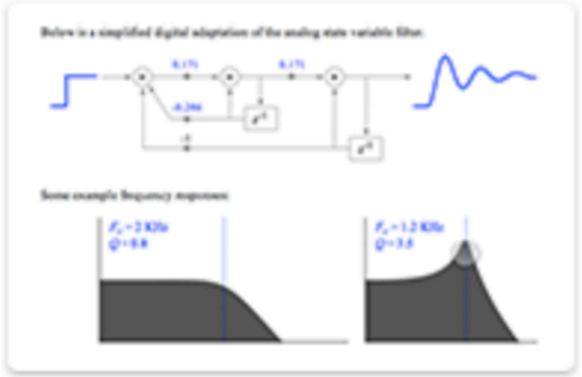
Explorable Explanations

Bret Victor / March 10, 2011

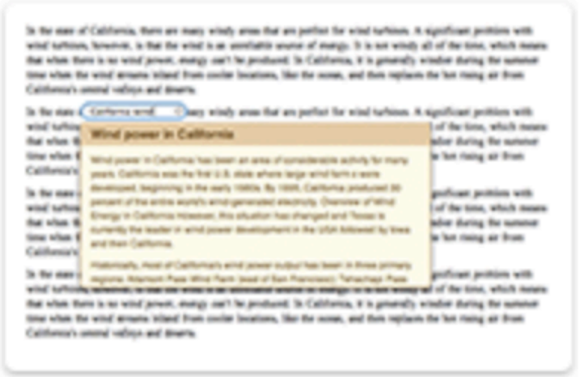
What does it mean to be an **active reader**?

A **reactive document** allows the reader to play with the author's assumptions and analyses, and see the consquences.

An **explorable example** makes the abstract concrete, and allows the reader to develop an intuition for how a system works.



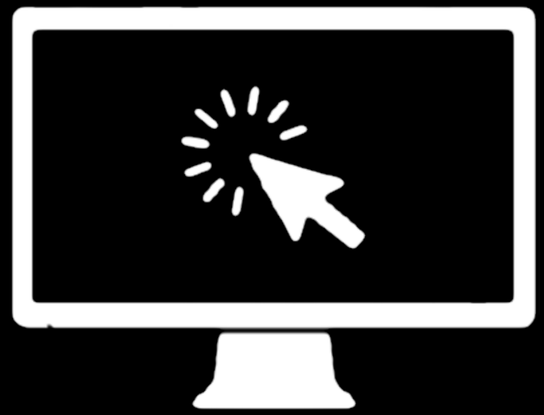
Contextual information allows the reader to learn related material just-in-time, and cross-check the author's claims.



http://worrydream.com/ExplorableExplanations/

EXPLORABLE MULTIVERSE ANALYSIS REPORT (**EMAR**)

https://explorablemultiverse.github.io/



MORE IN THE PAPER!

Increasing the Transparency of Research Papers with Explorable Multiverse Analyses

Pierre Dragicevic
Inria
Orsay, France
pierre.dragicevic@inria.fr

Yvonne Jansen
CNRS – Sorbonne Université
Paris, France
yvonne.jansen@sorbonne-universite.fr

Abhraneel Sarma
University of Michigan
Ann Arbor, MI, USA
abhsarma@umich.edu

Matthew Kay
University of Michigan
Ann Arbor, MI, USA
mjskay@umich.edu

Fanny Chevalier
University of Toronto
Toronto, Canada
fanny@cs.toronto.edu

ABSTRACT

We present *explorable multiverse analysis reports*, a new approach to statistical reporting where readers of research papers can explore alternative analysis options by interacting with the paper itself. This approach draws from two recent ideas: *i) multiverse analysis*, a philosophy of statistical reporting where paper authors report the outcomes of many different statistical analyses in order to show how fragile or robust their findings are; and *ii) explorable explanations*, narratives that can be read as normal explanations but where the reader can also become active by dynamically changing some elements of the explanation. Based on five examples and a design space analysis, we show how combining those two ideas can complement existing reporting approaches and constitute a step towards more transparent research papers.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Multiverse analysis, explorable explanation, statistics, transparent reporting, interactive documents.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300295

ACM Reference Format:
Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3290605.3300295

1 INTRODUCTION

The recent replication crisis in psychology and other disciplines has dealt a blow to the credibility of human-subject research and prompted a movement of methodological reform [70]. Much of this movement calls for more transparency in the way statistics are reported, so that findings become more trustworthy, more likely to be interpreted correctly, and easier to verify and replicate [29, 69, 72]. Concern for transparency in statistical reporting has spread to the HCI community, which has published several articles [24, 31, 58] and hosted several workshops [56, 57, 96] on the topic.

While much of the current discussions around transparent statistics in HCI focus on how the community can improve its practice, it has been suggested that HCI can do more than endorse and promote the transparent statistics movement—it can actively contribute to it by proposing novel user interfaces for better doing and better communicating statistics [31, 97]. In this article, we consider the research paper as a user interface, and seek to understand how we can enrich this user interface to better support and promote transparent statistics reporting.

While there are many ways a statistical report can lack transparency, a common and damaging form of opacity is *undisclosed flexibility* (see Figure 1a), i.e., not reporting the different options that have been tried during the analysis [85, 98], or the options that would have been chosen had the data been different [40]. Undisclosed flexibility is damaging because it substantially increases the chances of reporting erroneous findings, while being invisible to the reader.

Design space

Design considerations

Pitfalls

Future directions

MESSAGE

Promote and *support* transparent statistical reporting

More trustworthy

More interpretable

Easier to verify

Easier to replicate

RESULTS

Like the original paper we use an estimation approach, meaning that we report and interpret all results based on (unstandardized) effect sizes and their interval estimates [4]. We explain how to translate the results into statistical significance language to provide a point of reference, but we warn the reader against the pitfalls of dichotomous interpretations [5].

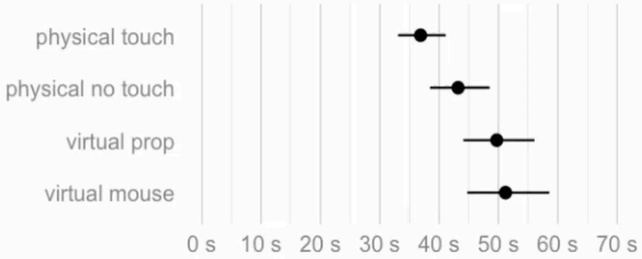


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the population mean 95% of the time across replications, under some assumptions [8]. In practice, if we assume we have very little prior knowledge about population means, each interval can be informally interpreted as a range of plausible values for the population mean, with the midpoint being about 7 times more likely than the endpoints [9].

Figure 3 shows the (geometric) mean completion time for each condition. At first sight, *physical touch* appears to be faster than the other conditions. However, since condition is a within-subject factor, it is preferable to examine within-subject differences [9], shown in Figure 4.

https://explorablemultiverse.github.io/

Explorable Multiverse Analyses

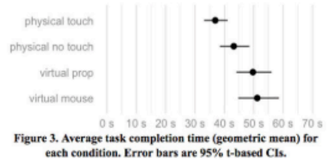
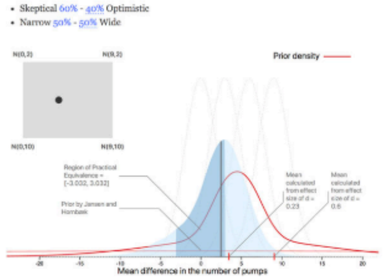


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the



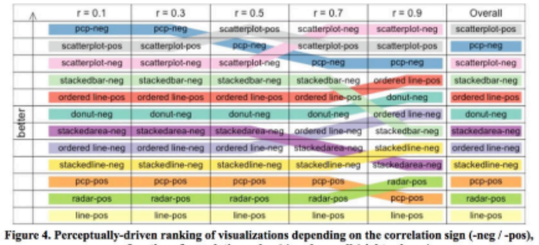
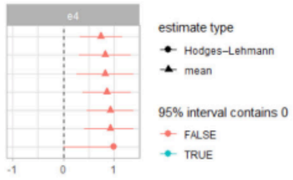


Figure 4. Perceptually-driven ranking of visualizations depending on the correlation sign (-neg / -pos), as a function of correlation value (r) and overall (right column).



Pierre Dragicevic (Inria), Yvonne Jansen (CNRS - Sorbonne Université), Abhraneel Sarma (University of Michigan)
Matthew Kay (University of Michigan), Fanny Chevalier (University of Toronto)

This is the companion website for the paper:

Dragicevic, Jansen, Sarma, Kay, and Chevalier. 2019. **Increasing the Transparency of Research Papers with Explorable Multiverse Analyses**. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA, 15 pages.

Useful tips before you try the examples below:

- You can animate any paper by **holding the A key**. Wait for all images to pre-load.
- Prefer the **Chrome browser**. You may experience layout issues with other browsers.