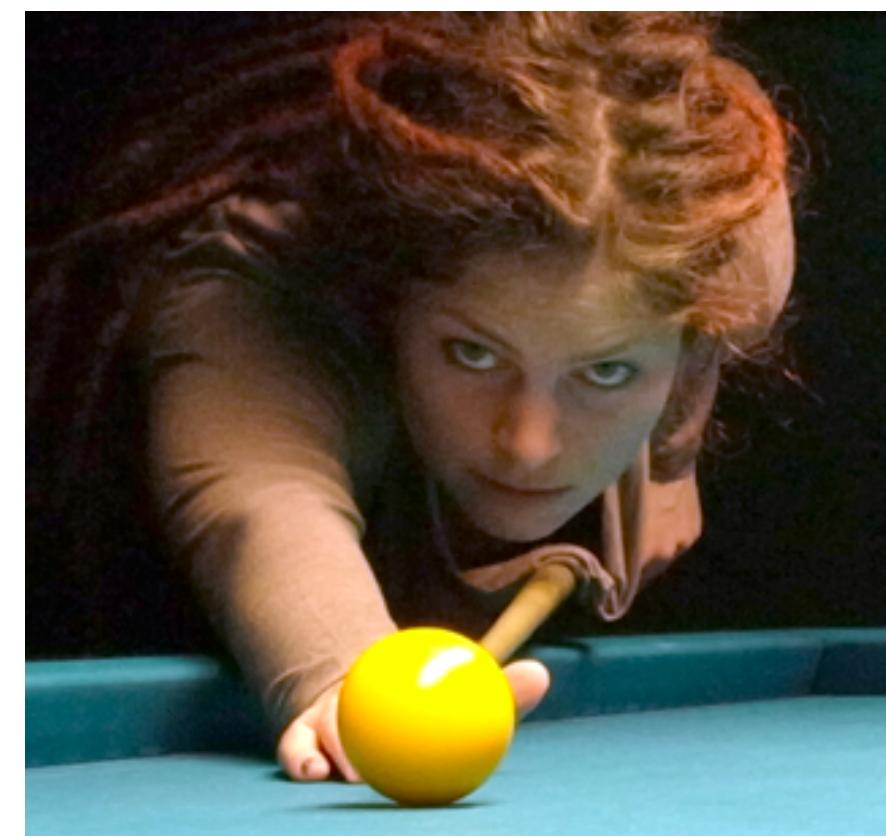


# Increasing the Transparency of Research Papers with Explorable Multiverse Analyses



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

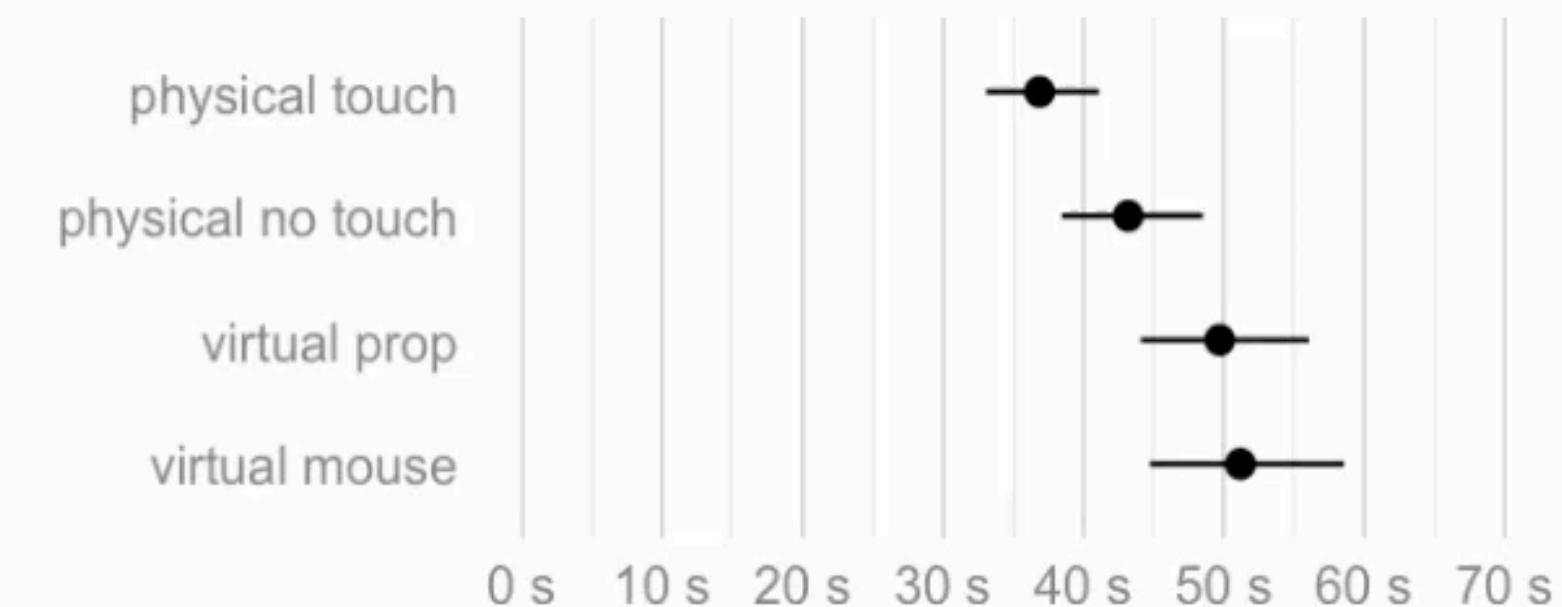
More interpretable

Easier to verify

Easier to replicate

## RESULTS

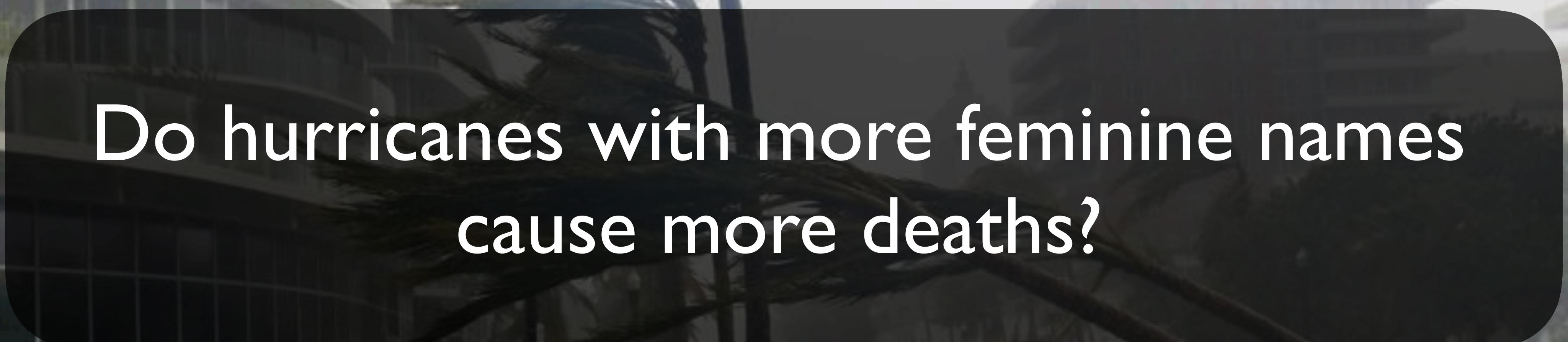
Like the original paper we use an estimation approach, meaning that we report and interpret all results based on (unstandardized) effect sizes and their interval estimates [4]. We explain how to translate the results into statistical significance language to provide a point of reference, but we warn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



Do hurricanes with more feminine names cause more deaths?

# Female hurricanes are deadlier than male hurricanes

Kiju Jung<sup>a,1</sup>, Sharon Shavitt<sup>a,b,1</sup>, Madhu Viswanathan<sup>a,c</sup>, and Joseph M. Hilbe<sup>d</sup>

<sup>a</sup>Department of Business Administration and <sup>b</sup>Department of Psychology, Institute of Communications Research, and Survey Research Laboratory, and <sup>c</sup>Women and Gender in Global Perspectives, University of Illinois at Urbana-Champaign, Champaign, IL 61820; and <sup>d</sup>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

**E**stimates 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<sup>a,1</sup>, Sharon Shavitt<sup>a,b,1</sup>, Madhu Viswanathan<sup>a,c</sup>, and Joseph M. Hilbe<sup>d</sup>

<sup>a</sup>Department of Business Administration and <sup>b</sup>Department of Psychology, Institute of Communications Research, and Survey Research Laboratory, and <sup>c</sup>Women and Gender in Global Perspectives, University of Illinois at Urbana-Champaign, Champaign, IL 61820; and <sup>d</sup>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

hurricanes to test this hypothesis. 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 unfortunate 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 aggression (23, 24). We extend this finding to hypothesize that the anticipated severity of a hurricane will

depend on its name. In other words, the perceived masculinity of a hurricane will affect the protective actions that people take. As a result, a hurricane with a feminine vs. masculine name will lead to less

— Ronald Coase, 1988 (*Nobel Laureate*)

## Abstract

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 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) to selective interpretation and the overstatement of their results from the archival study of their hypothesis (3), to the exclusion of their six behavioral experiments with populations in at-risk situations. In view of this letter,

The criticism of this 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 incl two-way interaction they or analysis. Model 1 reprodu main results. Models 2-4 s that female- and male-n were equally deadly cannot the interaction effect of a metric pressure and its r toll is included. A more a letter should have stated

Models 2-4 show lower barometric pressure tolls and that hurricanes

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 is insignificant and the damage toll relationship



# LETTER

## Are female hurricanes real male hurricanes?

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.



ELSEVIE

*Weather and Climate Extremes* 12 (2016) 80–84  
Contents lists available at ScienceDirect

Contents lists available at ScienceDirect  
Extremes 12 (2016) 80

er and cu  
es available at [ScienceDirect](#)

# Weather and Climate Extremes

*Journal homepage:* [www.elsevier.com/locate/watex](http://www.elsevier.com/locate/watex)

# Hurricane names: A bunch of hot air?

Gary Smit

*Department of Economics, Pomona College, United States*

## ARTICLE INFO



1

# 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

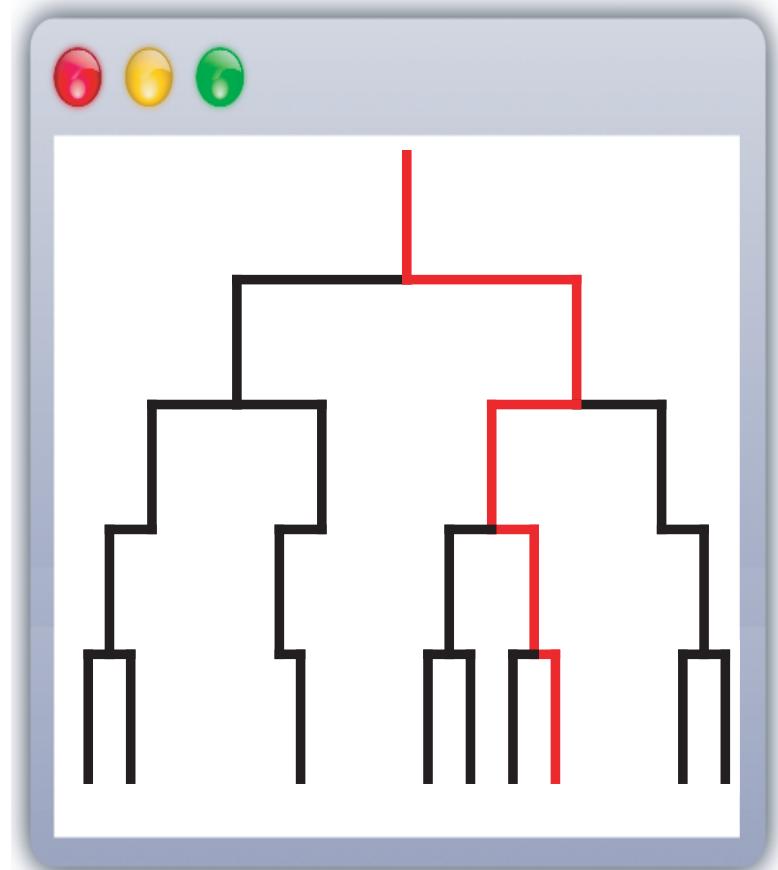
To conclude, the analyses given in ref. 1 are examples of the fact that prediction models using interaction terms have to be handled and interpreted carefully; in particular, using insignificant variables is not expedient and may lead to statistical artifacts.

-named hurricanes are deadlier because people do not take them seriously. This is based on a questionable statistical analysis of a narrowly defined dataset, not robust in that it is not confirmed by a straightforward analysis of another B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

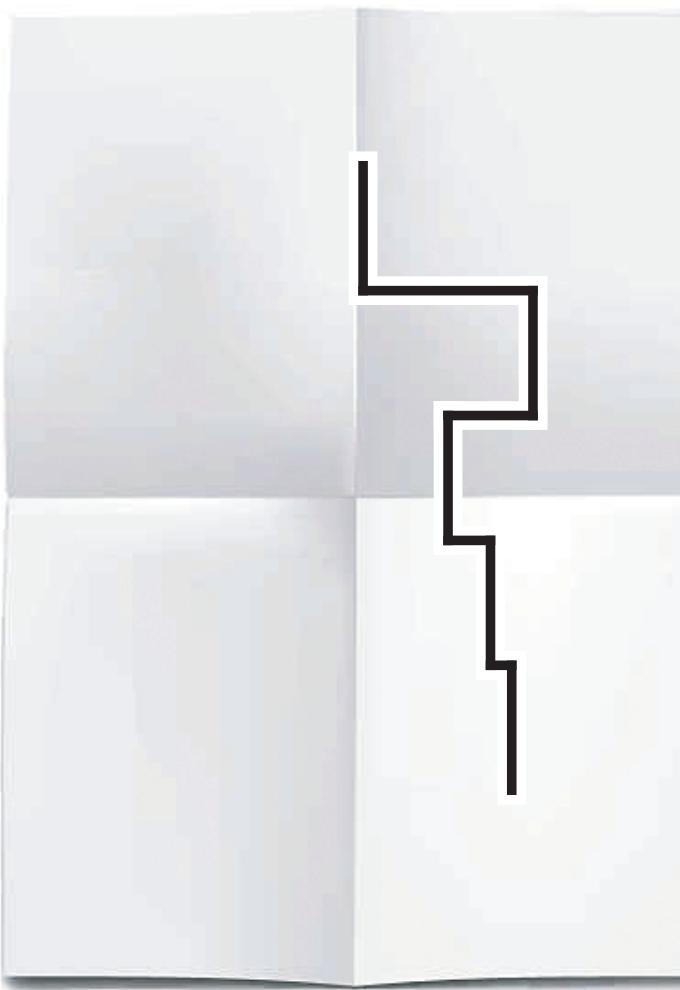
on (less than 39 mph), tropical storm (39–73 mph), hurricane (more than more than 73 mph), and major hurricane (more than Tropical storms and hurricanes are generally given Hurricane Sandy, but tropical depressions are not. al. (2014) examine a narrowly defined dataset: U.S. m Atlantic hurricanes that made landfall in the United n a strong, surprising conclusion is drawn from respect to the myriad decisions used to restrict the're are several issues: clude tropical storms? In 1994 Tropical Storm Al- Ifall near Destin, Florida, with maximum sustained h and caused historic flooding in Alabama and lted in at least 30 deaths and caused \$1.4 billion 4 dollars). Alberto was classified as a hurricane be mph, or

# STATISTICAL ANALYSIS & REPORTING

Analyzed



Reported

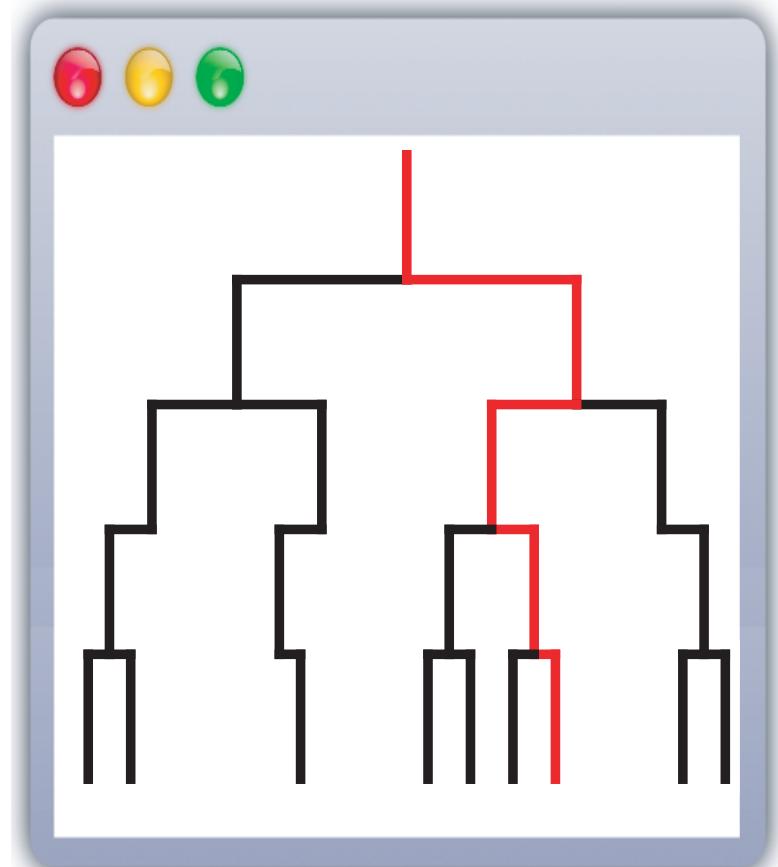


traditional analysis

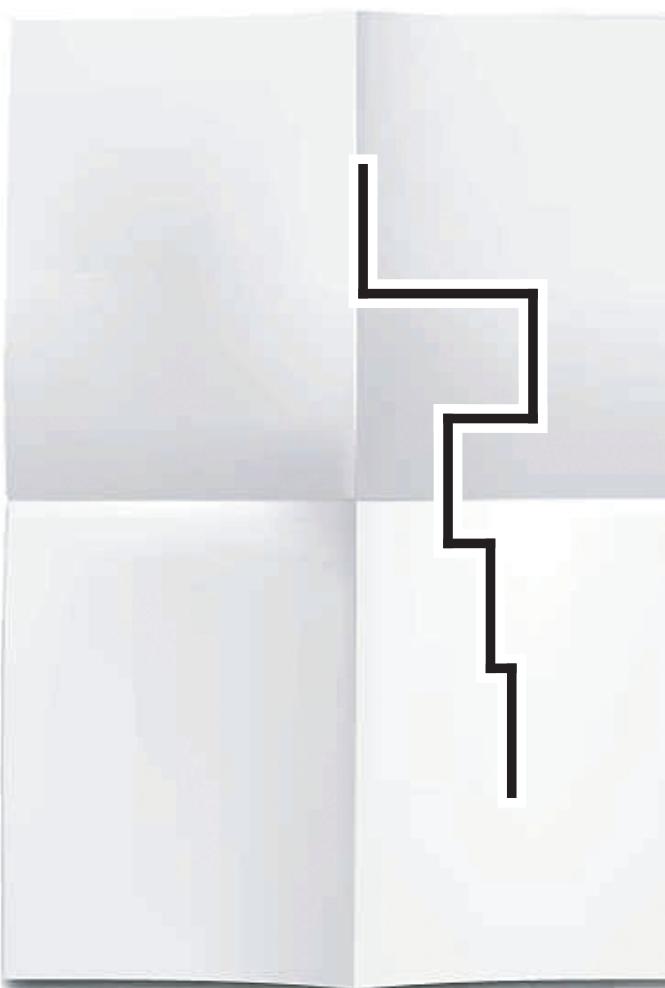
Low transparency

# STATISTICAL ANALYSIS & REPORTING

Analyzed



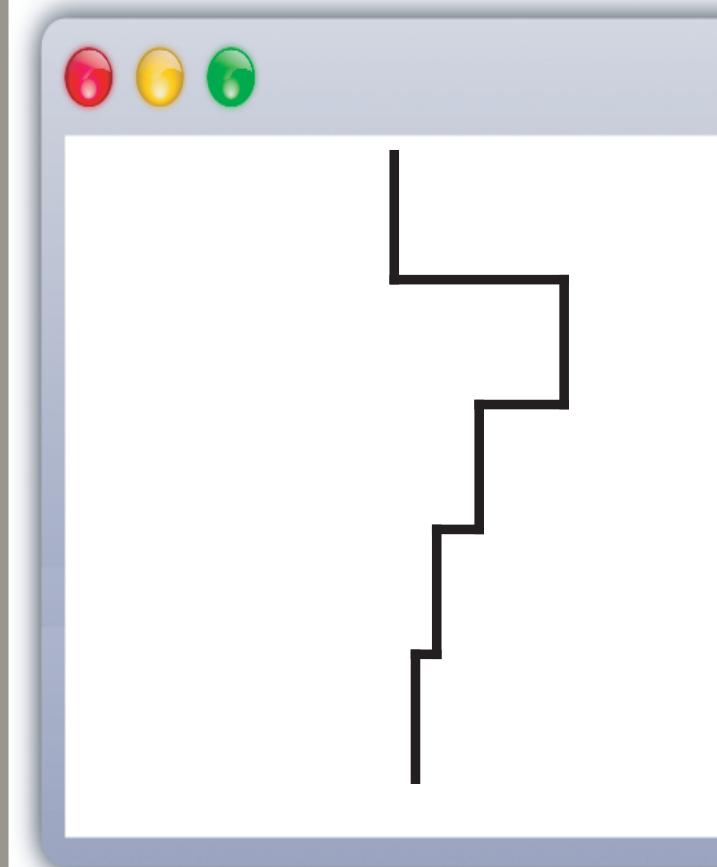
Reported



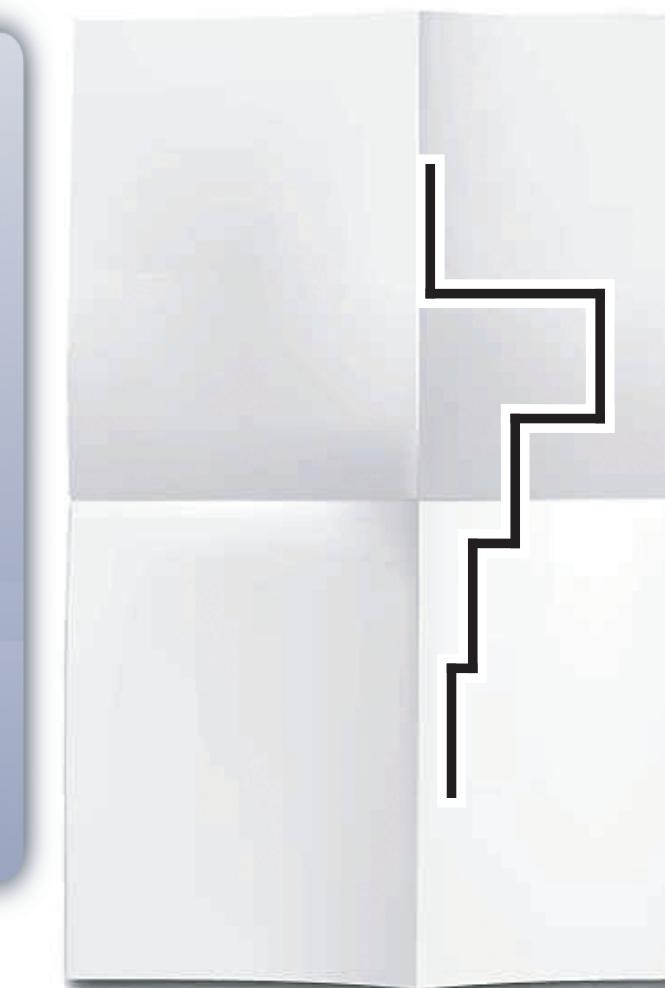
traditional analysis

Low transparency

Analyzed



Reported



planned analysis

High transparency



34 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

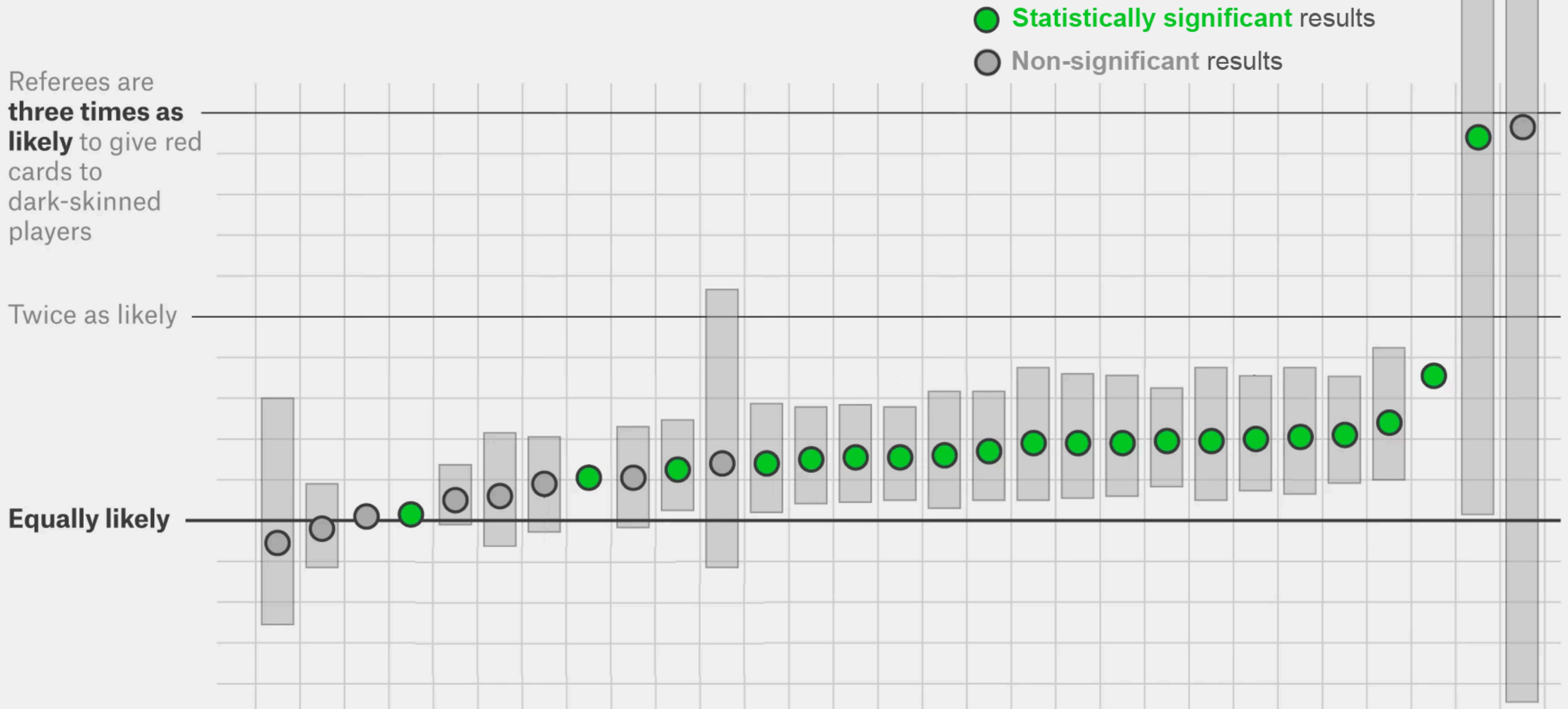
Twice as likely

**Equally likely**



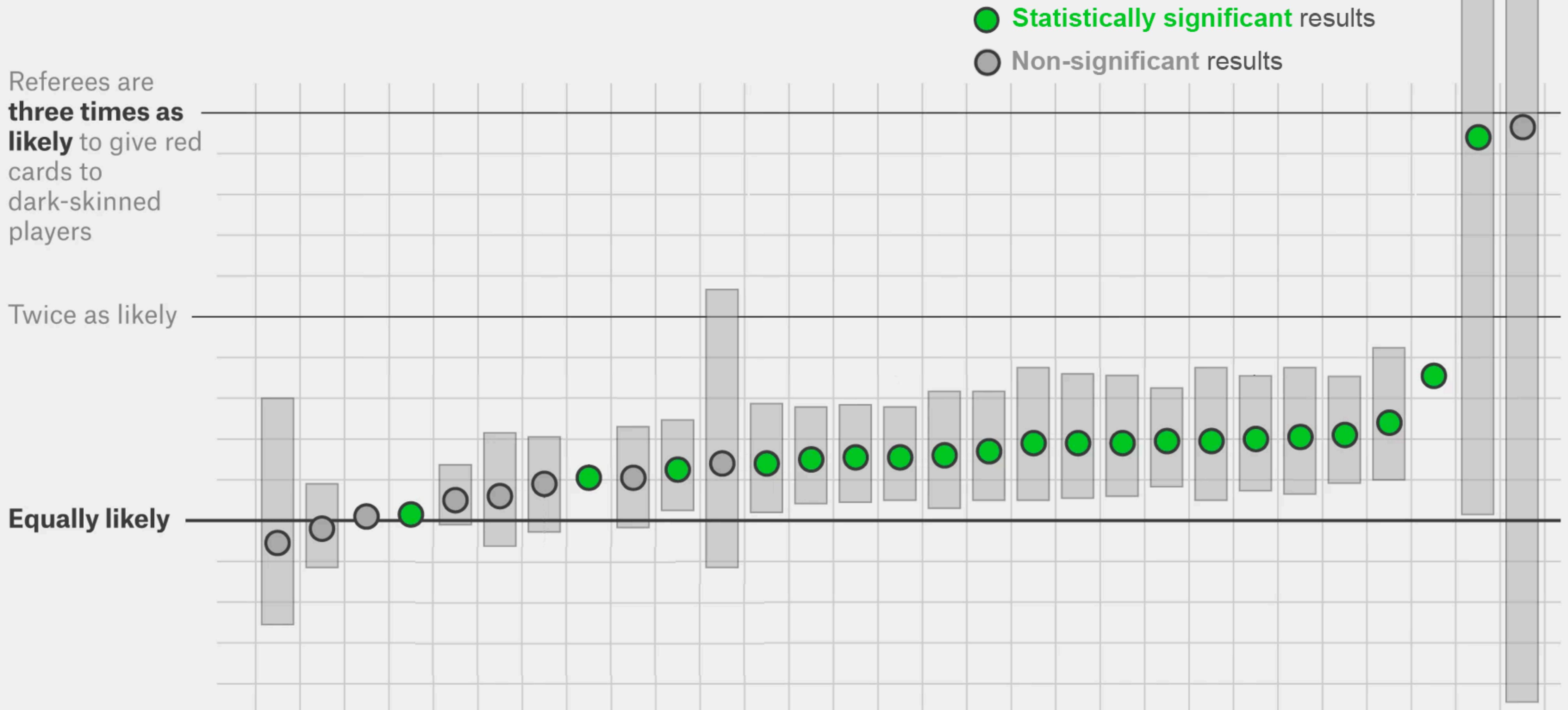
# Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players.



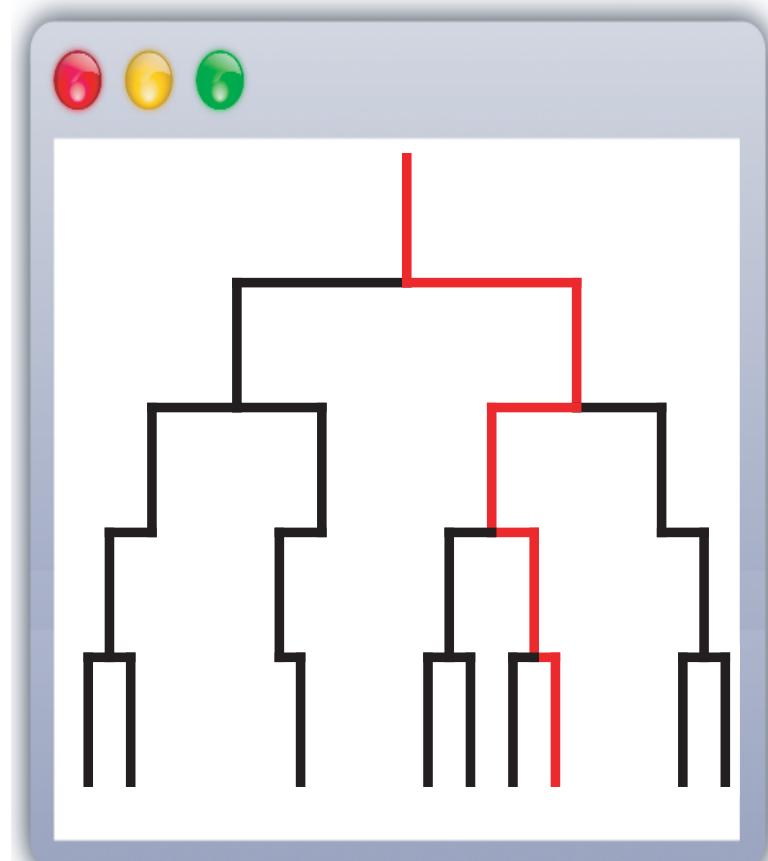
# Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players.

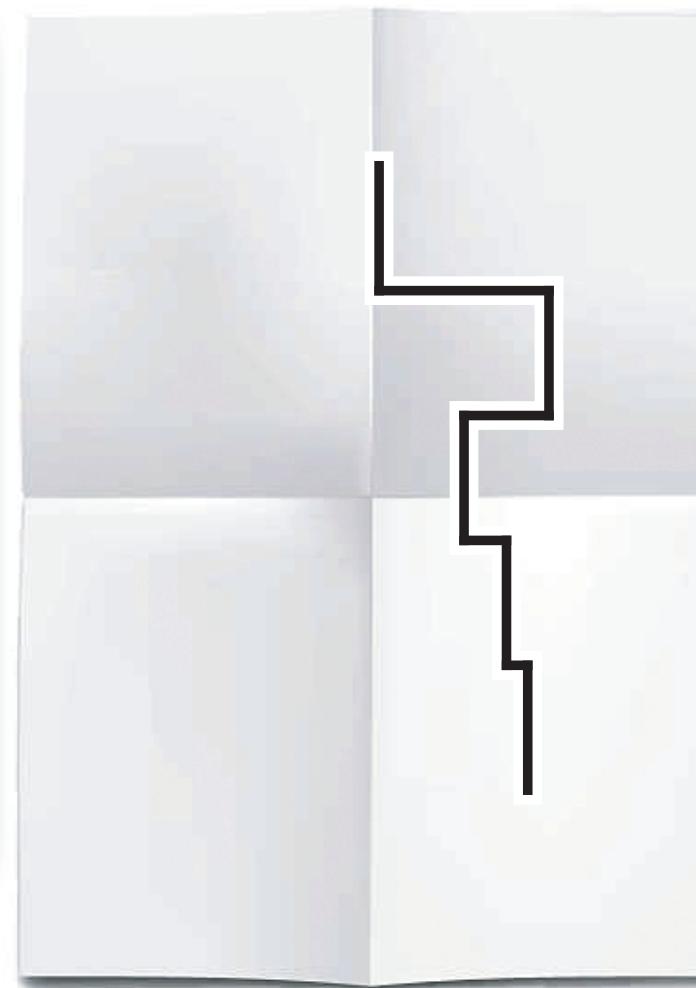


# STATISTICAL ANALYSIS & REPORTING

Analyzed



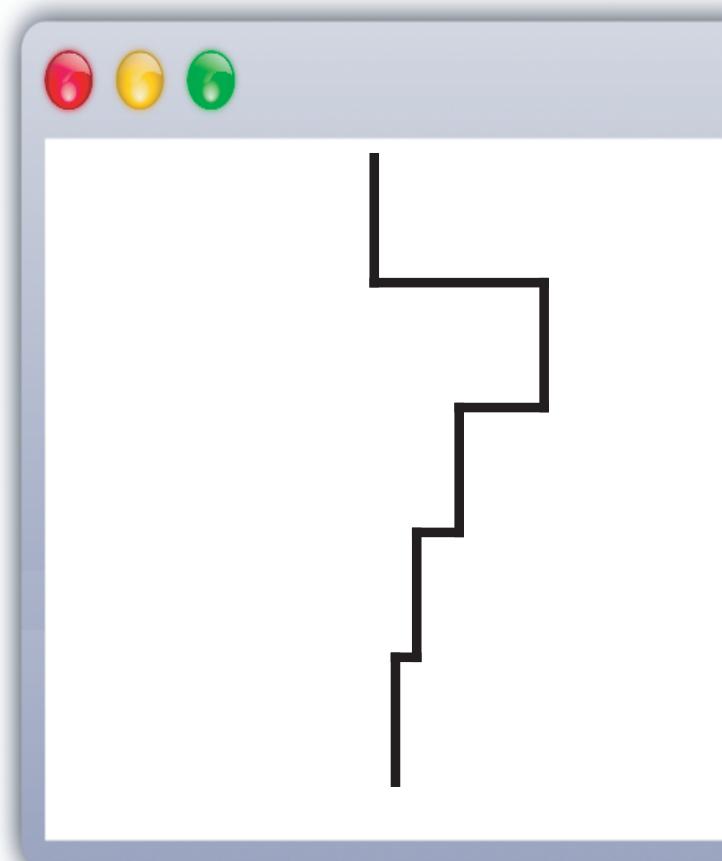
Reported



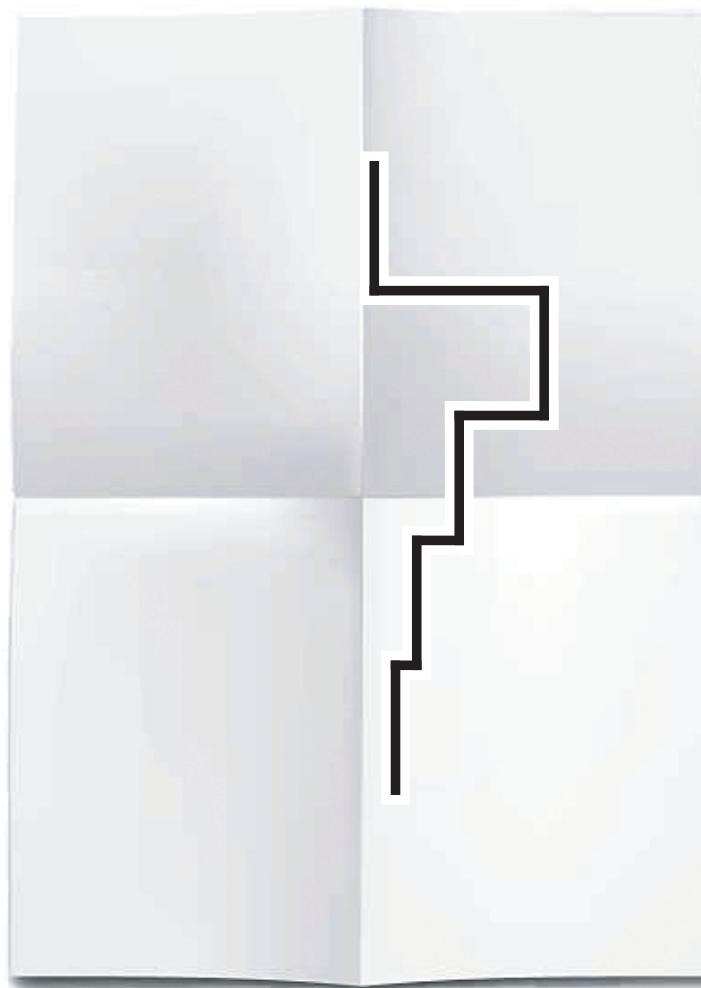
traditional analysis

Low transparency

Analyzed



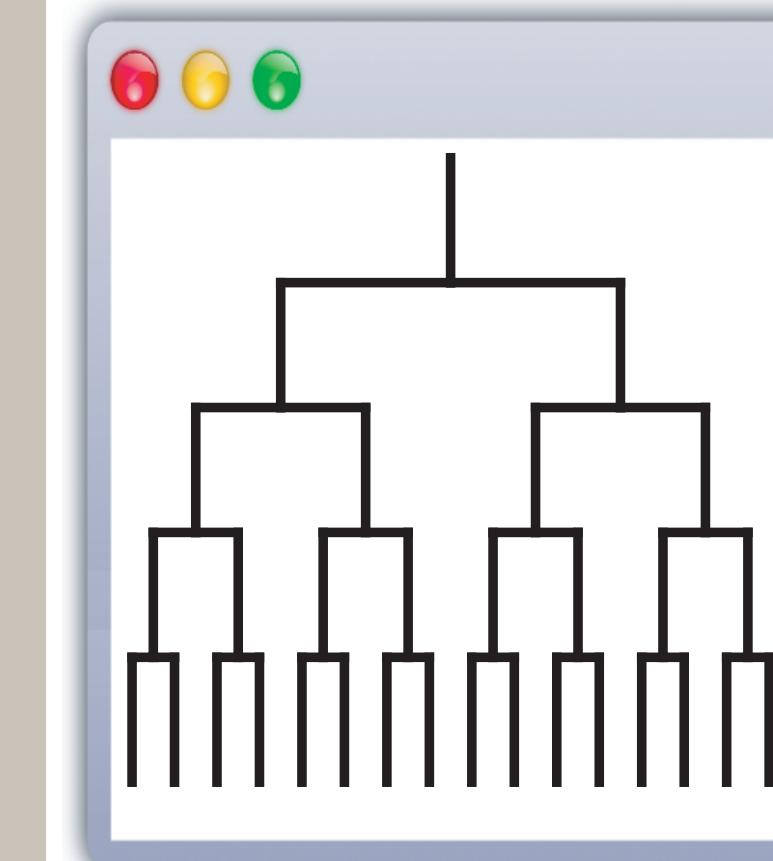
Reported



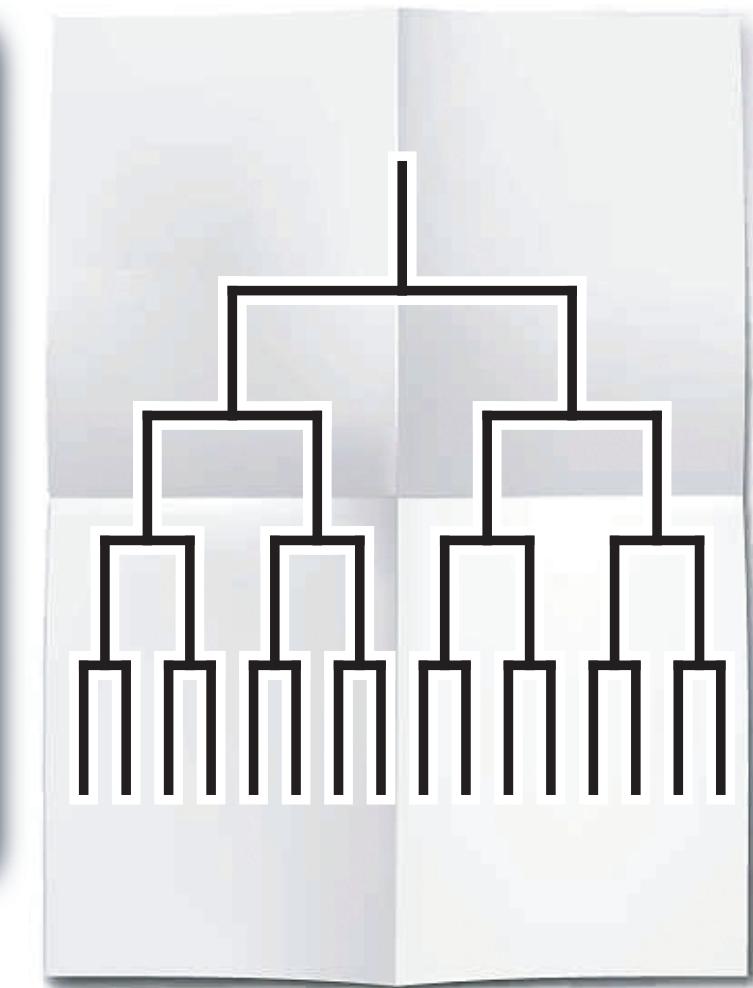
planned analysis

High transparency

Analyzed



Reported

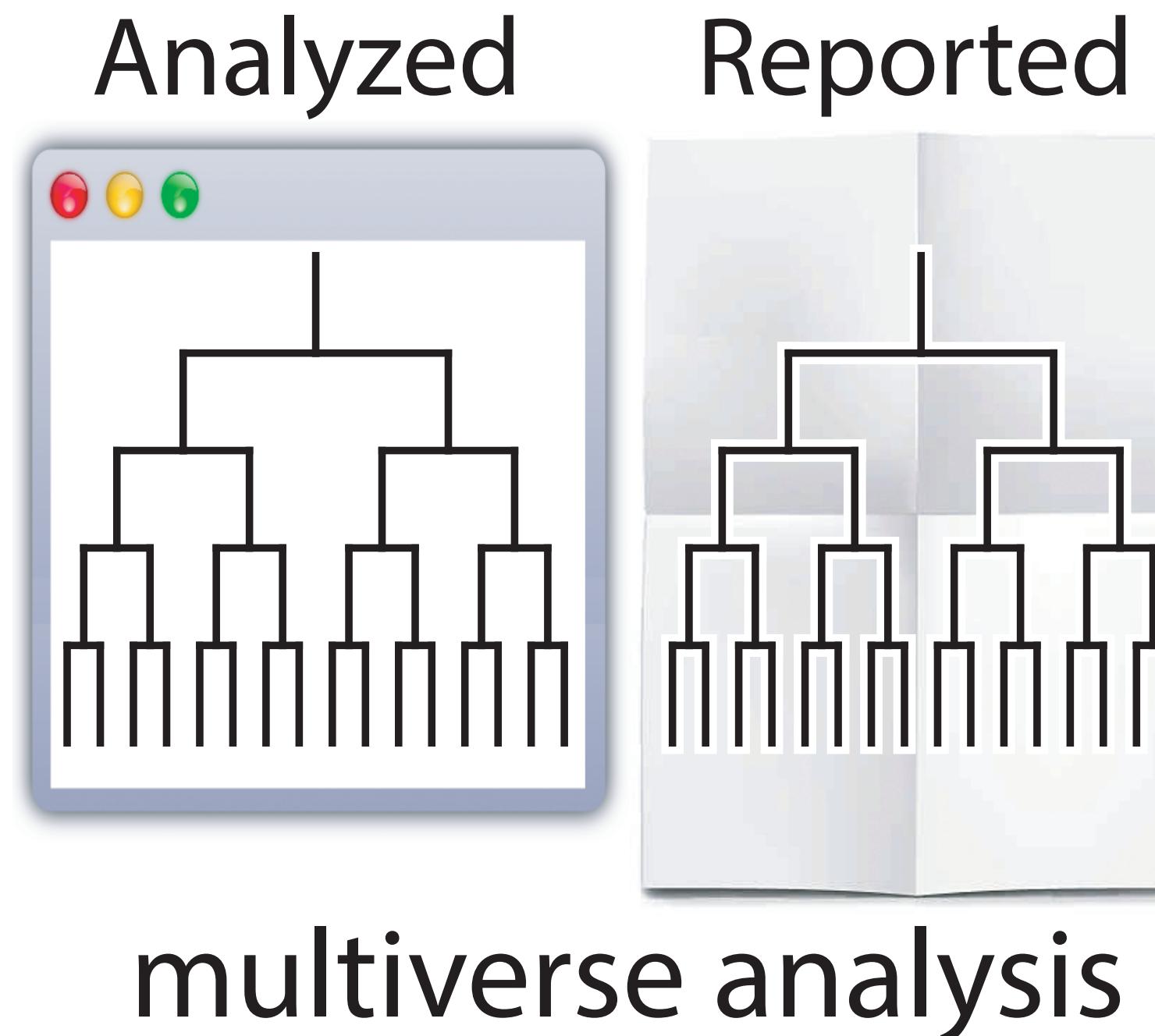


multiverse analysis

Even higher transparency

# MULTIVERSE ANALYSIS

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

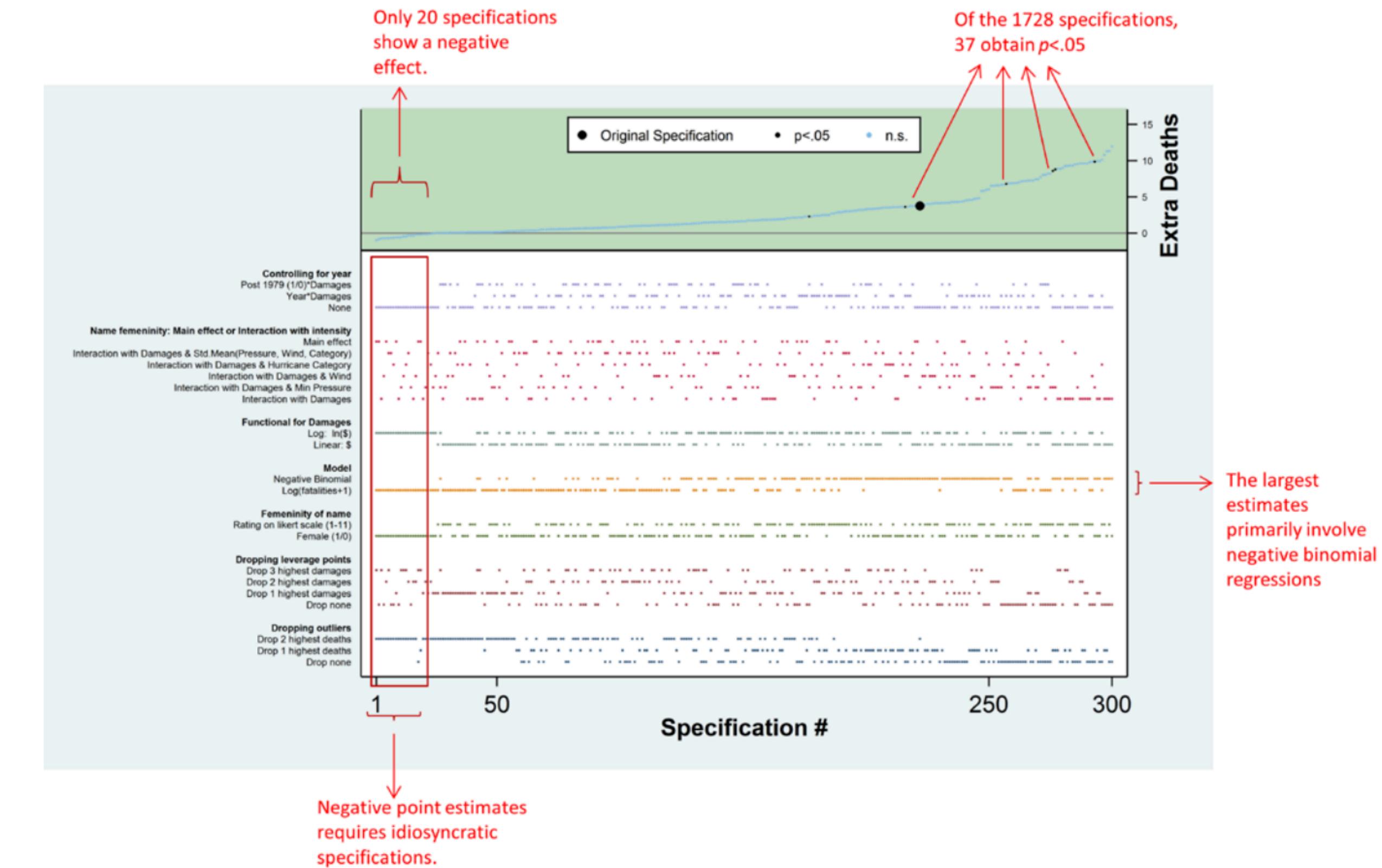
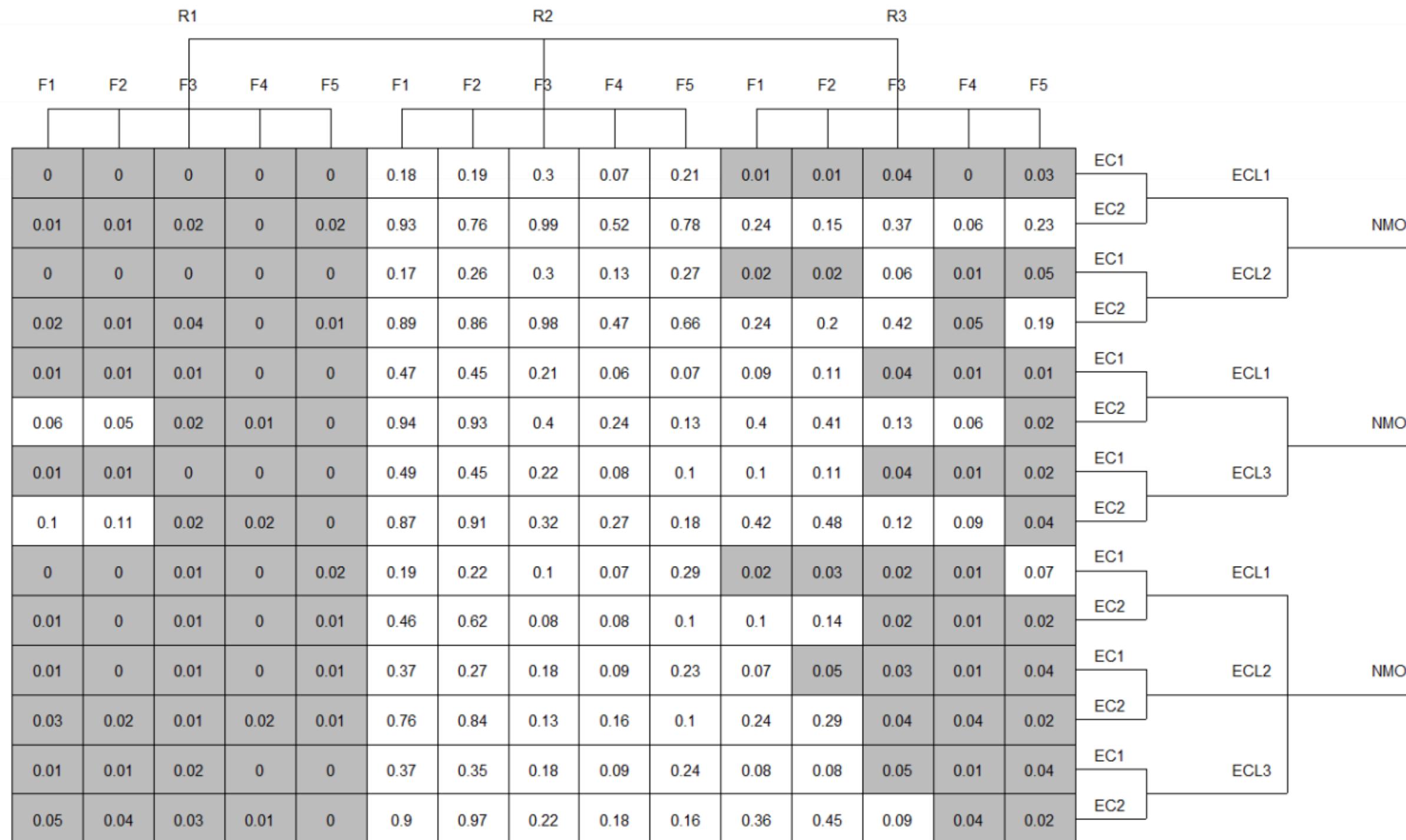


Steegen et al. (2016) Increasing Transparency Through a Multiverse Analysis

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 consequences.

**Result**  
The result was **8.8 TWh** per year.

**Context**

- That is the output of **10 nuclear reactors** (or 1/3rd of the total US nuclear reactors).
- That is the output of **50 fossil power** (or 1/3rd of the total US fossil power).
- That is **4.4%** of all electricity generated.
- That is **6.5%** of all residential electricity consumption.

**Data**

Resource	Capacity (GW)	Output (TWh)
Nuclear	10.80	8.81
Fossil	16.71	10.11
Hydro	9.01	6.11
Solar	0.00	7.01
Wind	1.04	6.11

**Description**

Resource first creates energy and passes it to **Wind** (via "Create All Wind" and "React to Wind"). Wind uses that energy to move from the Bureau of Labor Statistics' "Non-residential" to "Residential" (a value is added to the non-residential total). This creates an additional source of uniform distributed random values (0 to 1) between zero and one, which are then added to the residential total. This creates an additional source of uniform distributed random values (0 to 1) between zero and one, which are then added to the residential total.

Residential has a variable total value, the number of generated sources is uniform distributed between 1 and 10 (i.e. a uniform distribution in the range from 1 to 10 is generated).

Residential has average power generation: **8.8 TWh**.

**Result**

If every household uses its assumed source, home lighting would consume **8.8 TWh** per year. Home lighting currently consumes just 10% per year.

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

Below is a simplified digital adaptation of the scaling state variable filter.

Some example frequency responses:

$$F_c = 2 \text{ KHz}$$
$$Q = 0.8$$
$$F_c = 1.2 \text{ KHz}$$
$$Q = 3.8$$

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

In the case of California, there are many windy areas that are perfect for wind turbines. A significant portion with wind turbines, however, is that the wind is an unreliable source of energy. It is not windy all of the time, which means that when there is no wind power, energy can't be produced. In California, it is generally windy during the summer time when the wind comes in from cooler locations, like the ocean, and then replaces the 'hot rising air' from California's coastal valleys and deserts.

In the case of **California wind**, there are many windy areas that are perfect for wind turbines. A significant portion with wind turbines, however, is that the wind is an unreliable source of energy. It is not windy all of the time, which means that when there is no wind power, energy can't be produced. In California, it is generally windy during the summer time when the wind comes in from cooler locations, like the ocean, and then replaces the 'hot rising air' from California's coastal valleys and deserts.

Wind power in California has been an area of contentious activity for many years. California was the first U.S. state where large wind farms were developed, beginning in the early 1980s. By 1990, California produced 20 percent of the entire world's wind-generated electricity. (Source: [National Energy Policy Development Group](#))

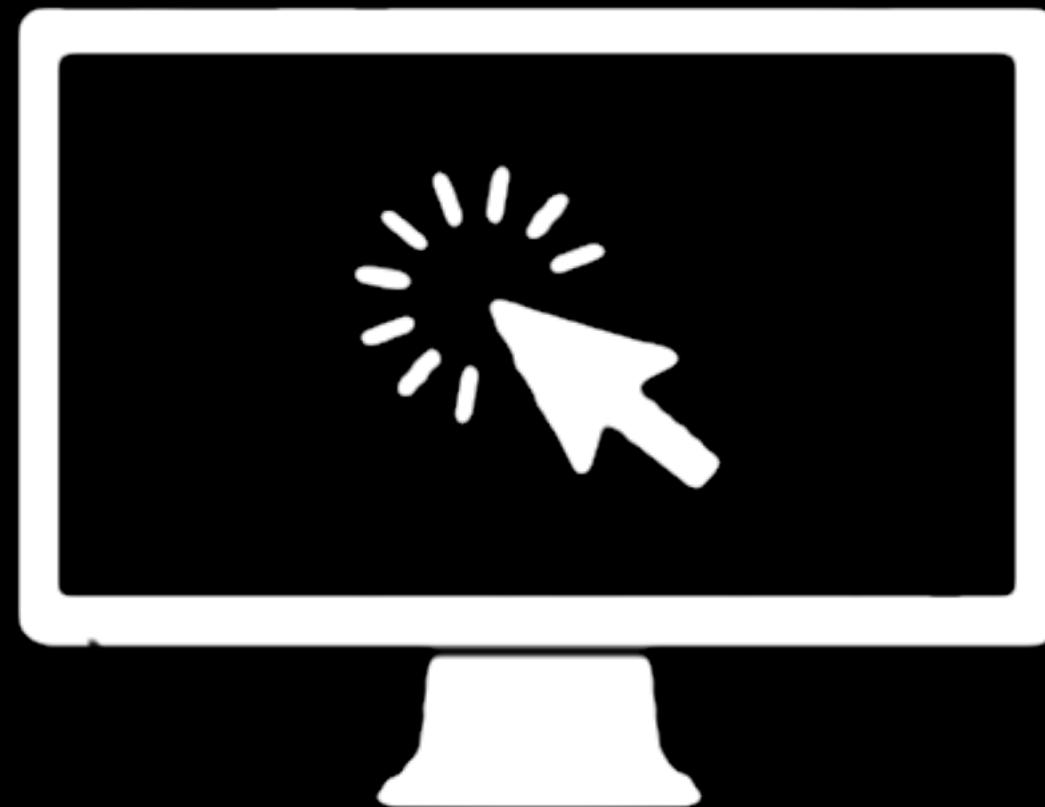
In the case of **wind turbines**, there are many windy areas that are perfect for wind turbines. A significant portion with wind turbines, however, is that the wind is an unreliable source of energy. It is not windy all of the time, which means that when there is no wind power, energy can't be produced. In California, it is generally windy during the summer time when the wind comes in from cooler locations, like the ocean, and then replaces the 'hot rising air' from California's coastal valleys and deserts.

Historically, most of California's wind power output has been in three primary regions: Altamont Pass Wind Farm (near San Francisco), Tehachapi Pass, and the San Gorgonio Pass. These three areas have been the dominant source of energy for the state since the late 1980s. However, more recently, there has been a shift towards smaller-scale wind farms, such as those found in the Mojave Desert and the High Sierra.

In the case of **California wind**, there are many windy areas that are perfect for wind turbines. A significant portion with wind turbines, however, is that the wind is an unreliable source of energy. It is not windy all of the time, which means that when there is no wind power, energy can't be produced. In California, it is generally windy during the summer time when the wind comes in from cooler locations, like the ocean, and then replaces the 'hot rising air' from California's coastal valleys and deserts.

# 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](mailto:pierre.dragicevic@inria.fr)

Yvonne Jansen  
CNRS – Sorbonne Université  
Paris, France  
[yvonne.jansen@sorbonne-universite.fr](mailto:yvonne.jansen@sorbonne-universite.fr)

Abhraneel Sarma  
University of Michigan  
Ann Arbor, MI, USA  
[abhsarma@umich.edu](mailto:abhsarma@umich.edu)

Matthew Kay  
University of Michigan  
Ann Arbor, MI, USA  
[mjskay@umich.edu](mailto:mjskay@umich.edu)

Fanny Chevalier  
University of Toronto  
Toronto, Canada  
[fanny@cs.toronto.edu](mailto: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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org). *CHI 2019, May 4–9, 2019, Glasgow, Scotland UK*

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

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].

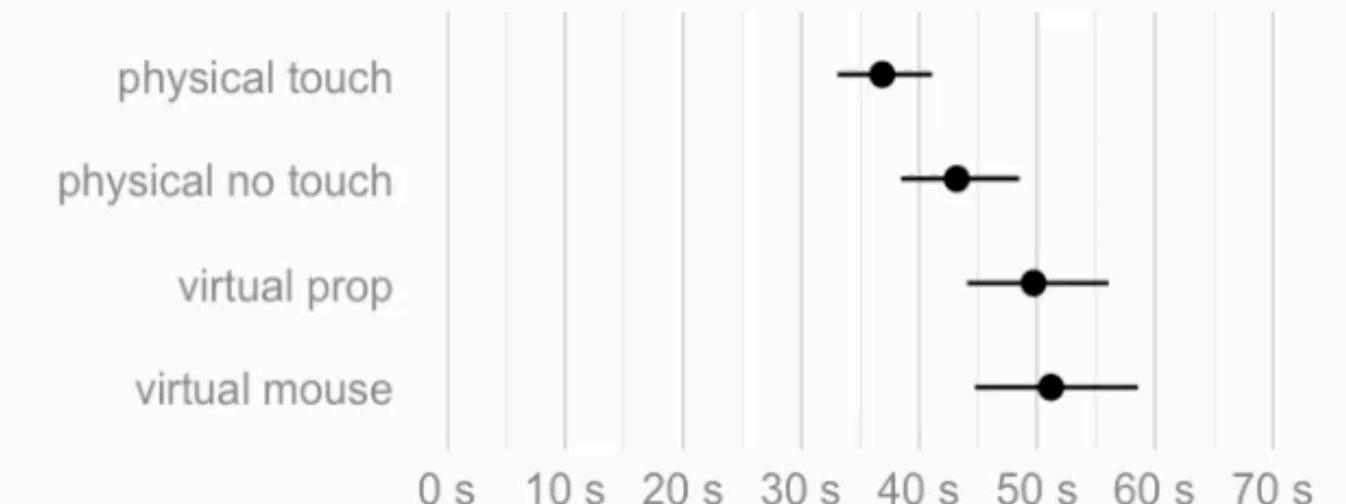


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

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.

physical no t  
virtual

Figure 4.  
(geometric )

Figure 4 pr  
faster on a  
physical no  
that both vi  
touch can fa  
these two p  
hard to fai  
we could no  
opposed 1  
performanc

## DISCUSSION

Our findin  
previously  
default ana  
previously  
default ide  
choices in  
together yi  
conclusions  
less clean  
abnormally  
weight as  
transformat  
outlier remo  
Meanwhile,  
slightly str  
CIs are slig

# Explorable Multiverse Analyses

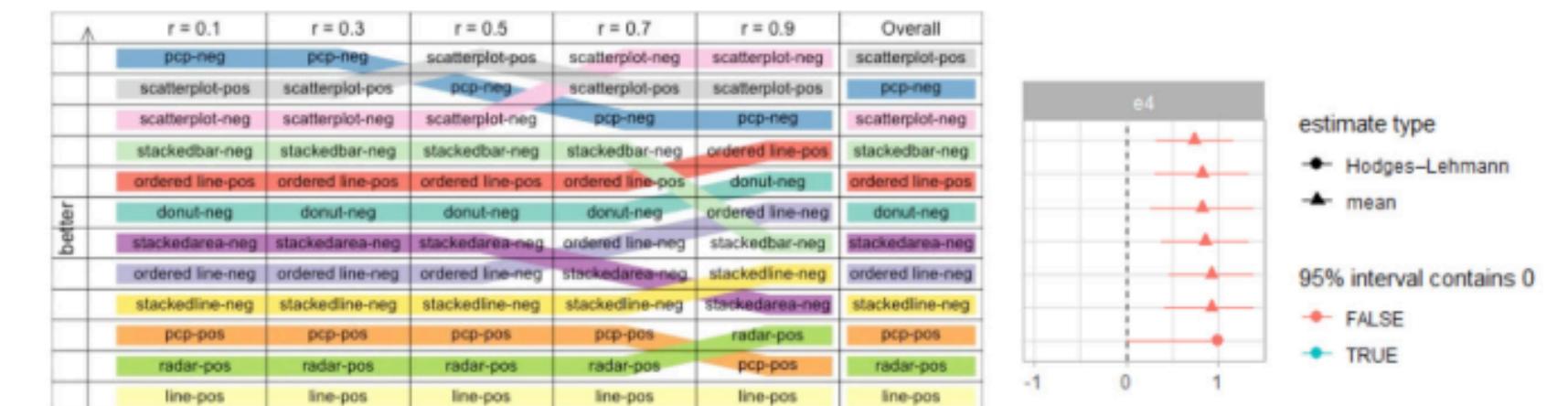
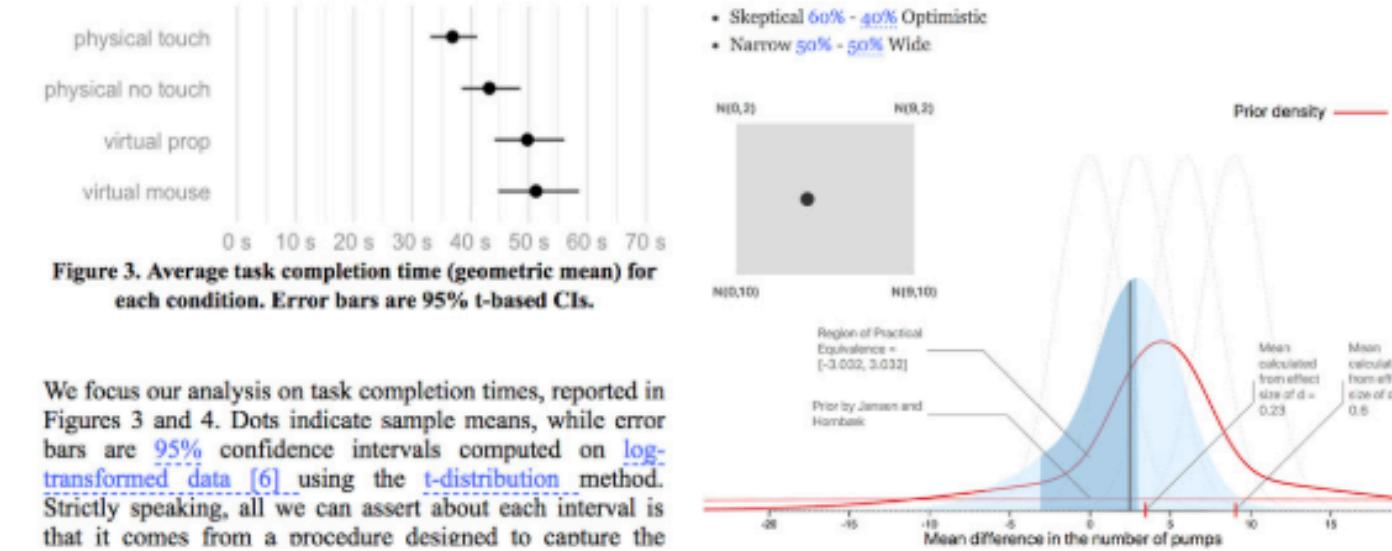


Figure 4. Perceptually-driven ranking of visualizations depending on the correlation sign (-neg / -pos), as a function of correlation value (r) and overall (right column).

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.

**Re-Evaluating the Efficiency of Physical Visualizations: A Simple Multiverse Analysis**

Anonymous Author  
Anonymous affiliation  
anonymous@gmail.org

Abstract: A previous study has shown that moving 3D data visualizations to the physical world can improve users' performance. We conducted a multiverse analysis of a subset of the experimental data using a multiverse analysis approach. Results from this multiverse analysis show that the effect of physicality on task completion times is not robust across different conditions.

Figure 3. 3D bar chart of task completion times.

These manipulations were meant to assess three questions: 1) How important is direct touch in the physical condition? 2) How important is rotation by direct manipulation? 3) How important is the physicality of the objects compared to the virtual objects, especially concerning depth cues. Figure 3 summarizes the three effects of interest.

Figure 4 shows the pairwise ratios between mean completion times. A value lower than 1 (i.e., on the left side of the dark line) means the condition on the left is better than the one on the right. The 95% confidence intervals are not corrected for multiplicity. Since the individual confidence level is 95%, the overall CI does not have a confidence level of 95%. Instead, it has a confidence level of 1 minus the probability of getting at least one false positive among the 3 comparisons (i.e., the familywise error rate) is 99%. Likewise, the simultaneous confidence level is 99%, meaning that if we repeat this experiment 100 times, we would expect the 3 confidence intervals to capture at least 99% of the time.

Figure 5 shows the results of the multiverse analysis.

**A Multiverse Reanalysis of Likert-Type Responses**

Anonymous Author  
Anonymous affiliation  
anonymous@gmail.org

Abstract: There is no consensus on how to best analyze responses to single Likert items. Therefore, studies involving Likert-type responses often include multiple statistical methods for researchers who disagree with the particular statistical analysis method used. We sought to reanalyze responses of 100 participants using 10 different statistical methods. The reanalyses are consistent with the original analysis, and are reasonably robust to the choice of statistical methods.

DATASET AND QUESTIONS: This study has four experiments. In each experiment, each participant is asked to answer a set of questions about a hypothesis along ("no graph" scenario) or the same set of questions about a hypothesis along ("graph" scenario). The response is a Likert scale from 1 to 10. We focus on the first data base. The first data base contains 100 responses to 10 different questions. The responses are consistent with the original analysis, and are reasonably robust to the choice of statistical methods.

Figure 3. Point estimates and 95% intervals for the differences between the two conditions for graph minus no graph.