DETECTING DATA FABRICATION

CHRIS HARTGERINK | @CHARTGERINK

WITH

JAN VOELKEL

JELTE WICHERTS

MARCEL VAN ASSEN

HTTP://BIT.LY/WCRI2017

WORKING MANUSCRIPT

HTTP://BIT.LY/WCRI2017-MANUSCRIPT

Funded by

THE OFFICE OF RESEARCH INTEGRITY

General Article

Psychological Science
© The Author(s) 2013
Reprints and permissions: sagepub.com/journalsPermissions.nav
[illegible]
pss.sagepub.com
SAGE

Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone

Uri Simonsohn
The Wharton School, University of Pennsylvania

Abstract
I argue that requiring authors to post the raw data supporting their published results has the benefit, among many others, of making fraud much less likely to go undetected. I illustrate this point by describing two cases of suspected fraud I identified exclusively through statistical analysis of reported means and standard deviations. Analyses of the raw data behind these published results provided invaluable confirmation of the initial suspicions, ruling out benign explanations (e.g., reporting errors, unusual distributions), identifying additional signs of fabrication, and also ruling out one of the suspected fraud's explanations for his anomalous results. If journals, granting agencies, universities, or other entities overseeing research promoted or required data posting, it seems inevitable that fraud would be reduced.

Keywords
judgment, decision making, scientific communication, fake data, data sharing, data posting

Received 7/20/12; Revision accepted 1/30/13

Academic misconduct is a rare event, but not rare enough. Its occurrence challenges the credibility of research, and the mission of science more generally. Although prevention is important, some misconduct is likely to occur no matter what steps are taken to prevent it. Measures that facilitate identifying such cases can help mitigate their negative consequences. Furthermore, the risk of detection may constitute the ultimate deterrent.

To undetectably fabricate data is difficult. It requires both (a) a good understanding of the phenomenon being studied (e.g., what measures of a construct tend to look like, which variables they correlate with and by how much) and (b) a good understanding of how sampling error is expected to influence the data (e.g., how much variation and the kind of variation the estimates of interest should exhibit given the observed sample size and design). In this article, I show that although means and standard deviations can be analyzed in light of these two criteria to identify likely cases of fraud, the availability of raw data makes the task of detection easier and more diagnostic, and hence that of fabrication more difficult and intimidating.

Posting data has many advantages unrelated to, and possibly more valuable than, prevention and detection of fraud. For example, as Wicherts and Bakker (2012) have noted, when raw data are posted, scientific evidence is preserved for longer periods of time, more researchers get to analyze and hence learn from a given amount of scientific evidence, and reporting errors become easier to prevent and detect.

In this article, I illustrate how raw data can be analyzed for identifying likely fraud through two case studies. Each began with the observation that summary statistics reported in a published article were too similar across conditions to have originated in random samples, an approach to identifying problematic data that has been employed before (Carlisle, 2012; Fisher, 1936; Gaffan & Gaffan, 1992; Kalai, McKay, & Bar-Hillel, 1998; Roberts, 1987; Sternberg & Roberts, 2006).[1] These preliminary analyses of excessive similarity motivated me to contact the authors and request the raw data behind their results. Only when the raw data were analyzed did these suspicions rise to a level of confidence that could trigger

Corresponding Author:
Uri Simonsohn, The Wharton School, University of Pennsylvania, 3730 Walnut St., 500 Huntsman Hall, Philadelphia, PA 19104
E-mail: [illegible]@wharton.upenn.edu

Flawed science:
The fraudulent research practices of social psychologist Diederik Stapel

Levelt Committee

Noort Committee

Drenth Committee

This document is an English translation of the Dutch report '*Falende wetenschap: De frauduleuze onderzoekspraktijken van social-psycholoog Diederik Stapel*'. In the event of any differences between the Dutch report and the translation, the Dutch report will prevail.

28 november 2012

Funded by
THE OFFICE OF RESEARCH INTEGRITY ®

General Article

aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
24(10) 1875–1888
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613480366
pss.sagepub.com
SAGE

Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone

Uri Simonsohn
The Wharton School, University of Pennsylvania

Flawed science:

The fraudulent research practices of social psychologist Diederik Stapel

Abstract
I argue that
others, of n
fraud. I ider
raw data b
explanation
out one of
other entiti

METHODS DEVELOPED IN CASU

Keywords
judgment, c

Received 7/20

APPLICABLE AS GENERIC METHODS?

Academic n
Its occurrence challenges the credibility of research, and the mission of science more generally. Although prevention is important, some misconduct is likely to occur no matter what steps are taken to prevent it. Measures that facilitate identifying such cases can help mitigate their negative consequences. Furthermore, the risk of detection may constitute the ultimate deterrent.

To undetectably fabricate data is difficult. It requires both (a) a good understanding of the phenomenon being studied (e.g., what measures of a construct tend to look like, which variables they correlate with and by how much) and (b) a good understanding of how sampling error is expected to influence the data (e.g., how much variation and the kind of variation the estimates of interest should exhibit given the observed sample size and design). In this article, I show that although means and standard deviations can be analyzed in light of these two criteria to identify likely cases of fraud, the availability of raw data makes the task of detection easier and more diagnostic, and hence that of fabrication more difficult and intimidating.

Posting data has many advantages unrelated to, and possibly more valuable than, prevention and detection of noted, when raw data are posted, scientific evidence is preserved for longer periods of time, more researchers get to analyze and hence learn from a given amount of scientific evidence, and reporting errors become easier to prevent and detect.

In this article, I illustrate how raw data can be analyzed for identifying likely fraud through two case studies. Each began with the observation that summary statistics reported in a published article were too similar across conditions to have originated in random samples, an approach to identifying problematic data that has been employed before (Carlisle, 2012; Fisher, 1936; Gaffan & Gaffan, 1992; Kalai, McKay, & Bar-Hillel, 1998; Roberts, 1987; Sternberg & Roberts, 2006).[1] These preliminary analyses of excessive similarity motivated me to contact the authors and request the raw data behind their results. Only when the raw data were analyzed did these suspicions rise to a level of confidence that could trigger

Corresponding Author:
Uri Simonsohn, The Wharton School, University of Pennsylvania, 3730 Walnut St., 500 Huntsman Hall, Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

This document is an English translation of the Dutch report '*Falende wetenschap: De frauduleuze onderzoekspraktijken van social-psycholoog Diederik Stapel*'. In the event of any differences between the Dutch report and the translation, the Dutch report will prevail.

28 november 2012

Funded by
THE OFFICE OF RESEARCH INTEGRITY ®

	GENUINE	FABRICATED
"GENUINE"		
"FABRICATED"		

Funded by
THE OFFICE OF RESEARCH INTEGRITY

	GENUINE	FABRICATED
"GENUINE"	??	??
"FABRICATED"	??	??

Funded by
THE OFFICE OF RESEARCH INTEGRITY

HOW CAN DATA BE FABRICATED?

Funded by

THE OFFICE OF RESEARCH INTEGRITY ®

MANY LABS (N=36)

	GENUINE	FABRICATED
"GENUINE"	??	??
"FABRICATED"	??	??

Funded by
THE OFFICE OF RESEARCH INTEGRITY

MANY LABS (N=36)

WE ASKED RESEARCHERS (N=36)

	GENUINE	FABRICATED
"GENUINE"	??	??
"FABRICATED"	??	??

Funded by
THE OFFICE OF RESEARCH INTEGRITY

YES, WE ASKED RESEARCHERS TO FABRICATE DATA

Funded by

THE OFFICE OF RESEARCH INTEGRITY

YES, WE ASKED RESEARCHERS TO FABRICATE DATA

AND WE EVEN PAID THEM

Funded by

THE OFFICE OF RESEARCH INTEGRITY ®

YES, WE ASKED RESEARCHERS TO FABRICATE DATA

AND WE EVEN PAID THEM

AND WE PAID THEM EVEN MORE IF THEY WERE TOP 3

Funded by

THE OFFICE OF RESEARCH INTEGRITY

			Mean (true distance: 2,906.5 miles)	Standard Deviation
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female	2562.12	956.35
		Male	2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female	3421.25	845.21
		Male	3380.98	932.56

Funded by

THE OFFICE OF RESEARCH INTEGRITY ®

Anchoring study - distance from San Francisco to New York					
Expectations			**Current result**		**Supported**
Main effect of condition			$F(1, 96) = 21.33, p < .001$		✓
No main effect of gender			$F(1, 96) = 0.03, p = 0.867$		✓
No interaction effect of gender * condition			$F(1, 96) = 0, p = 0.96$		✓
				Mean (true distance: 2,906.5 miles)	**Standard Deviation**
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female		2562.12	956.35
		Male		2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female		3421.25	845.21
		Male		3380.98	932.56

Funded by THE OFFICE OF RESEARCH INTEGRITY

FABRICATION HYPOTHESES

1. SIGNIFICANT CONDITION EFFECT
2. NONSIGNIFICANT GENDER EFFECT
3. NONSIGNIFICANT INTERACTION EFFECT

Funded by
THE OFFICE OF RESEARCH INTEGRITY

Anchoring study - distance from San Francisco to New York

Expectations			Current result	Supported
Main effect of condition			F(1, 96) = 21.33, p < .001	✓
No main effect of gender			F(1, 96) = 0.03, p = 0.867	✓
No interaction effect of gender * condition			F(1, 96) = 0, p = 0.96	✓

			Mean (true distance: 2,906.5 miles)	Standard Deviation
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female	2562.12	956.35
		Male	2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female	3421.25	845.21
		Male	3380.98	932.56

Funded by
THE OFFICE OF RESEARCH INTEGRITY

Anchoring study - distance from San Francisco to New York

Expectations			Current result	Supported
Main effect of condition			F(1, 96) = 21.33, p < .001	✓
No main effect of gender			F(1, 96) = 0.03, p = 0.867	✓
No interaction effect of gender * condition			F(1, 96) = 0, p = 0.96	✓

			Mean (true distance: 2,906.5 miles)	Standard Deviation
Low anchor	The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?	Female	2562.12	956.35
		Male	2540.36	942.14
High anchor	The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?	Female	3421.25	845.21
		Male	3380.98	932.56

WANT TO TRY IT YOURSELF?

HTTP://BIT.LY/TRY-FABRICATION

Funded by

THE OFFICE OF RESEARCH INTEGRITY ®

DATA

EACH OF FOUR STUDIES PROVIDES

1. FOUR VARIANCES
2. TWO NONSIGNIFICANT RESULTS
3. ONE SIGNIFICANT RESULT

Funded by
THE OFFICE OF RESEARCH INTEGRITY

DATA

EACH OF FOUR STUDIES PROVIDES

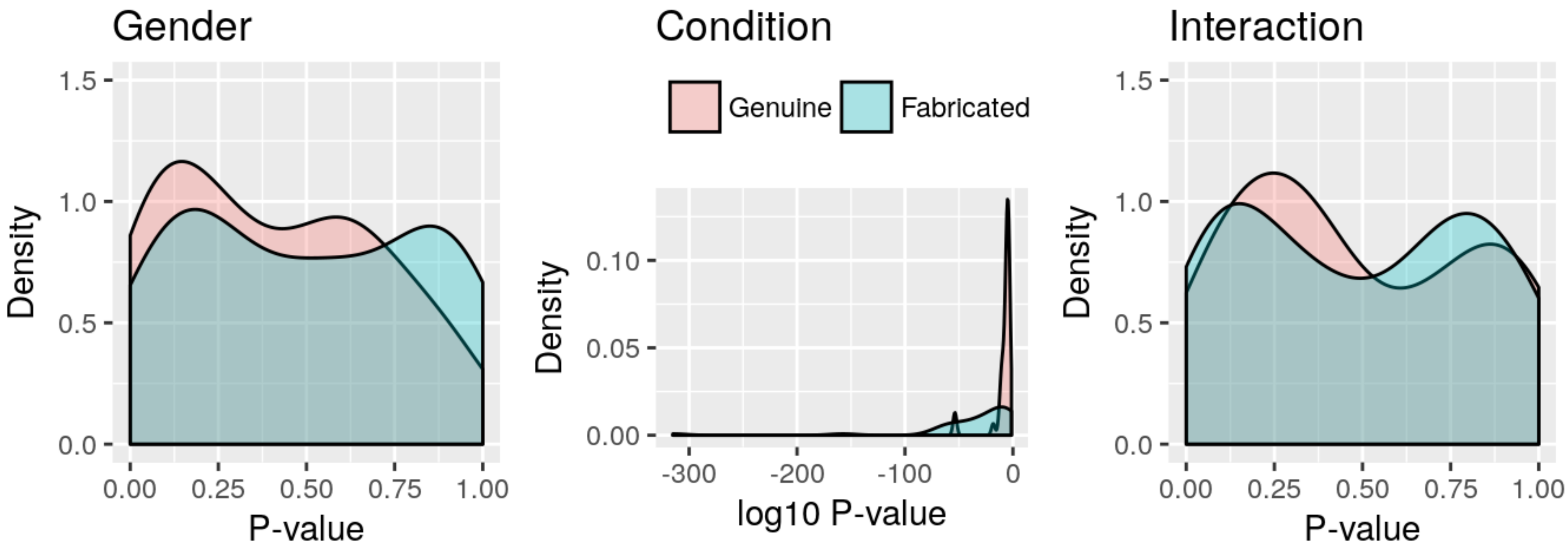
1. FOUR VARIANCES
2. TWO NONSIGNIFICANT RESULTS
3. ONE SIGNIFICANT RESULT

PER RESPONDENT:

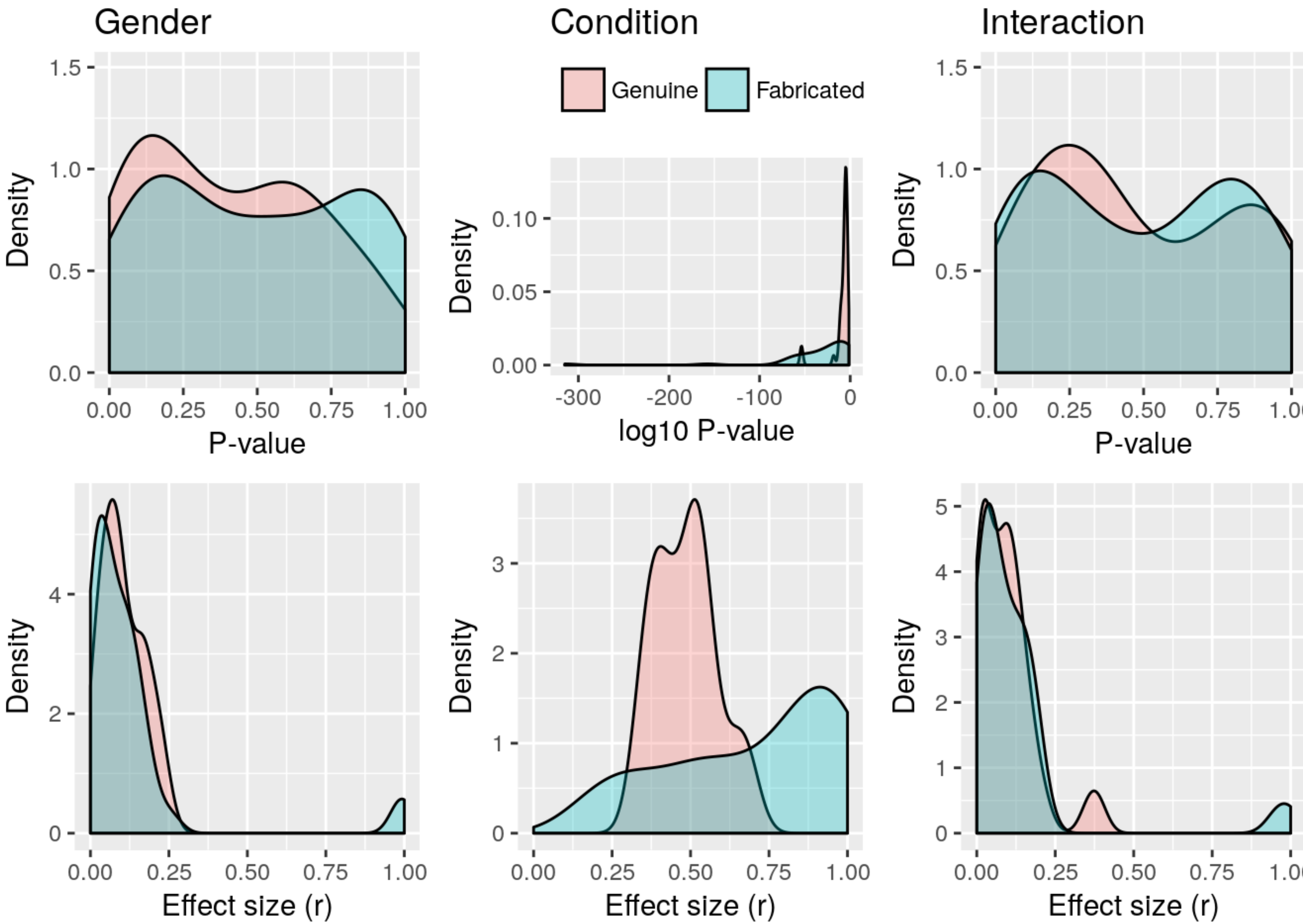
1. SIXTEEN VARIANCES
2. EIGHT NONSIGNIFICANT RESULTS
3. FOUR SIGNIFICANT RESULTS

Funded by

THE OFFICE OF RESEARCH INTEGRITY ®



Funded by
THE OFFICE OF RESEARCH INTEGRITY


Gender
Condition
Interaction
Genuine
Fabricated
Density
P-value
log10 P-value
Effect size (r)
1.5
1.0
0.5
0.0
0.10
0.05
0.00
0.00
0.25
0.50
0.75
1.00
-300
-200
-100
0
5
4
3
2
1
0

Funded by

EFFECTIVENESS METHODS

METHOD	AUROC

Funded by

THE OFFICE OF RESEARCH INTEGRITY

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770

Funded by
THE OFFICE OF RESEARCH INTEGRITY

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770

$$z_j^2 \sim \left(\frac{\chi^2_{N_j - 1}}{N_j - 1} \right) / MS_w$$

$$MS_w = \frac{\sum_{j=1}^{k} (N_j - 1) s_j^2}{\sum_{j=1}^{k} (N_j - 1)}$$

Funded by

THE OFFICE OF RESEARCH INTEGRITY

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770
NONSIGNIFICANT P-VALUES GENDER EFFECT	.521
NONSIGNIFICANT P-VALUES INTERACTION EFFECT	.535

Funded by
THE OFFICE OF RESEARCH INTEGRITY

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770
NONSIGNIFICANT P-VALUES GENDER EFFECT	.521
NONSIGNIFICANT P-VALUES INTERACTION EFFECT	.535

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \ln(1 - \frac{p_i - t}{1 - t})$$

Funded by

THE OFFICE OF RESEARCH INTEGRITY

EFFECTIVENESS METHODS

METHOD	AUROC
VARIANCE ANALYSIS, ASSUMES = VARIANCE ACROSS CONDITIONS	.423
VARIANCE ANALYSIS, ASSUMES != VARIANCE ACROSS CONDITIONS	.770
NONSIGNIFICANT P-VALUES GENDER EFFECT	.521
NONSIGNIFICANT P-VALUES INTERACTION EFFECT	.535
EFFECT SIZE (1- CORRELATION)	.744

Funded by

THE OFFICE OF RESEARCH INTEGRITY

BUT ASSUMES 50/50 PREVALENCE

Funded by

THE OFFICE OF RESEARCH INTEGRITY

BUT ASSUMES 50/50 PREVALENCE

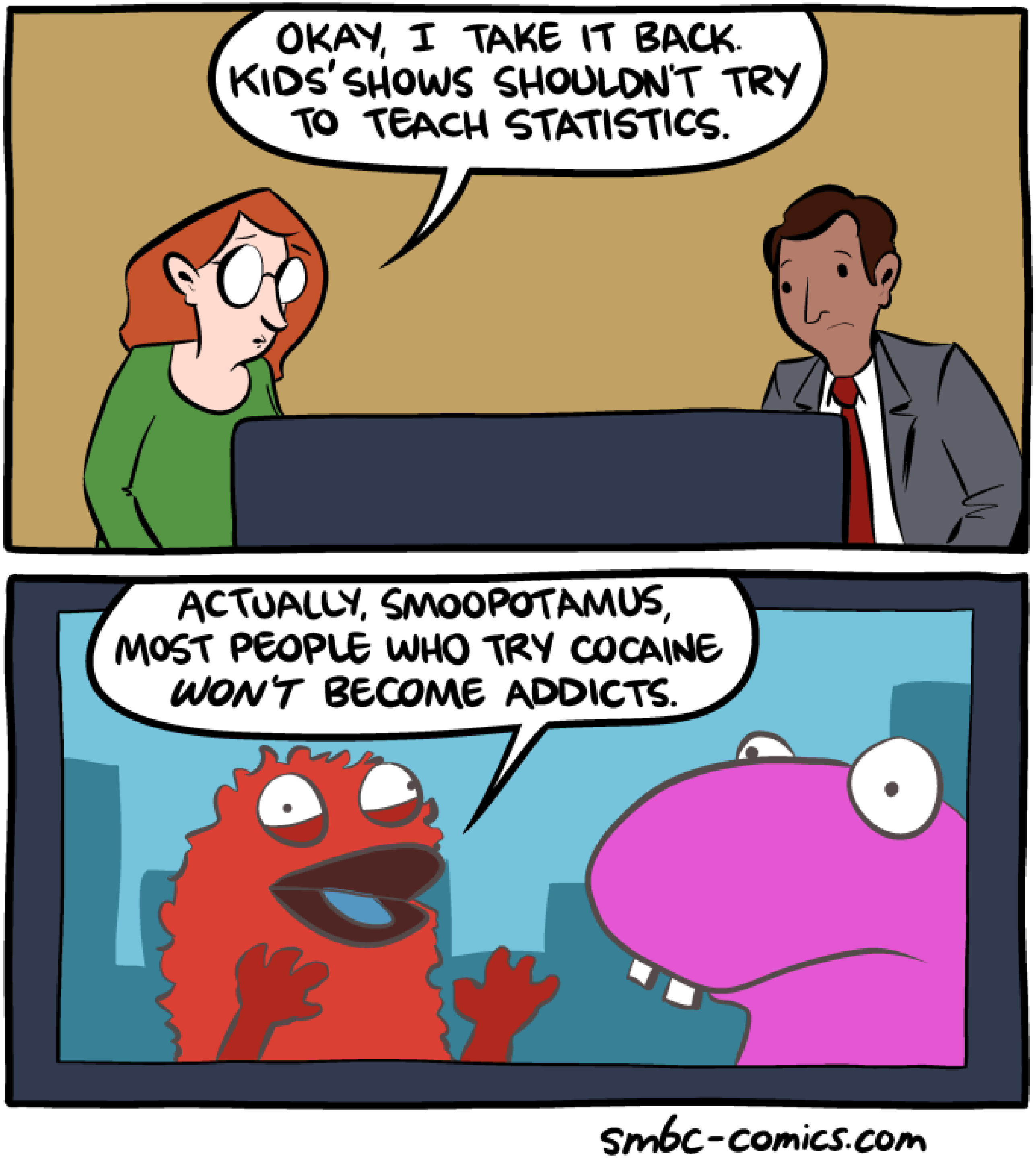
SO ONLY USE AS FLAG, NOT EVIDENCE OF

Funded by

THE OFFICE OF RESEARCH INTEGRITY

DUAL USE PROBLEM

Funded by
THE OFFICE OF RESEARCH INTEGRITY



Funded by THE OFFICE OF RESEARCH INTEGRITY

LESSONS

1. FABRICATED NONSIGNIFICANT DATA LOOK RATHER GENUINE
2. VARIANCE ANALYSIS SENSITIVE TO POPULATION VARIANCES BEING HETERO- OR HOMOGENEOUS (IMPORTANT IN APPLICATION)
3. LARGE EFFECT SIZES EASY INDICATOR (R>.7)
4. USE AS FLAG, NOT EVIDENCE

Funded by
THE OFFICE OF RESEARCH INTEGRITY

THIS WAS JUST STUDY ONE

Funded by
THE OFFICE OF RESEARCH INTEGRITY

THIS WAS JUST STUDY ONE

STUDY TWO REPLICATES WITH RAW DATA

Funded by
THE OFFICE OF RESEARCH INTEGRITY

THIS WAS JUST STUDY ONE

STUDY TWO REPLICATES WITH RAW DATA

AND ADDS INTERVIEWS!

Funded by

THE OFFICE OF RESEARCH INTEGRITY

ONGOING, BUT YOU CAN FIND INCOMING TRANSCRIPTS @

HTTP://BIT.LY/WCRI2017-TRANSCRIPTS

Funded by

THE OFFICE OF RESEARCH INTEGRITY

ONGOING, BUT YOU CAN FIND INCOMING TRANSCRIPTS @
HTTP://BIT.LY/WCRI2017-TRANSCRIPTS

FEEL FREE TO USE FOR RESEARCH!

Funded by
THE OFFICE OF RESEARCH INTEGRITY