Psychological and sociological motivations and challenges for blind analyses

Rob MacCoun

Scientia nihil aliud est quam veritatis imago. (Science is but an image of the truth.)

- Francis Bacon

Scientia nihil aliud est quam veritatis imago. (Science is but an image of the truth.)

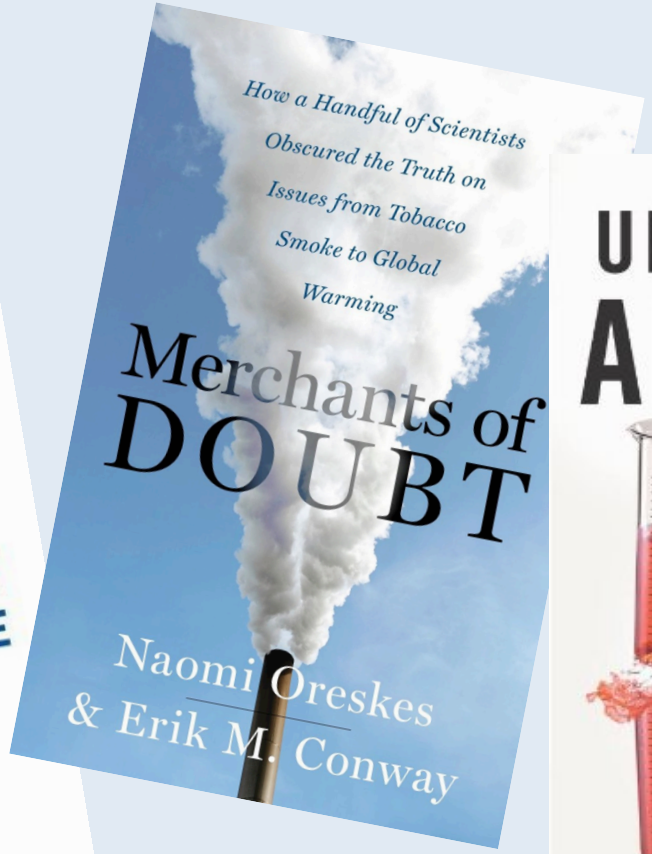
- Francis Bacon

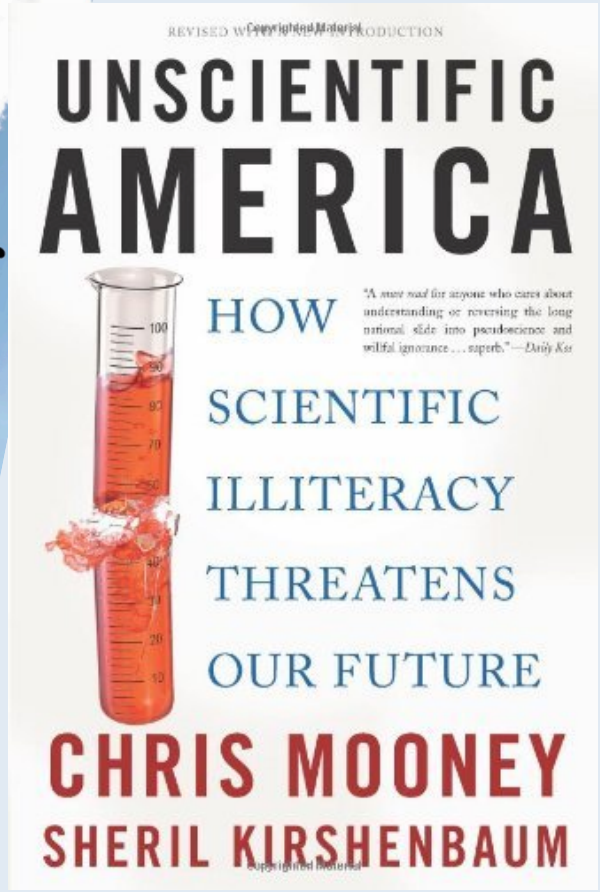
Back off man. I'm a scientist.

- Bill Murray, playing Dr. Peter Venkman, in the movie *Ghostbusters*.


command
We have
Charts & Graphs
to back us up
so fuck off


Copyrighted Material
NEW YORK TIMES BESTSELLER
"Nothing short of a landmark in contemporary political reporting."
—Salon.com
The REPUBLICAN WAR on SCIENCE
REVISED AND UPDATED
CHRIS MOONEY
Copyrighted Material


How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming
Merchants of DOUBT
Naomi Oreskes & Erik M. Conway


Copyrighted Material
UNSCIENTIFIC AMERICA
HOW SCIENTIFIC ILLITERACY THREATENS OUR FUTURE
CHRIS MOONEY
SHERIL KIRSHENBAUM


SOCIAL SELECTION
Popular articles on social media
A proposal to cure sloppy science
A debate this week on way
quickly spread to social m
College London, called 'Is
we fix it?', included claims
practices, such as tweakin
results seem significant, ar
com/4imbij). One suggest
to register their ex
with a journal bef
support. Audienc
publisher at the op
NATURE.COM
For more on
popular papers:
go.nature.com/mjyiia
SOCIAL SELECTION
Popular topics on social media
Science reporting flaws exposed
Researchers often complain about inaccurate science
stories in the popular press, but few air their grievances in a
The New York Times
http://nyti.ms/1JIfMXV
The Opinion Pages
OP-ED CONTRIBUTORS
What's Behind Big Science Frauds?
By ADAM MARCUS and IVAN ORANSKY
MAY 22, 2015
NATURE.COM
For more on
popular papers:
go.nature.com/awqdox
Canada, posted on Facebook:
"Journalists and scientists should both
strive for accuracy in their claims."
Front. Psychol. 6, 00988 (2015)

The New York Times | http://nyti.ms/1JoE6N6

SCIENCE

Many Psychology Findings Not as Strong as Claimed, Study Says

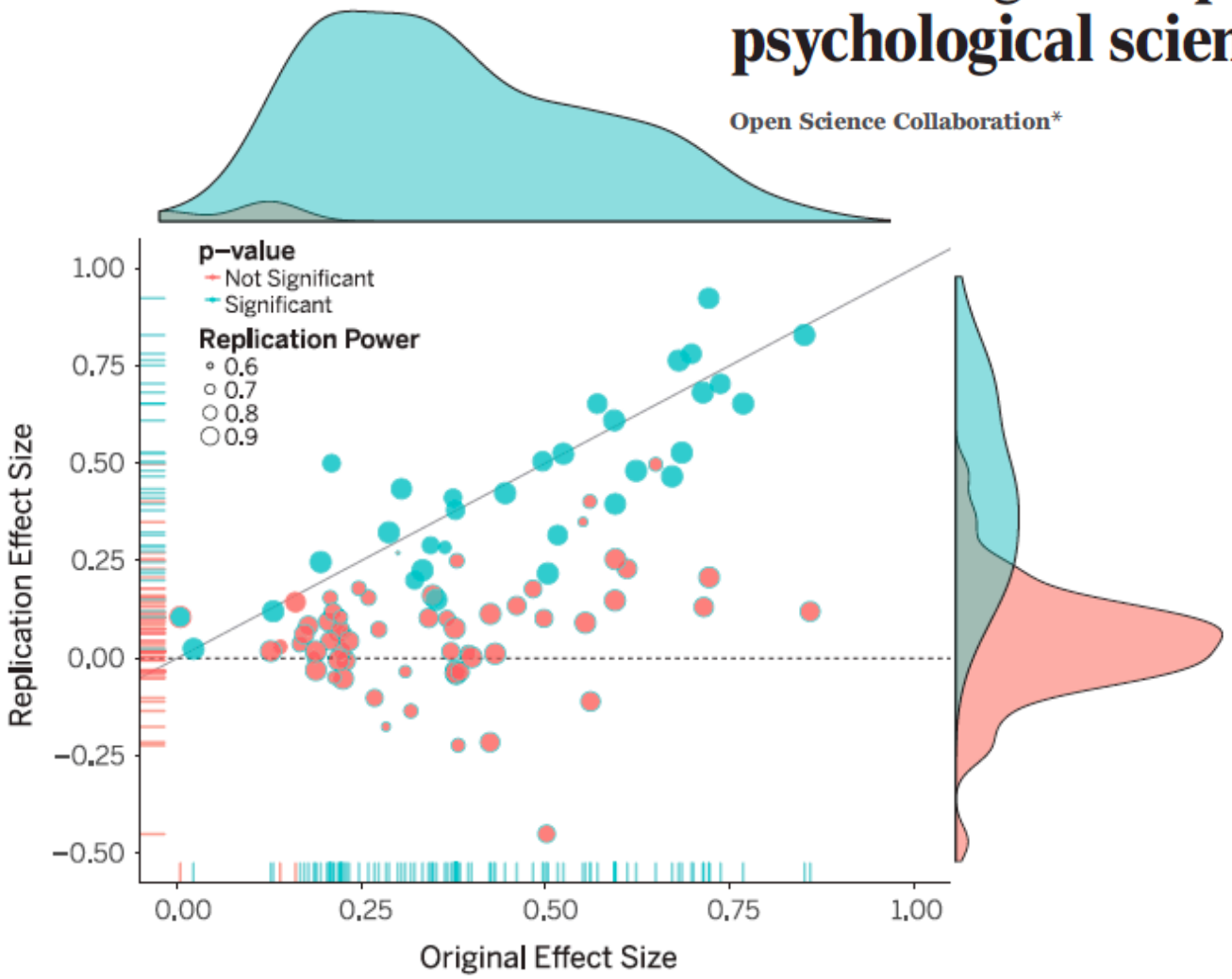
By BENEDICT CAREY AUG. 27, 2015

Estimating the reproducibility of psychological science

Open Science Collaboration*

Estimating the reproducibility of psychological science

Open Science Collaboration*



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Rescuing US biomedical research from its systemic flaws

Bruce Alberts[a], Marc W. Kirschner[b], Shirley Tilghman[c,1], and Harold Varmus[d]

[a]Department of Biophysics and Biochemistry, University of California, San Francisco, CA 94158; [b]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; [c]Department of Molecular Biology, Princeton University, Princeton, NJ 08540; and [d]National Cancer Institute, Bethesda, MD 20892

NATURE | NEWS

Irreproducible biology research costs put at $28 billion per year

Study calculates cost of flawed biomedical research in the United States.

Monya Baker

09 June 2015

Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias

Kerry Dwan[1]*, Douglas G. Altman[2], Juan A. Arnaiz[3], Jill Bloom[4], An-Wen Chan[5], Eugenia Cronin[6], Evelyne Decullier[7], Philippa J. Easterbrook[8], Erik Von Elm[9,10], Carrol Gamble[1], Davina Ghersi[11], John P. A. Ioannidis[12,13], John Simes[14], Paula R. Williamson[1]

Annu. Rev. Psychol. 1998. 49:259–87

BIASES IN THE INTERPRETATION AND USE OF RESEARCH RESULTS

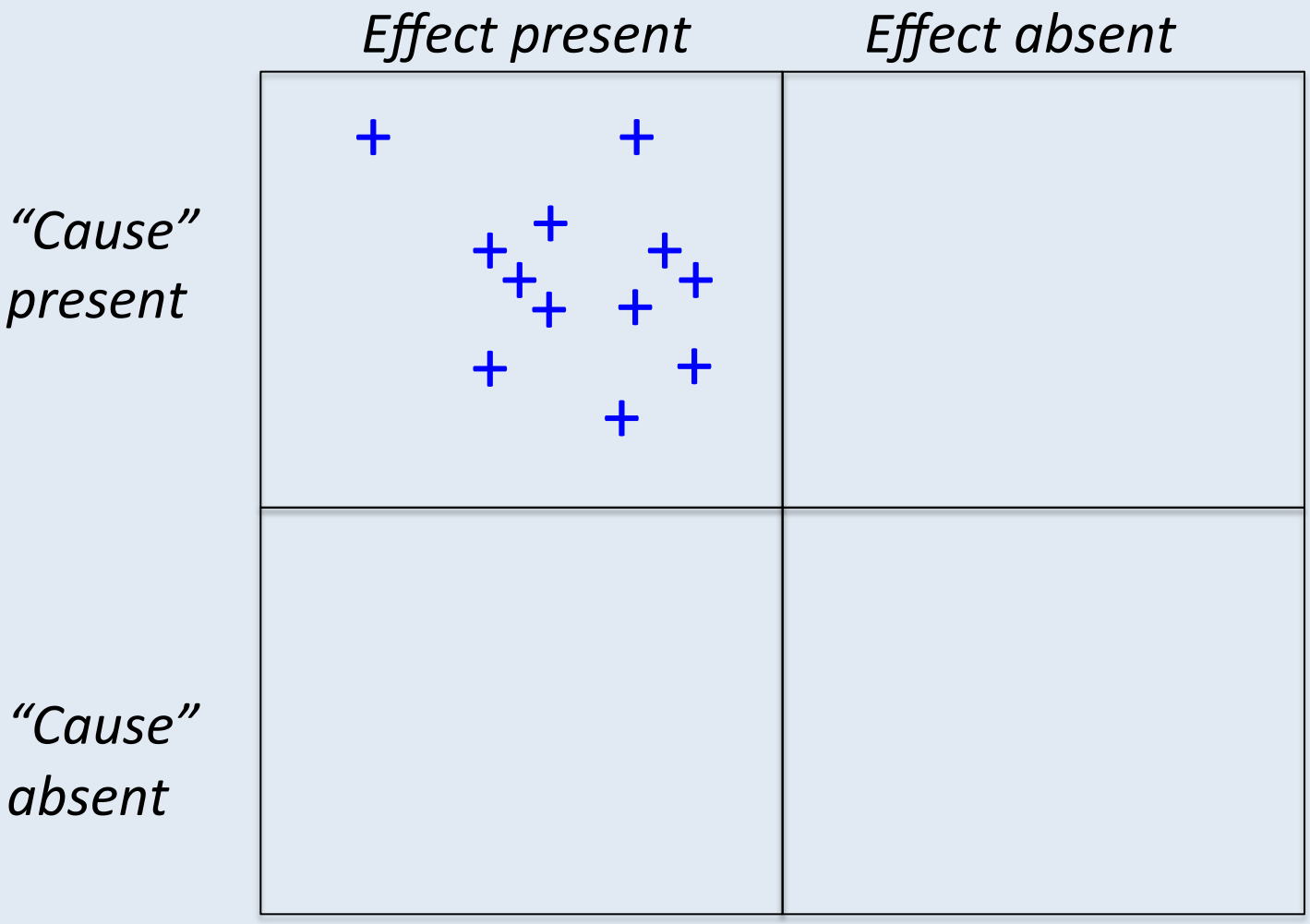
Robert J. MacCoun

Bias Prototypes

	Intentional?	*Motivated?*	*Justifiable?*
Fraud	Yes	Yes	No
Advocacy	Yes	Yes	Maybe
Cold Bias	No	No	No
Hot Bias	No	Yes	No
Skepticism	Maybe	Yes	Yes

MacCoun (1998, *Ann.Rev.Psy.*)

Confirmatory bias in evidence search



Confirmation bias in evidence evaluation: Lord et al (1979)

- Participants were either *pro or anti capital punishment*
- Read 1 of 4 versions of a research article:

	Found deterrent effect	Found no deterrent effect
Longitudinal		
Cross-sectional		

Lord et al (1979): Findings

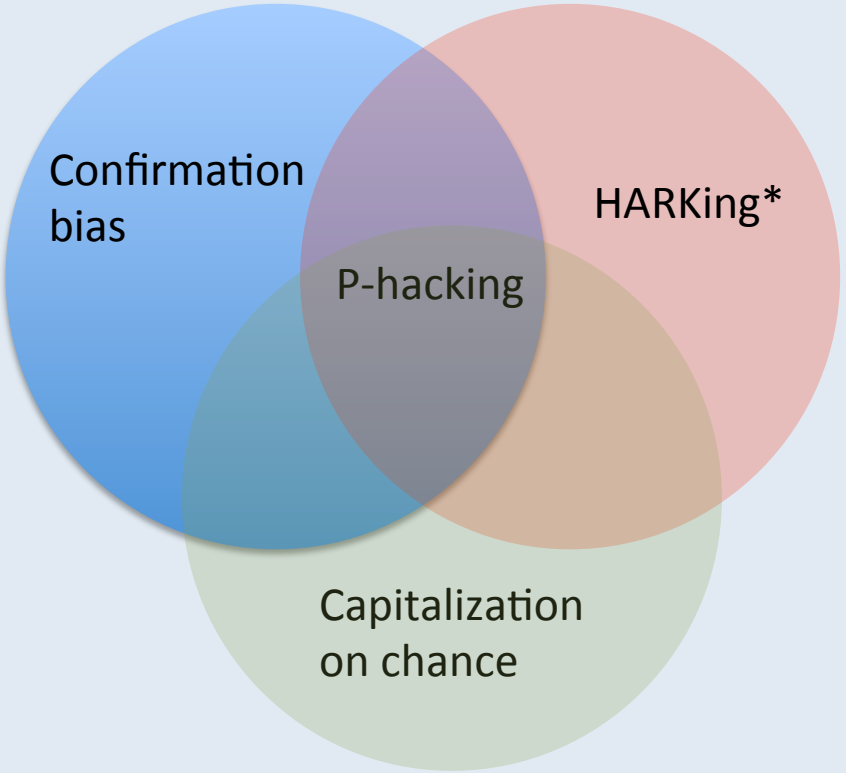
- *Biased assimilation:* Favored whichever study supported one's own views
- *Attitude polarization:* Groups grew farther apart after reading research

Disconfirmation Bias

- "But in *our* lab, we constantly seek out errors and artifacts, and we correct them."
- Surely true,
  - But we are more likely to correct the results we didn't expect…
  - …while overlooking errors that might be producing data that looks like what we expected.

"p hacking" to achieve statistical significance

- *Trimming outliers*
- *Collecting more cases*
- *Throwing out cases*
- *Subgroup analyses*

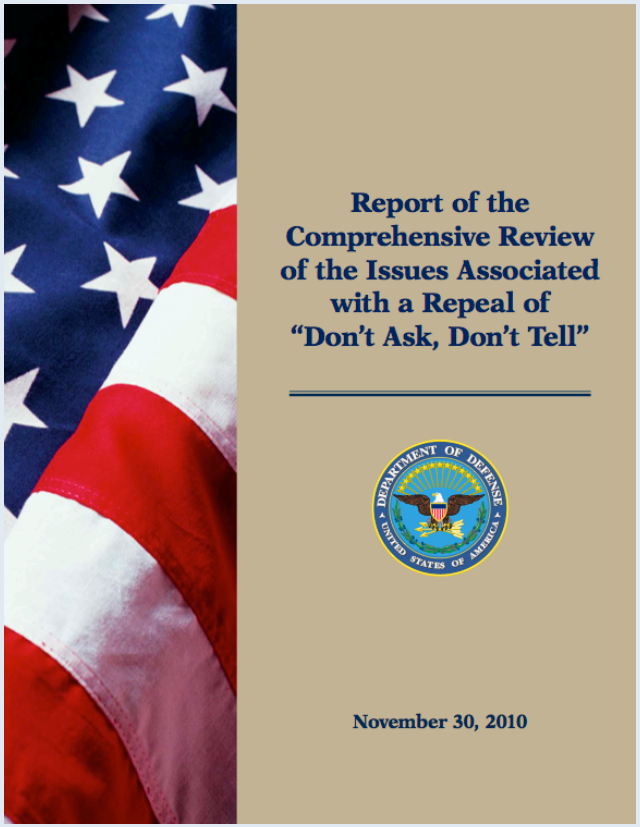


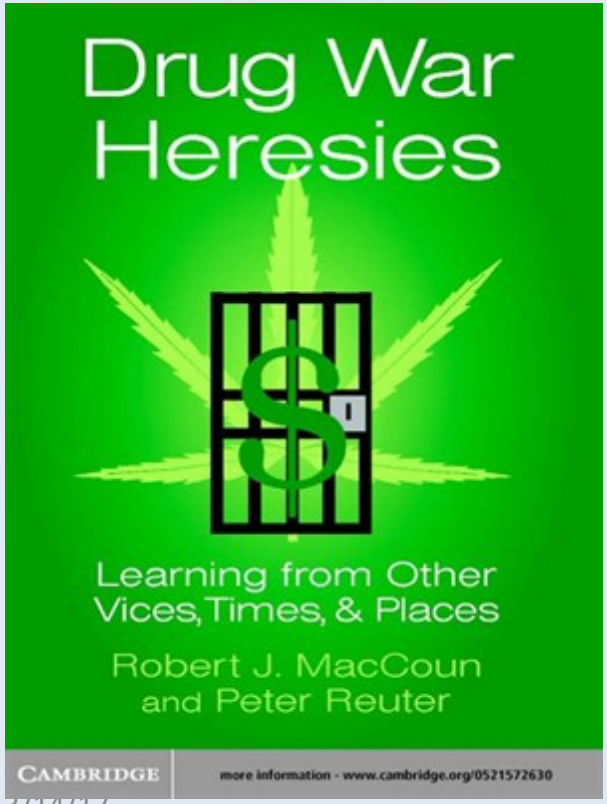
HARKing = "Hypothesizing After the Results are Known"

Data collection → Analysis → Interpretation

SEXUAL ORIENTATION AND U.S. MILITARY PERSONNEL POLICY

An Update of RAND's 1993 Study



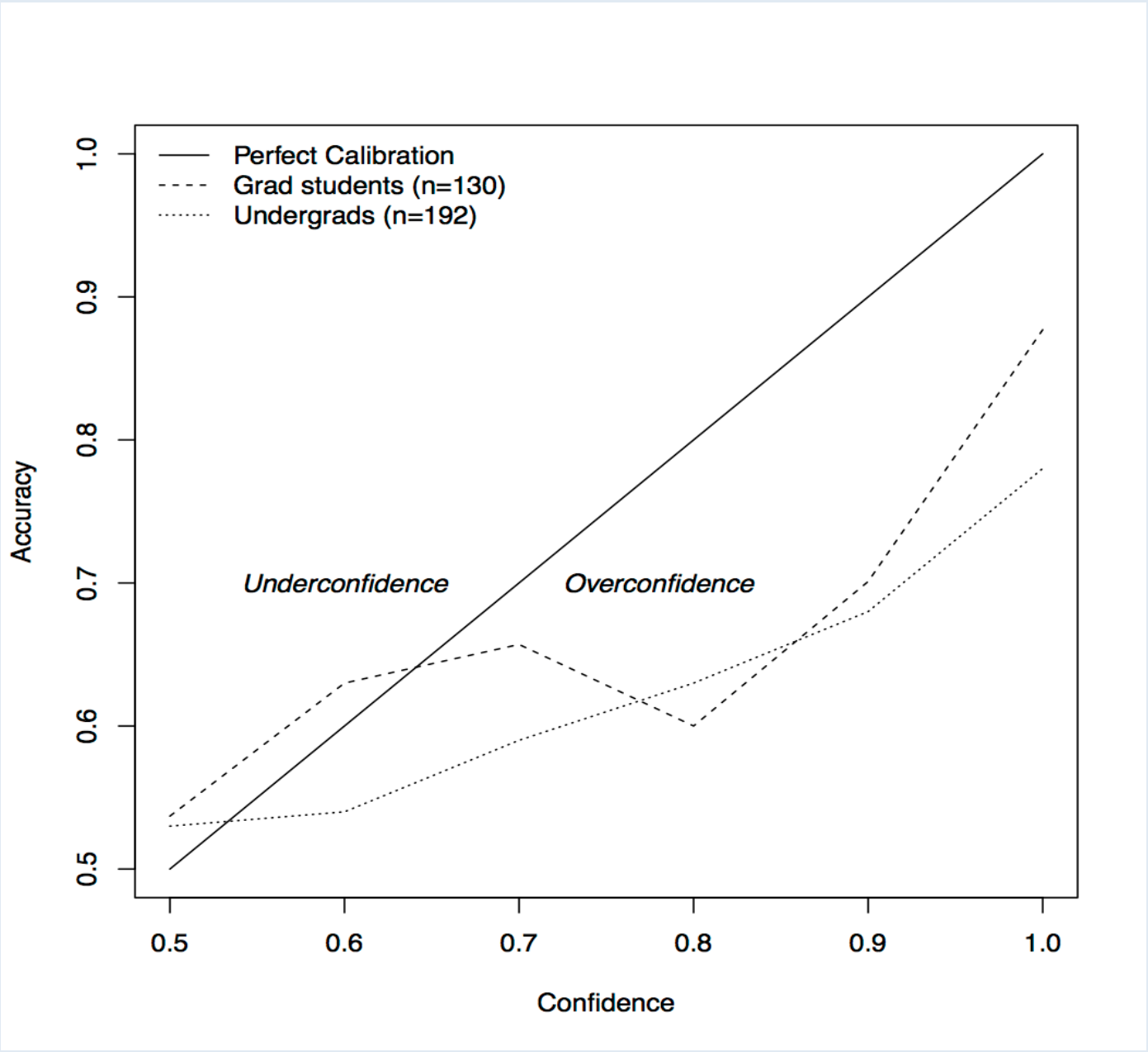


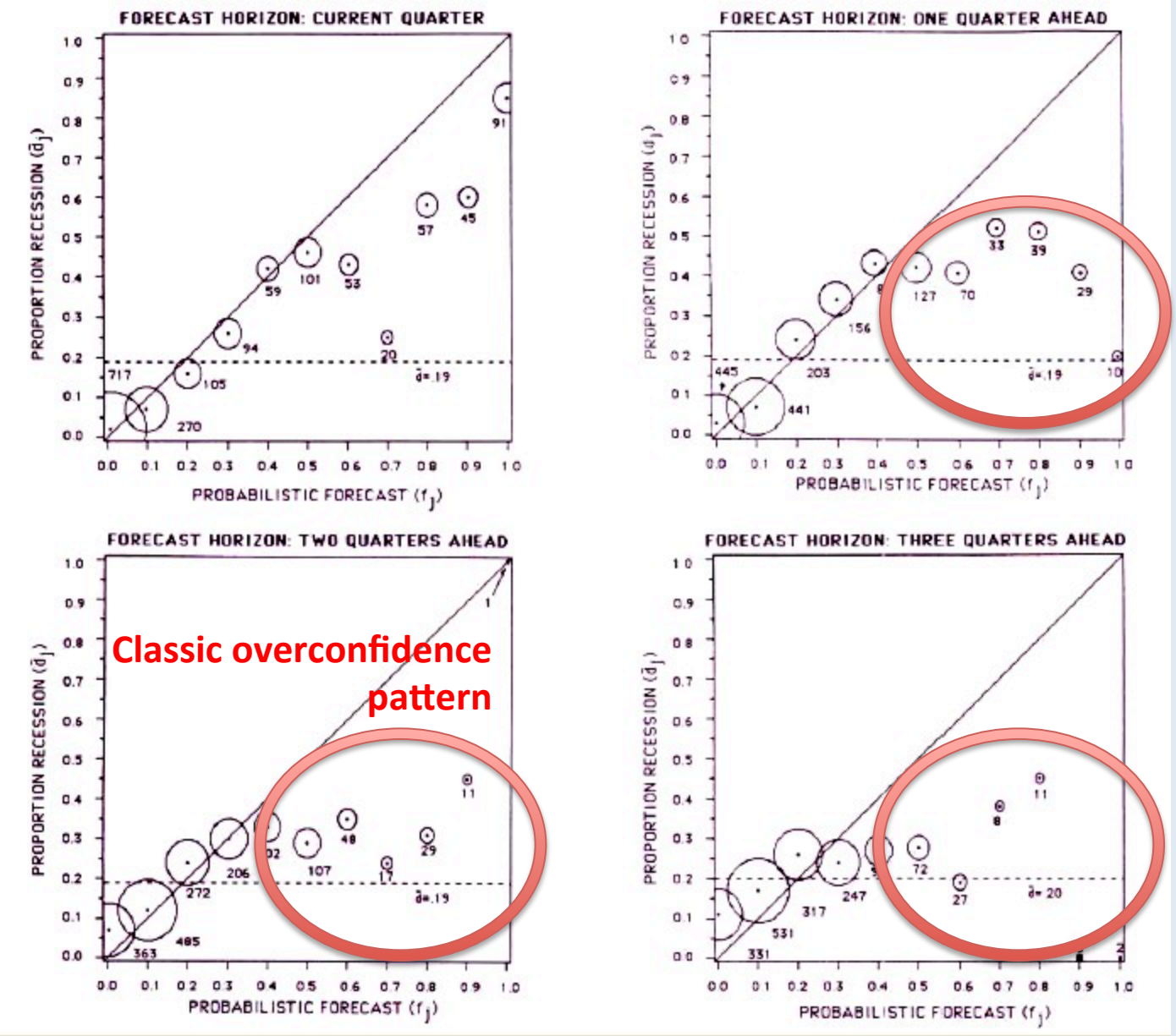
Interpreting Dutch Cannabis Policy: Reasoning by Analogy in the Legalization Debate

Robert MacCoun and Peter Reuter

Expert overconfidence (Tetlock, 1998)

- Predictions made by professional foreign policy experts were “only slightly more accurate than one would expect from chance.”
- Average confidence rating across domains (where 9 = maximum confidence):
  - *Experts who were right:* 6.5 to 7.6
  - *Experts who were wrong:* 6.3 to 7.1

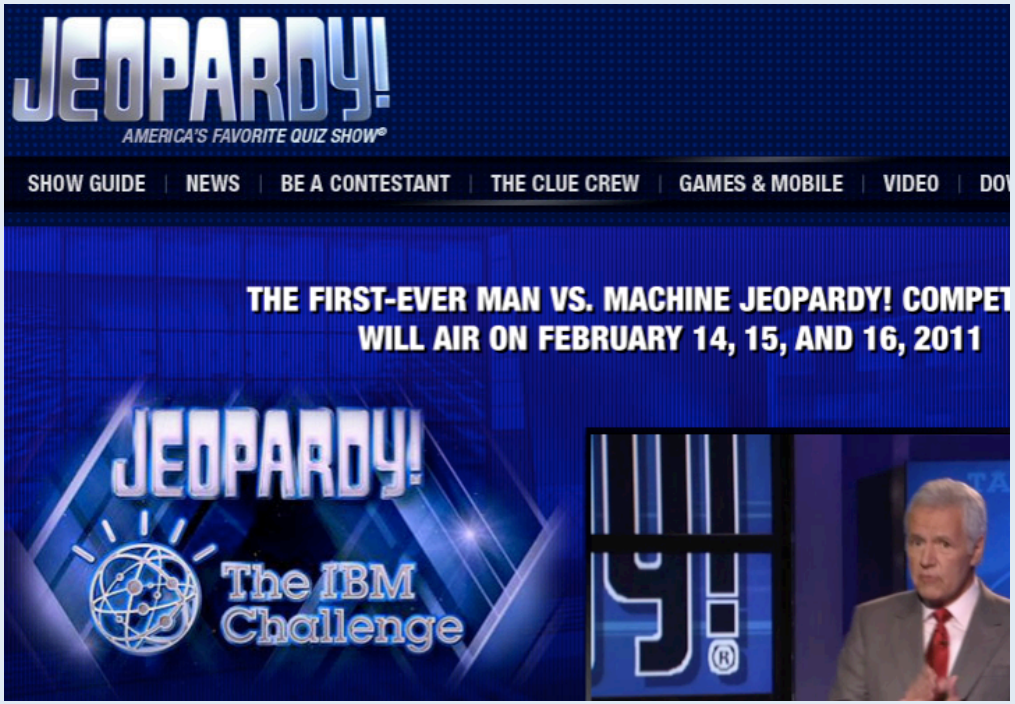

Perfect Calibration
Grad students (n=130)
Undergrads (n=192)
Underconfidence
Overconfidence
Accuracy
Confidence
0.5
0.6
0.7
0.8
0.9
1.0



Braun & Yaniv, 1992

Survey panel of **professional economic forecasters** conducted quarterly by NBER and the Am. Statistical Assoc.

To play "Jeopardy!", Watson…builds an evidence profile to determine what are the most likely correct answers. Each answer has a confidence level.



For some questions, one answer will have a high confidence level. This is when Watson is most likely to buzz in. **For other questions, none of the answers will have a high confidence level and Watson will not buzz in.**

“That’s the interesting thing,” McQueeney said. “The machine knows when it doesn’t know the answer.”

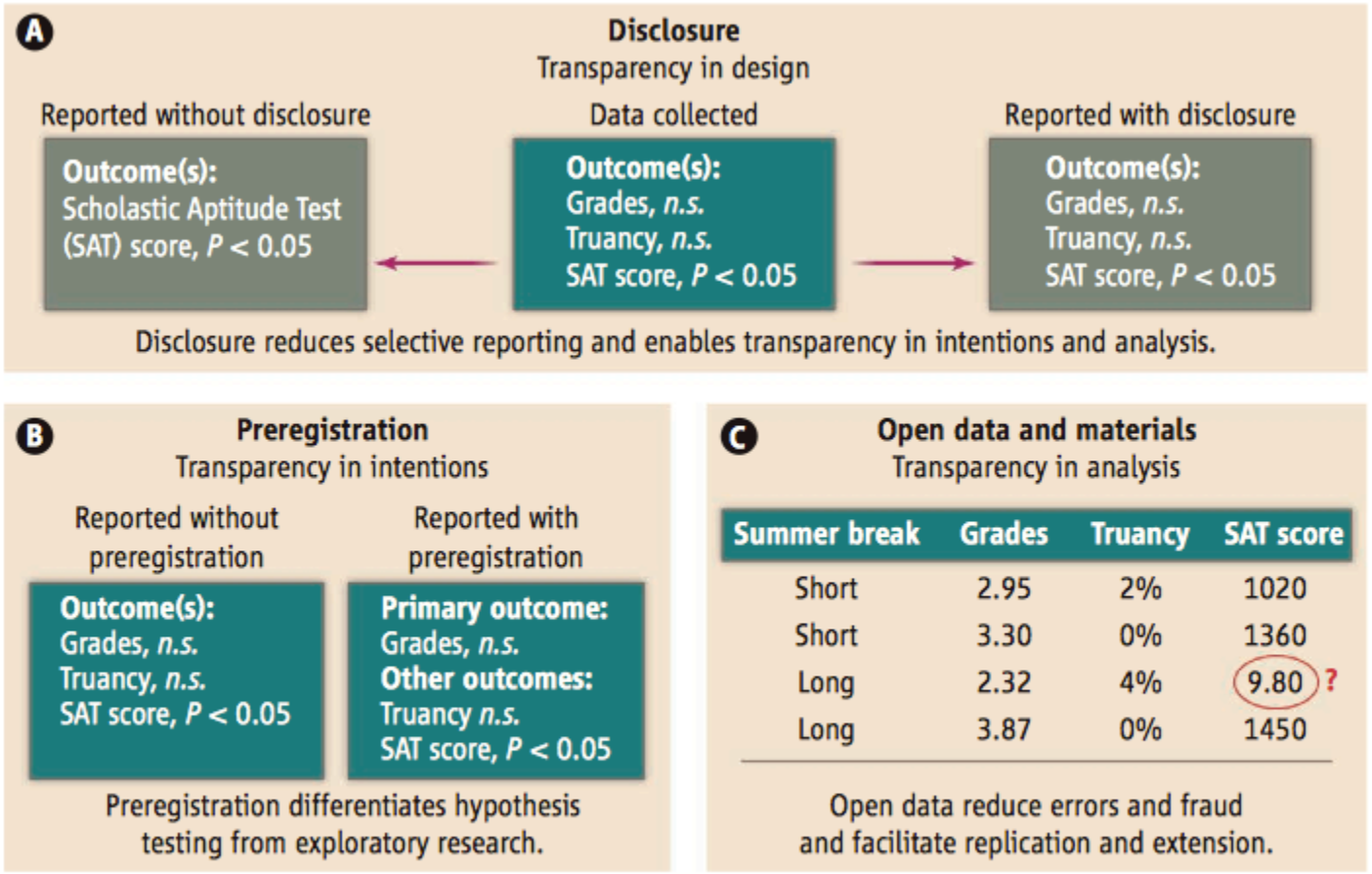
Corrective Practices

- Traditional practices
  - *peer reviewing*
  - *occasional "hired gun" replication*
  - *expert panels*
- New practices
  - *"Open science"*
  - *Preregistration of hypotheses*
  - *Routine "many labs" replication*
  - *Blind analysis*

Promoting Transparency in Social Science Research

3 JANUARY 2014 VOL 343 SCIENCE

E. Miguel,[1]* C. Camerer,[2] K. Casey,[3] J. Cohen,[3] K. M. Esterling,[4] A. Gerber,[5] R. Glennerster,[6] D. P. Green,[7] M. Humphreys,[7] G. Imbens,[3] D. Laitin,[3] T. Madon,[1] L. Nelson,[1] B. A. Nosek,[8,9] M. Petersen,[1] R. Sedlmayr,[10] J. P. Simmons,[11] U. Simonsohn,[11] M. Van der Laan[1]



Three mechanisms for increasing transparency in scientific reporting. Demonstrated with a research question: "Do shorter summer breaks improve educational outcomes?" n.s. denotes $P > 0.05$.

8 OCTOBER 2015 | VOL 526 | NATURE

Hide results to seek the truth

More fields should, like particle physics, adopt blind analysis to thwart bias, urge **Robert MacCoun** and **Saul Perlmutter**.

BLINDING STRATEGIES

Technique examples	Perturbation	Potential application
Noising $\theta_{ij}=y_{ij}+n_{ij}$ or $\theta_{ij}=\beta_k+n_{ij}$	Add a random number (from an appropriate statistical distribution) to data points or model parameters.	Testing which of several prevention messages is most effective in reducing smoking.
Biasing $\theta_{ij}=y_{ij}+b_j$	Obscure differences in experimental conditions by adding a hidden value that is biased in a particular direction.	Estimating whether the costs of a controversial safety regulation exceed its benefits.
Cell scrambling $\theta_{ij}=y_{\#}$	Shuffle labels for experimental conditions, so that it is unclear which set of results matches which conditions.	Testing a prediction that hard-copy books are better comprehended than audiobooks.
Item scrambling $\theta_{ij}=y_{\#\#}$	Randomly relabel each data point to de-identify experimental conditions.	Analysing group differences that might be easy to recognize even with noise and bias (for example, effects of neighbourhood and school on crime victimization).
Various combinations	Row scrambling: keep pairs of variables together to preserve correlation. Variable blinding: swap labels of various variables.	

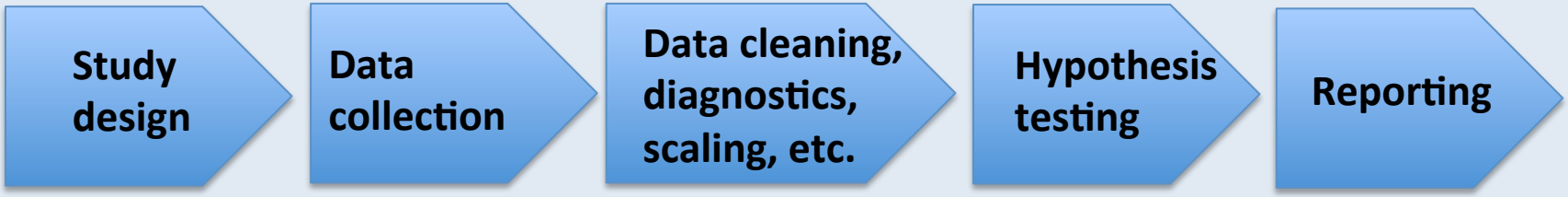
y_{ij} is the ith observation in the jth condition ('cell') of the study; β_k is the kth parameter of a model; θ_{ij} is y_{ij} or β_k after blinding; n_{ij} is random error, b_j is a bias term, and # denotes a randomly swapped subscript.

Different methods blind different features of the data

% of simulations where feature was <u>undetectable:</u>

	Add noise	Add bias	Scrambled cases	Scrambled cells
Range	100	100	0	0
Effect size	100	100	100	0
Exact p value	100	98	100	0
Stat. signif.	36	38	66	0
Direction	6	39	50	49

When should we blind?

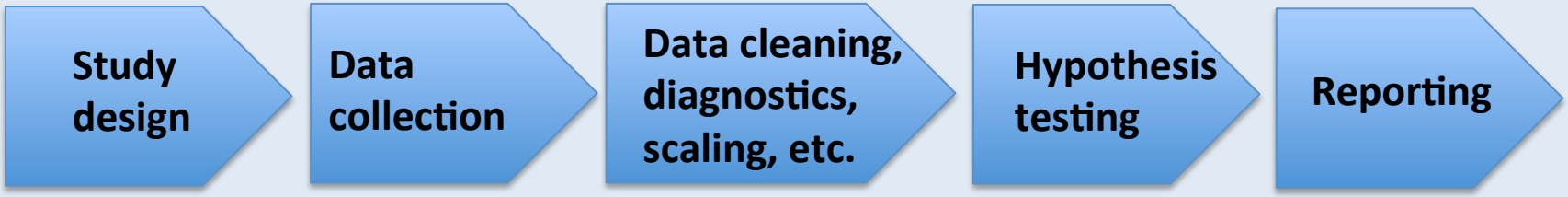


When should we blind?

Key point to remember: The computer isn't blinded – you are.

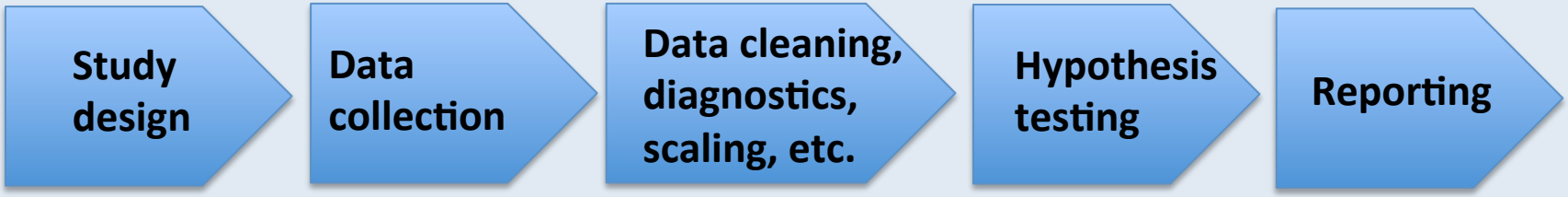
The blinding can be limited to what's displayed rather than what's being analyzed.

When should we blind?



Or we can choose not to use blinding at all.

When should we blind?



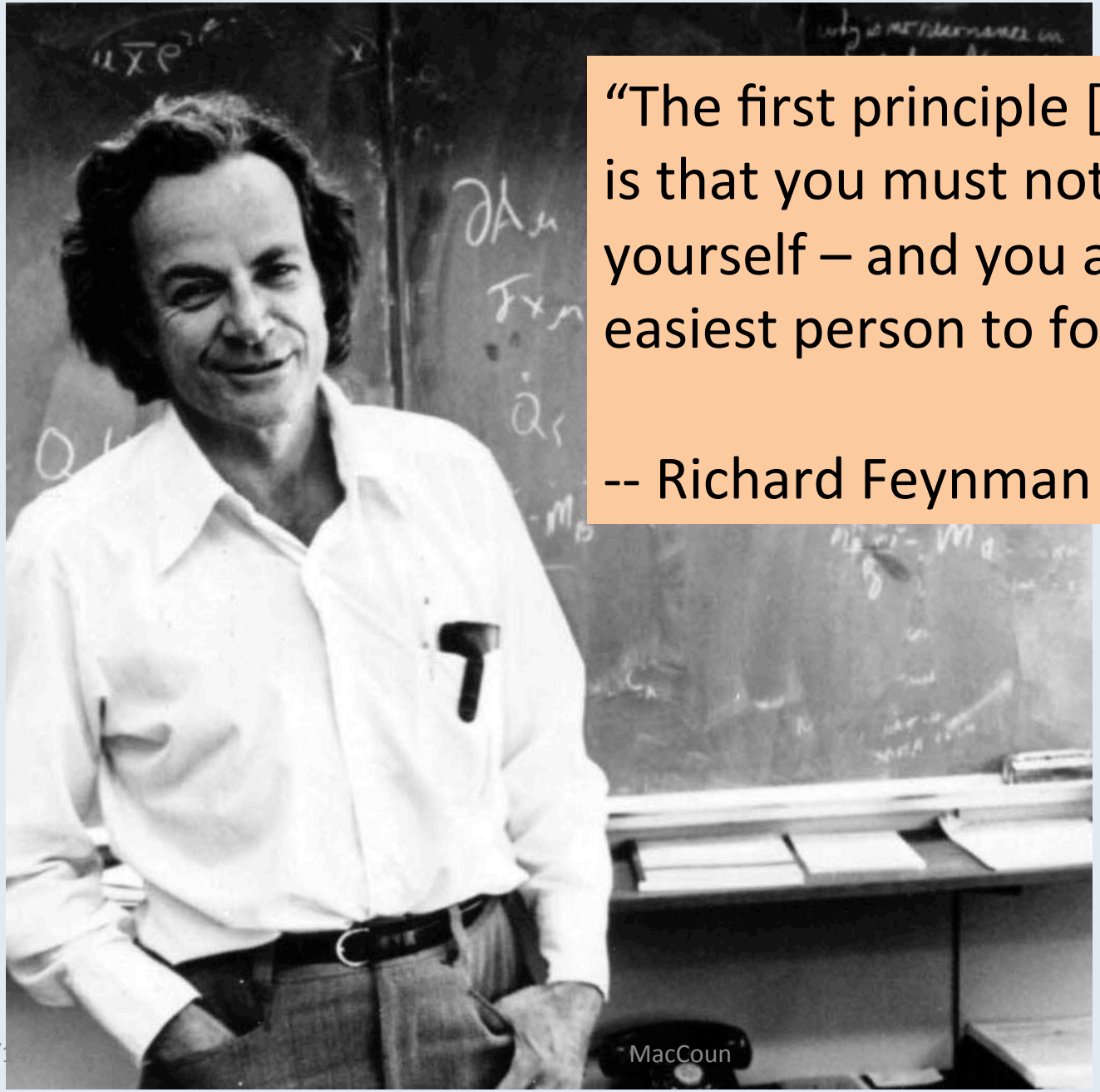
Or we can choose not to use blinding at all….

...but maybe we should start saying so:

“The data analysis was performed without blinding...”

Blind analysis should be tested empirically

- Benefits > costs?
- Comparison to other methods
  - Peer review, cross-validation, pre-registered analyses, cross-lab replication
- Does it impair discovery? Does it facilitate it?



"The first principle [of science] is that you must not fool yourself – and you are the easiest person to fool."

-- Richard Feynman

Forbes

A Poker Player Explains Why You Make Bad Business Decisions

What she just described is "outcome blind" analysis of decisions. Duke insists that all great poker players think this way. In fact she says she had a small peer group of elite players who regularly got together and talked shop. "When we told each other 'the story of a hand,' we wouldn't mention whether we won or lost. It was all about the behaviors, the variables, the decisions – not the outcome."

(Image courtesy of Annie Duke)

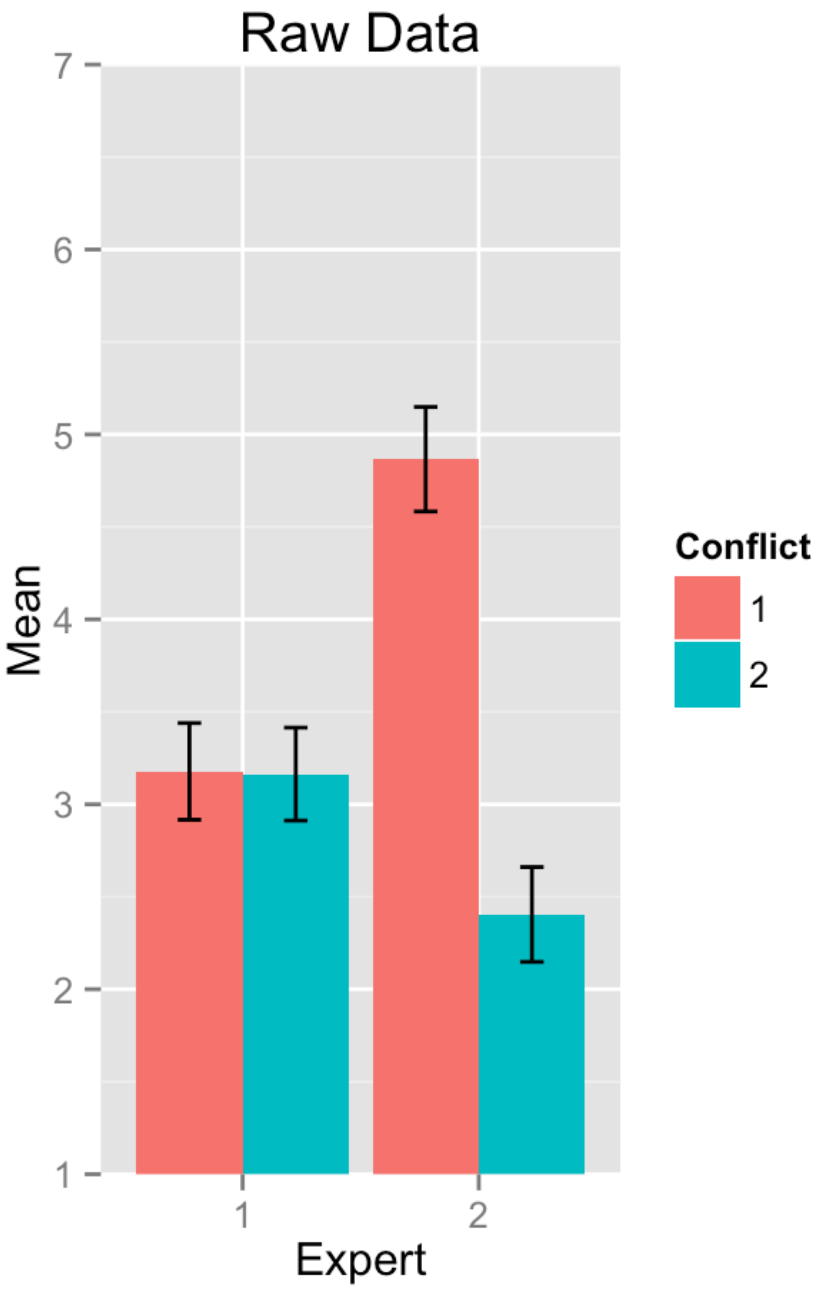
UNUSED SLIDES BELOW…

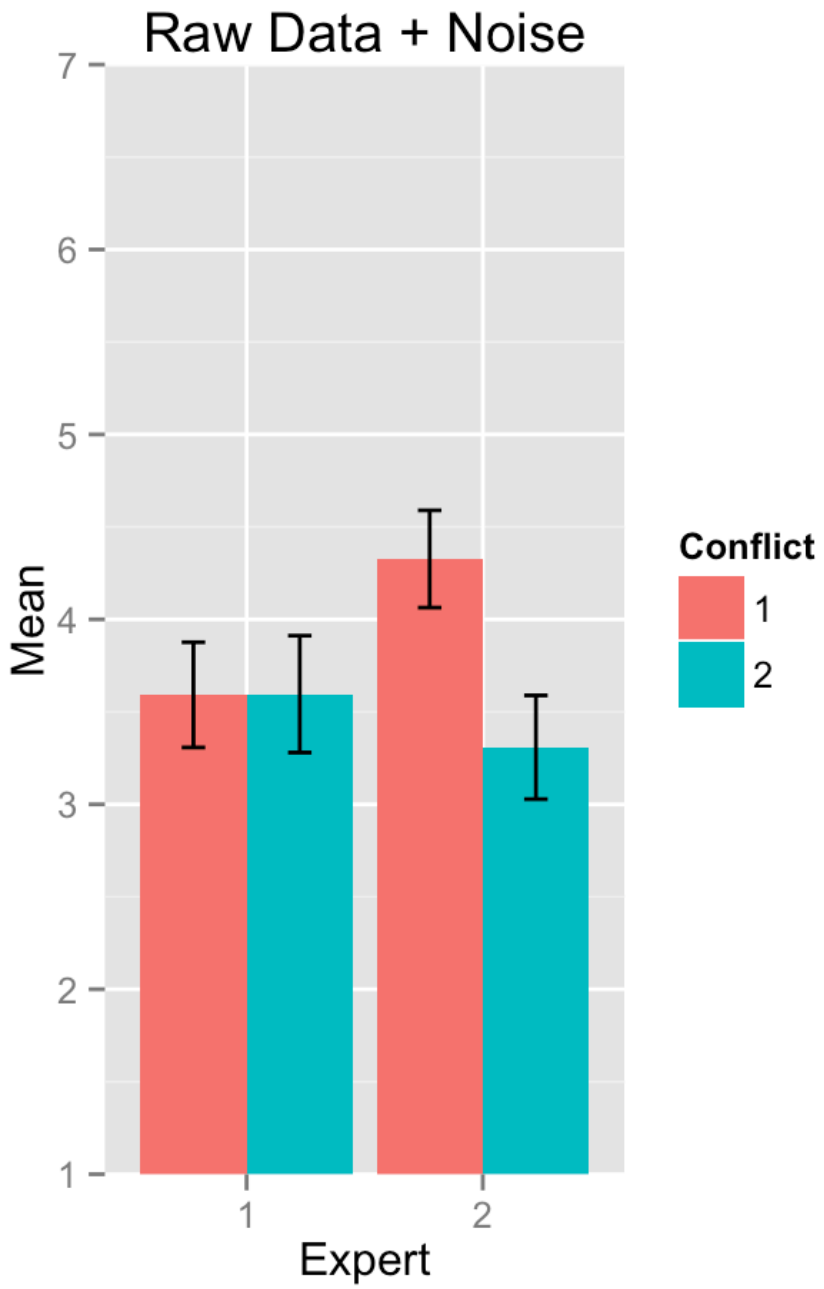
Canonical psychology experiment

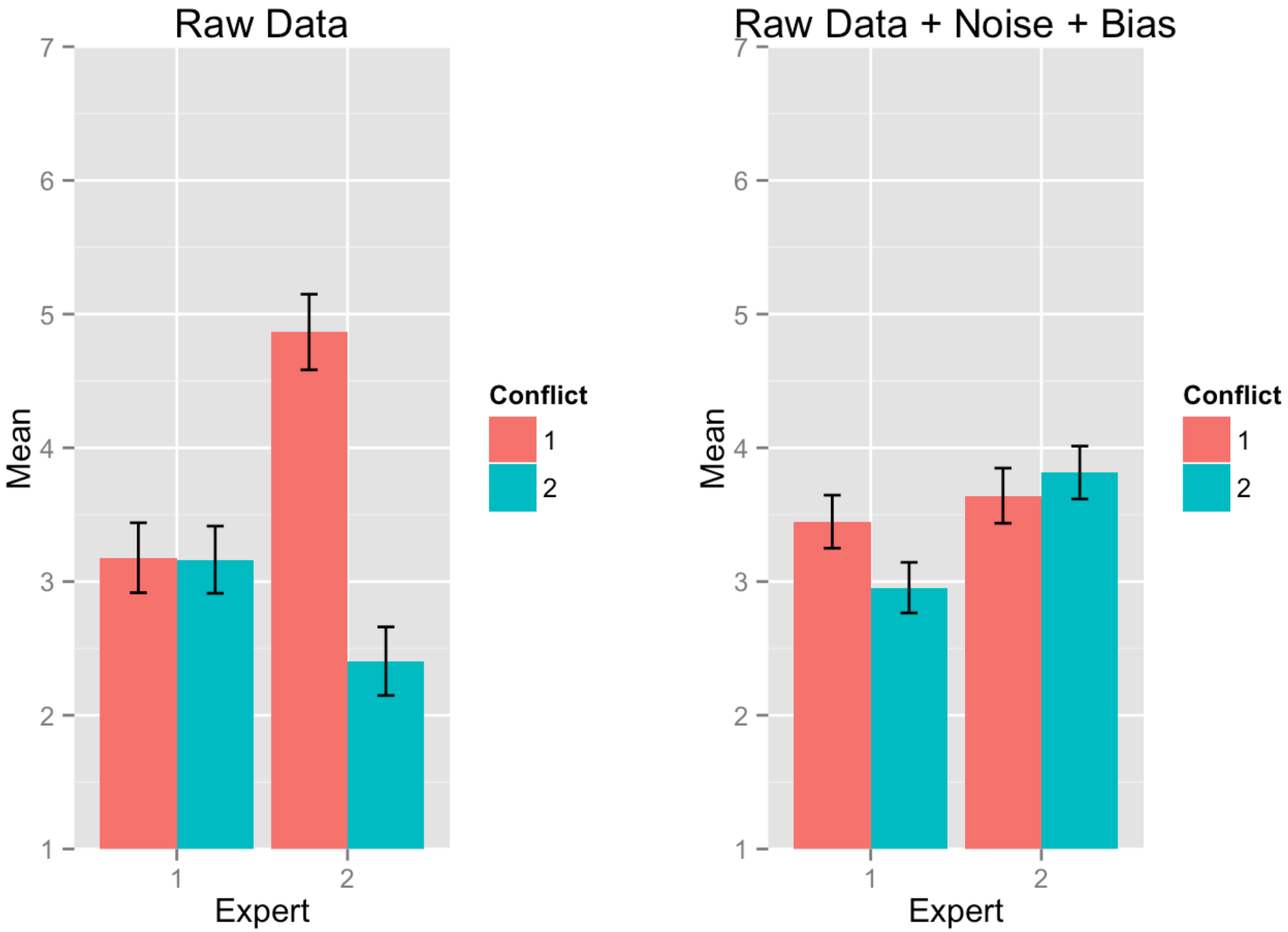
- Advertisement urging people to vote for wetlands protection ballot initiative
- 2 x 2 factorial experiment (50 subjects per cell):
  - SOURCE EXPERTISE: 1=Low (source has a BA in Biology) vs. 2=High (source is a Yale Professor of Biology)
  - CONFLICT OF INTEREST: 3=None, 4=Yale will get new wetlands science center if initiative passes

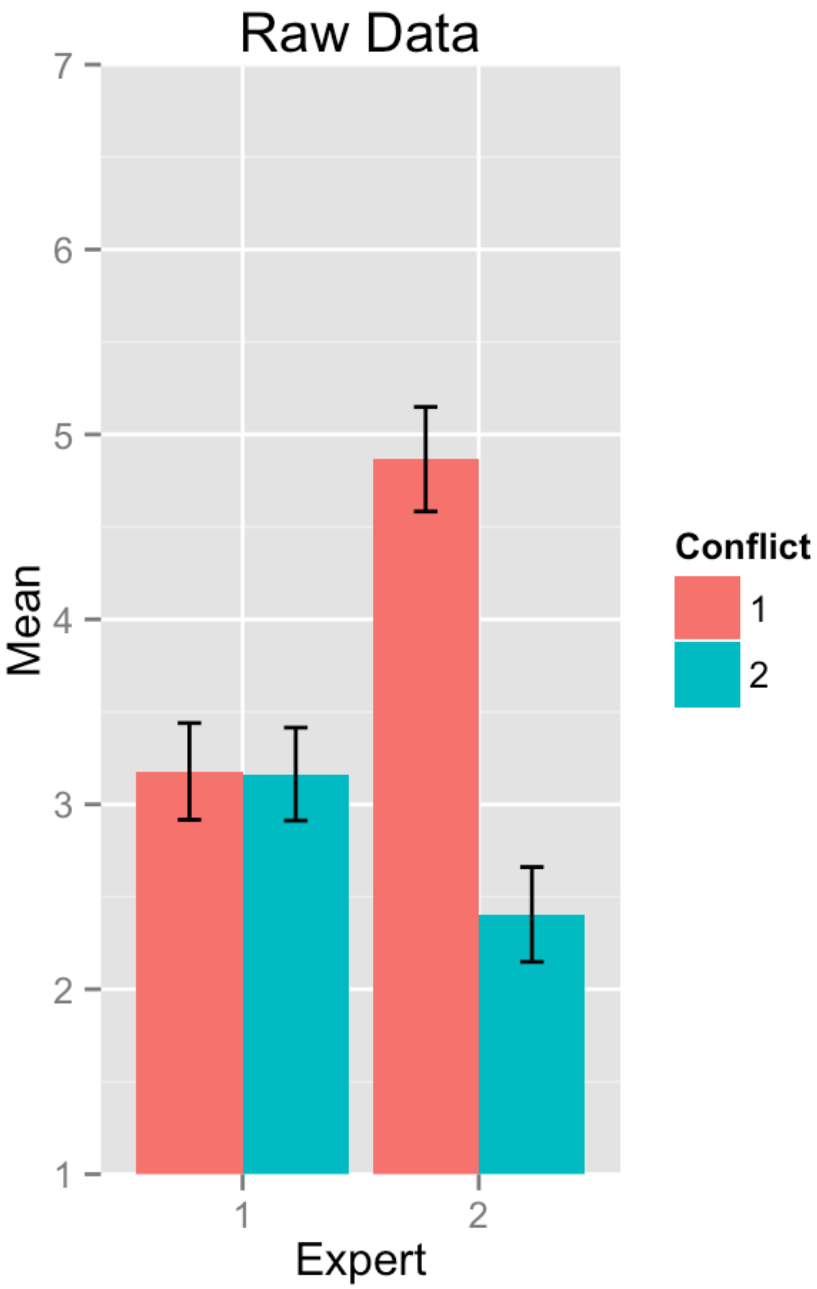
To simulate raw data, we sampled from normal distribution:

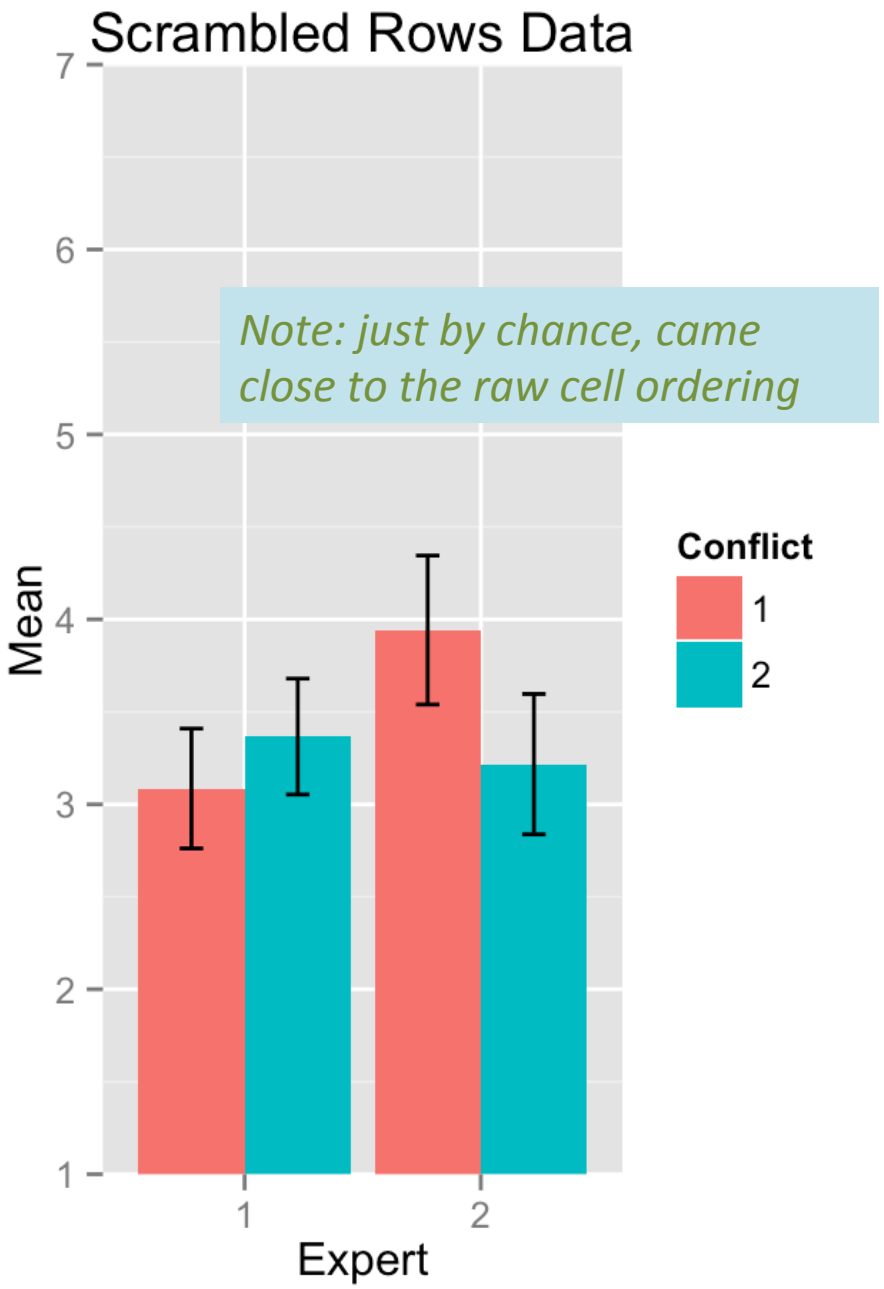
```
raw.DV00 <- rnorm(n=50,mean=3, sd=1) # Low Expert, No Conflict

raw.DV01 <- rnorm(n=50,mean=3, sd=1) # Low Expert, Conflict

raw.DV10 <- rnorm(n=50,mean=4.5, sd=1) # High Expert, No Conflict

raw.DV11 <- rnorm(n=50,mean=2.5, sd=1) # High Expert, Conflict
```


Raw Data
Mean
Expert
Conflict
1
2
7
6
5
4
3
2
1

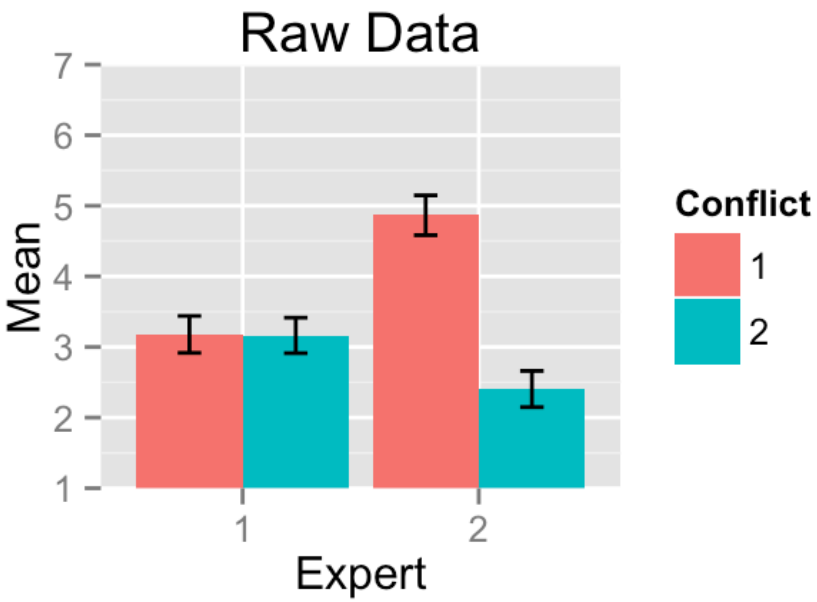

Raw Data + Noise
Mean
Expert
Conflict
1
2
7
6
5
4
3
2
1

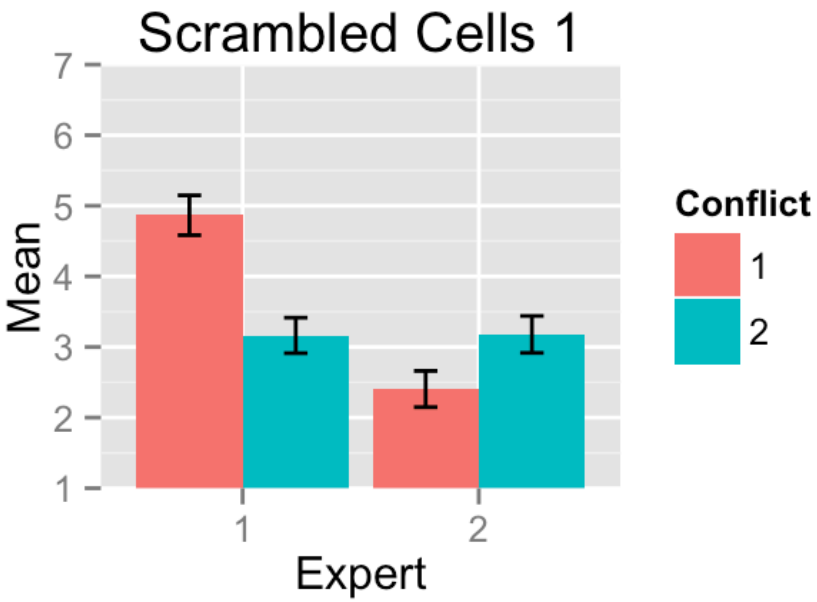

Raw Data
Mean
Expert
Conflict
1
2
Raw Data + Noise + Bias
Mean
Expert
Conflict
1
2

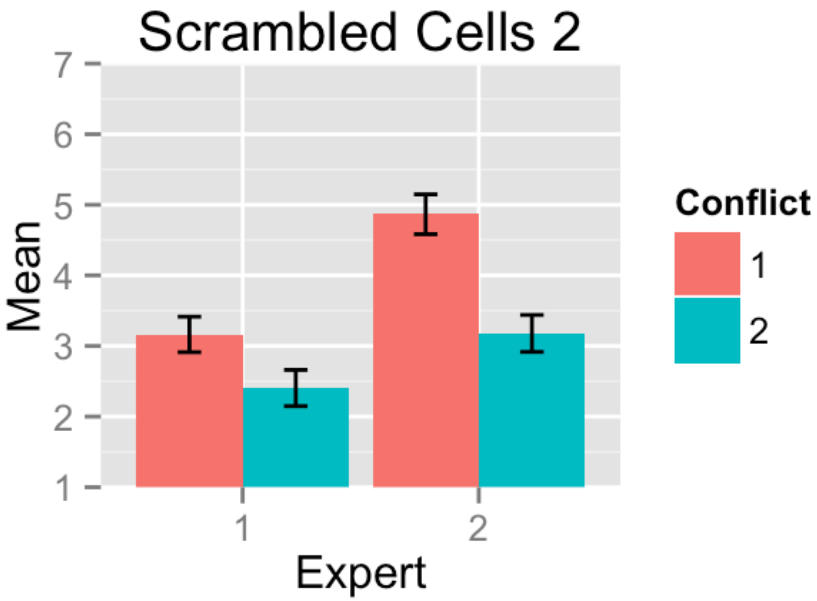


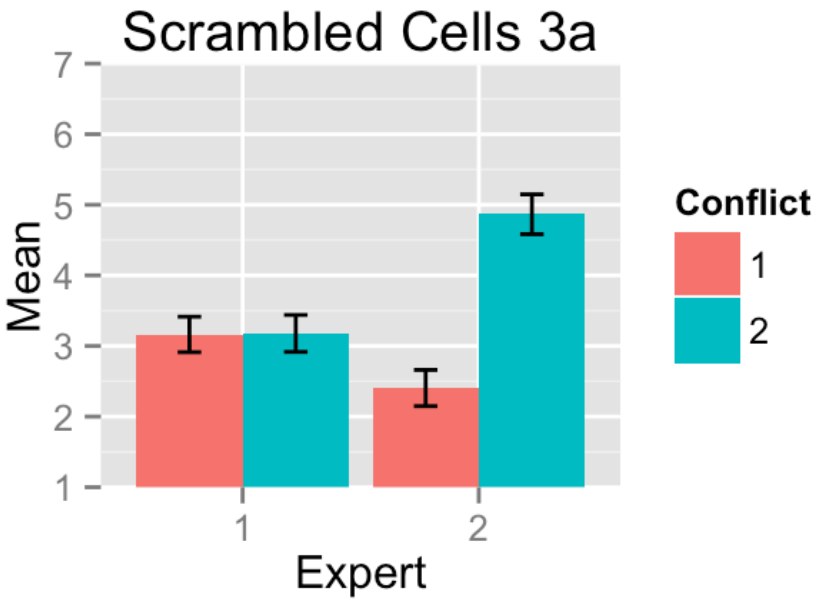


Note: just by chance, came close to the raw cell ordering

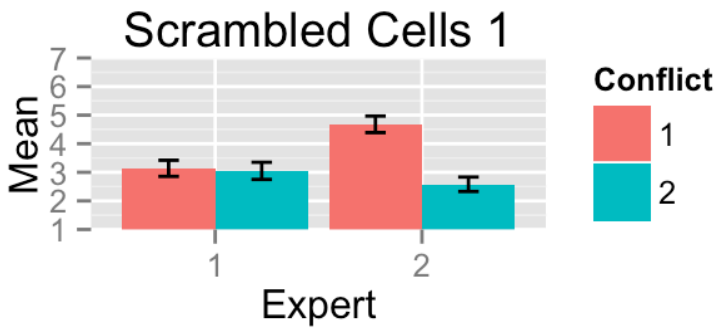

Raw Data
Mean
Expert
Conflict
1
2

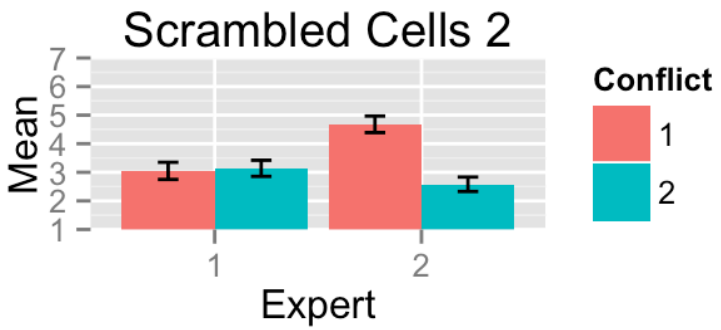

Scrambled Cells 1
Mean
Expert
Conflict
1
2

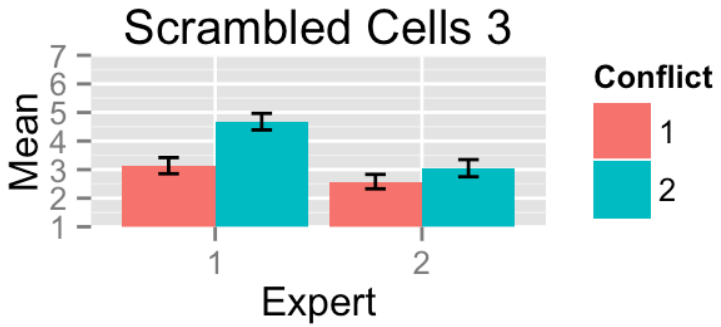

Scrambled Cells 2
Mean
Expert
Conflict
1
2

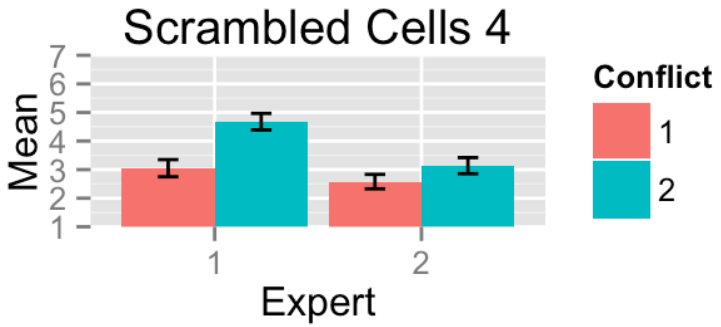

Scrambled Cells 3a
Mean
Expert
Conflict
1
2

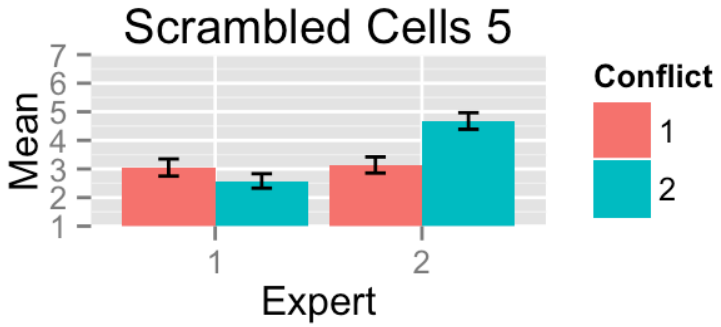
Panel of 6 outcome sets, one of which is true….

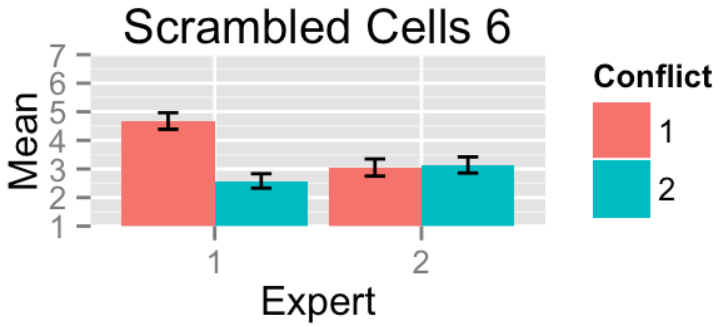












Differences between physics and psychology research

- Physics has far higher degree of precision
- Physics has strong theory predicting quantities, and/or theoretical implications that can differ dramatically as a function of measured quantities
- Physics tends to have large research teams
- Physics data are sometimes sparse, difficult to obtain
- Random assignment to condition often impossible

2 x 2 ANOVA

	df	SS	MS	F	p
Expert	1	10.8	10.8	12.6	<.001
Conflict	1	76.6	76.6	89.3	<.001
E*C	1	74.9	74.9	87.3	<.001
Residual	196	168.1	0.9		

Blinding method 1: Add noise

- 200 random numbers sampled from uniform (min=1,max=7)
- blind.DV = mean(raw.DV, randnum)

Raw data

	df	SS	MS	F	p
Expert	1	10.8	10.8	12.6	<.001
Conflict	1	76.6	76.6	89.3	<.001
E*C	1	74.9	74.9	87.3	<.001
Residual	196	168.1	0.9		

Blind data (raw + noise)

	df	SS	MS	F	p
Expert	1	2.5	2.5	2.5	0.119
Conflict	1	12.9	12.9	12.6	<.001
E*C	1	13.1	13.1	12.8	<.001
Residual	196	199.4	1.0		

Blinding method 2: Add noise + bias

- Same vector of random numbers as method 1
- Average together with a cell-specific bias term
  - Four bias terms, each sampled from normal(mean=3, sd=1)

Raw data

	df	SS	MS	F	p
Expert	1	10.8	10.8	12.6	<.001
Conflict	1	76.6	76.6	89.3	<.001
E*C	1	74.9	74.9	87.3	<.001
Residual	196	168.1	0.9		

Blind data (raw + noise + bias)

	df	SS	MS	F	p
Expert	1	13.9	13.9	28.7	<.001
Conflict	1	1.3	1.3	2.7	0.105
E*C	1	5.6	5.6	11.5	<.001
Residual	196	95.1	0.5		

Blinding method 3: Row scrambling

- There are n=200 rows in the data matrix
- “Row scrambling” simply randomly sorts their dependent variables, keeping the independent variables the same

Raw data

	df	SS	MS	F	p
Expert	1	10.8	10.8	12.6	<.001
Conflict	1	76.6	76.6	89.3	<.001
E*C	1	74.9	74.9	87.3	<.001
Residual	196	168.1	0.9		

Blind data (row scrambling)

	df	SS	MS	F	p
Expert	1	6.3	6.3	4.0	<.05
Conflict	1	2.5	2.5	1.6	0.21
E*C	1	12.7	12.7	8.0	<.01
Residual	196	309.1	1.6		

Comment on first three methods

- All three are *regressive* in the sense that the differences between conditions get weakened
- Problem:
  - *If data are noisy enough, no different than pre-registered data analysis (in the absence of any data)*
  - *Retaining more of the actual differences stimulates more thinking about the data*
  - *How to stimulate creativity without stimulating confirmatory bias?*

Blinding method 4: Cell scrambling

- Rather than scrambling individual data points, this method keeps cell data intact, but scrambles the four cells of the design
- In this case, there are 4! = 24 possible orderings, and we've randomly sampled three of them…
- In all 3 cases, the main effects and interaction are significant ($p<.001$), but the interpretation differs…

Analytic Method	What Could Be Blinded?
Analysis of variance for randomized controlled trials	• Outcome values • Statistical significance • Experimental condition
Ordinary least-square regression analysis	• Outcome values • Statistical significance • Variable labels
Factor analysis	• Outcome values • Statistical significance • Variable labels
Time series analysis	• Outcome values • Statistical significance • Event dates

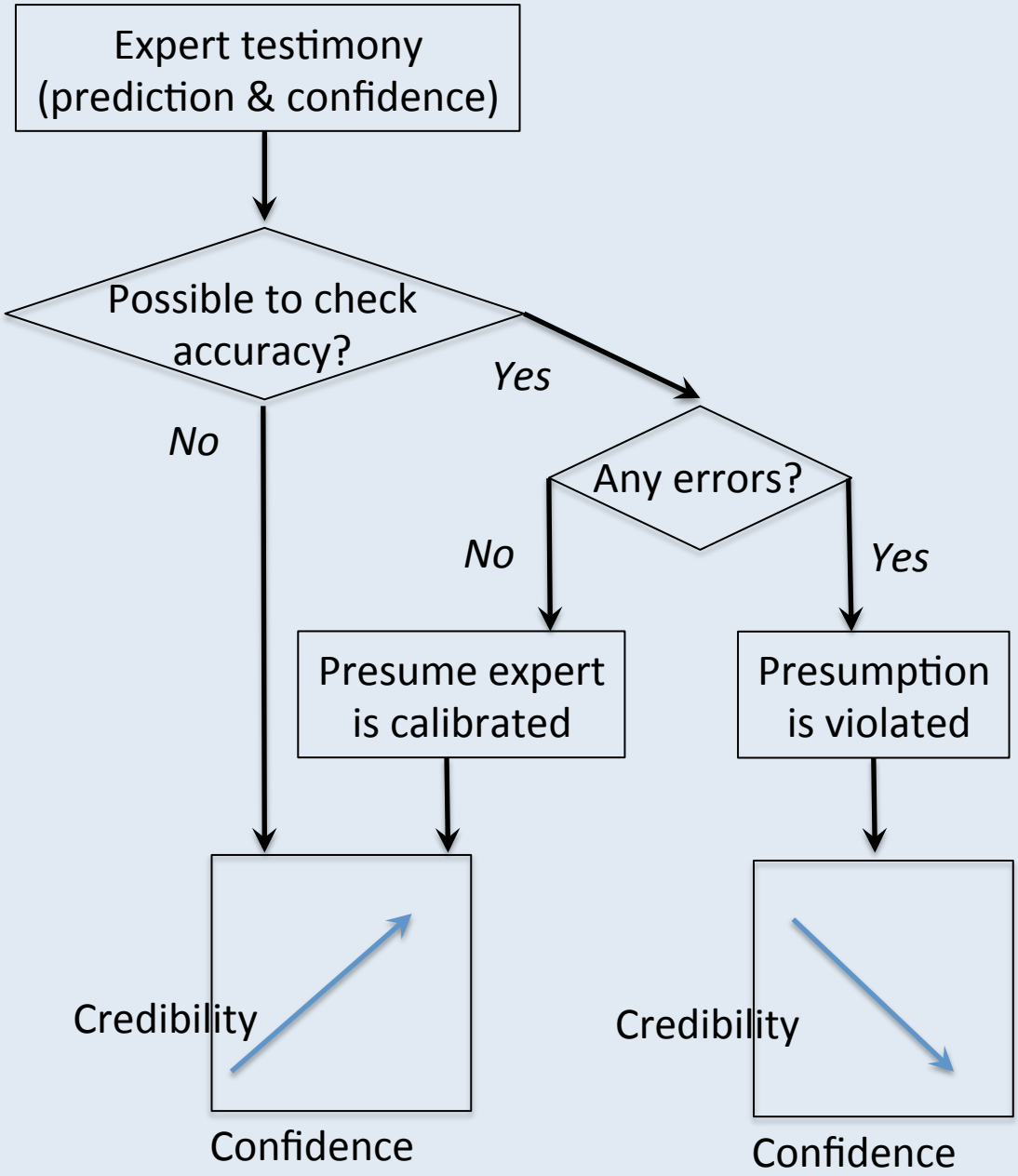

Expert testimony
(prediction & confidence)
Possible to check
accuracy?
Yes
No
Any errors?
No
Yes
Presume expert
is calibrated
Presumption
is violated
Credibility
Confidence
Credibility
Confidence

Table 1. Citizens' motives and their perceptions of expert motives.

		Perceived expert motive	
		Inquisitorial	*Adversarial*
Citizen's motives	*Inquisitorial*	Who is more likely to be correct?	Who is more honest?
	Adversarial	Does this help or hurt our side?	

Political Psychology

Political Psychology, Vol. 30, No. 1, 2009

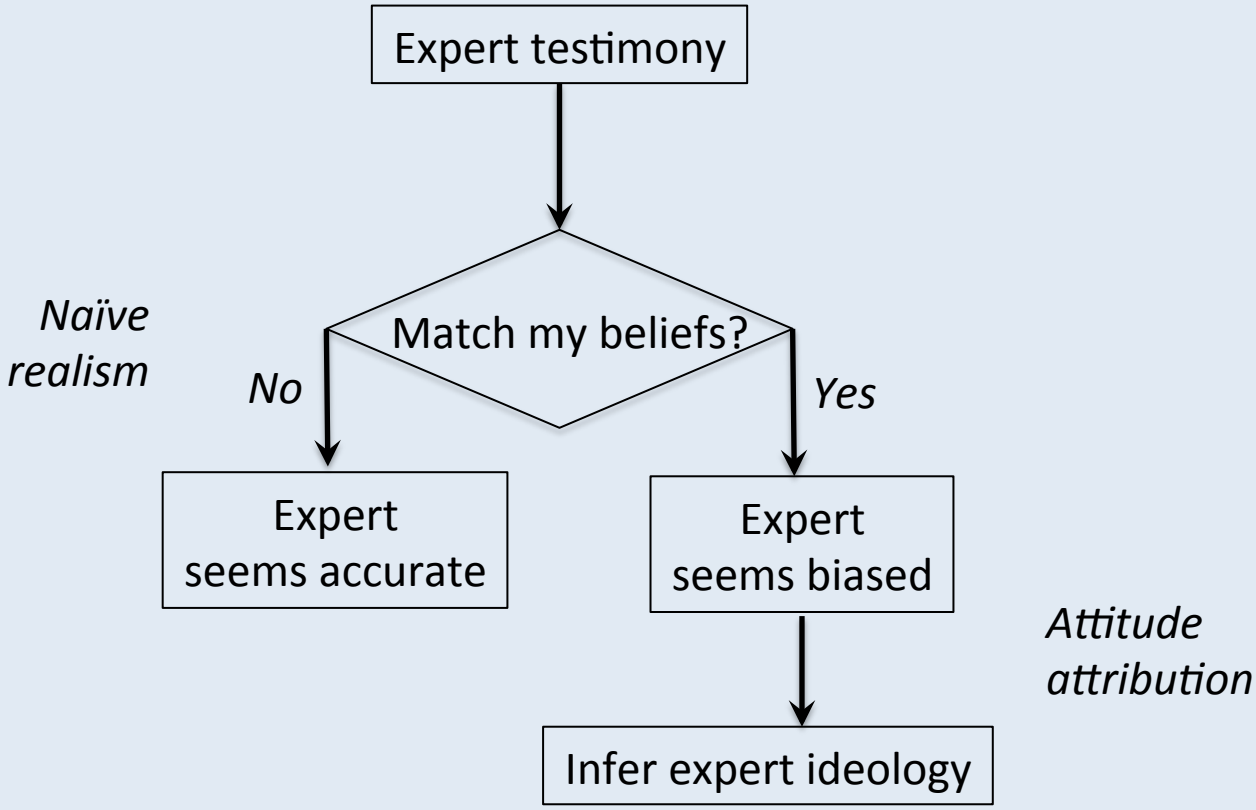
Citizens' Perceptions of Ideological Bias in Research on Public Policy Controversies

Robert J. MacCoun
University of California, Berkeley

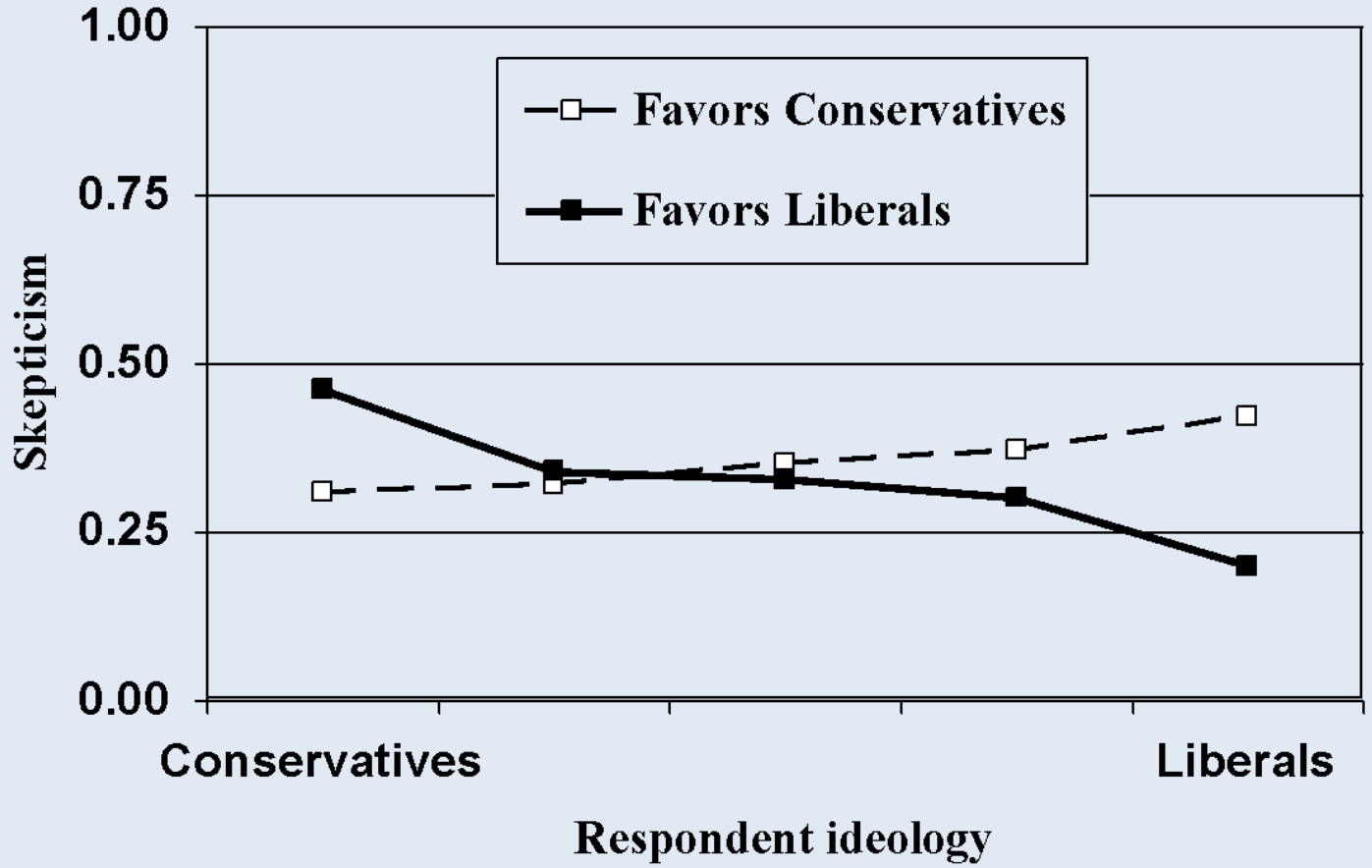
Susannah Paletz
University of California, Berkeley

Table 1. Research Design and Cell Sizes

	Finding favors conservative position	Finding favors liberal position
Gun control	Fails to reduce crime (n = 120)	Reduces crime (n = 104)
Death penalty	Deters crime (n = 90)	Fails to deter crime (n = 122)
Medical marijuana	No medical value (n = 93)	Has medical value (n = 88)
School vouchers	Improves education (n = 124)	Fails to improve education (n = 105)
Nutrition ads	TV Actor (n = 102) vs. Real Doctor (n = 102)	


Expert testimony
Match my beliefs?
No
Yes
Naïve realism
Expert seems accurate
Expert seems biased
Attitude attribution
Infer expert ideology

How doubtful is result?



How liberal is the researcher?