Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling

Leslie John *(Harvard Business School)*

Presenting work done in collaboration with:
George Loewenstein *(Carnegie Mellon)*
Drazen Prelec *(MIT)*

Prominent cases of research fraud

How one man got away with mass fraud by saying 'trust me, it's science'

Resveratrol researcher faked data, report says; what drives academic fraud?

Scientist Under Inquiry Resigns From Harvard

Researcher Faked Evidence of Human Cloning, Koreans Report

Published: July 7, 2010

'Lying Dutchman' Could Cast Doubt on Ad Research

Suspect Study Raises Questions About What Should Be Done to Prevent Abuses

This paper is not about these clear-cut cases of fraud.

Questionable research practices (QRPs)

- The “grey zone” of acceptable practice
- Practices that are sometimes justified, but often not
- Provide considerable latitude for rationalization
- Can increase false positives (Simmons, Nelson, & Simonsohn, 2011)
- QRPs might be surprisingly common

Goals of this project

1. Estimate the prevalence of QRPs among psychologists
2. Test the effect of providing truth-telling incentives on admission rates

Procedure

- Emailed faculty in U.S. research-oriented psychology departments; respondents asked about:
  - prevalence of QRPs (in various ways; more on next slide)
  - defensibility of QRPs
  - whether they had doubts about research integrity
  - demographic questions
- Anonymity and participation tracking
- 2 conditions:
  - Control: donation based on response rate
  - Truth-incentivized: donation based on truthfulness
    - Based on Bayesian Truth Serum (Prelec, 2004)

Multiple measures of QRP prevalence

- For each of 10 QRPs, Ss anonymously indicated:
  1. Whether they had engaged in the practice (yes/no)
     ➔ *Measure #1: admission rate*
  2. The % of psychologists that have engaged in the practice
     ➔ *Measure #2: prevalence estimate*
  3. Admission estimate: among psychologists who engaged in practice, % who would admit to having done so
     ➔ *Measure #3: implied prevalence estimate: admission rate ÷ admission estimate*
     - Example:
       - On average, Ss think that 60% of people who have done the behavior will admit to it (admission estimate)
       - 40% of Ss admit to the behavior (admission rate)
       - Therefore, implied prevalence estimate is .40/.60 = 67%

Items

(order of presentation was randomized)

1. In a paper, failing to report all of a study's dependent measures.
2. Deciding whether to collect more data after looking to see whether the results were significant.
3. In a paper, failing to report all of a study's conditions.
4. Stopping collecting data earlier than planned because one found the result that one had been looking for.
5. In a paper, 'Rounding off' a p value (e.g. reporting that a p value of .054 is less than .05)
6. In a paper, selectively reporting studies that 'worked.'
7. Deciding whether to exclude data after looking at the impact of doing so on the results.
8. In a paper, reporting an unexpected finding as having been predicted from the start.
9. In a paper, claiming that results are unaffected by demographic variables (e.g. gender) when one is actually unsure (or knows that they do).
10. Falsifying data.

Response and completion rates

- Response rate: 36% (2,155 out of 5,964)
- Attrition rate: 33% (719 out of 2,155)
- Completed response rate: 24% (1,436 out of 5,964)

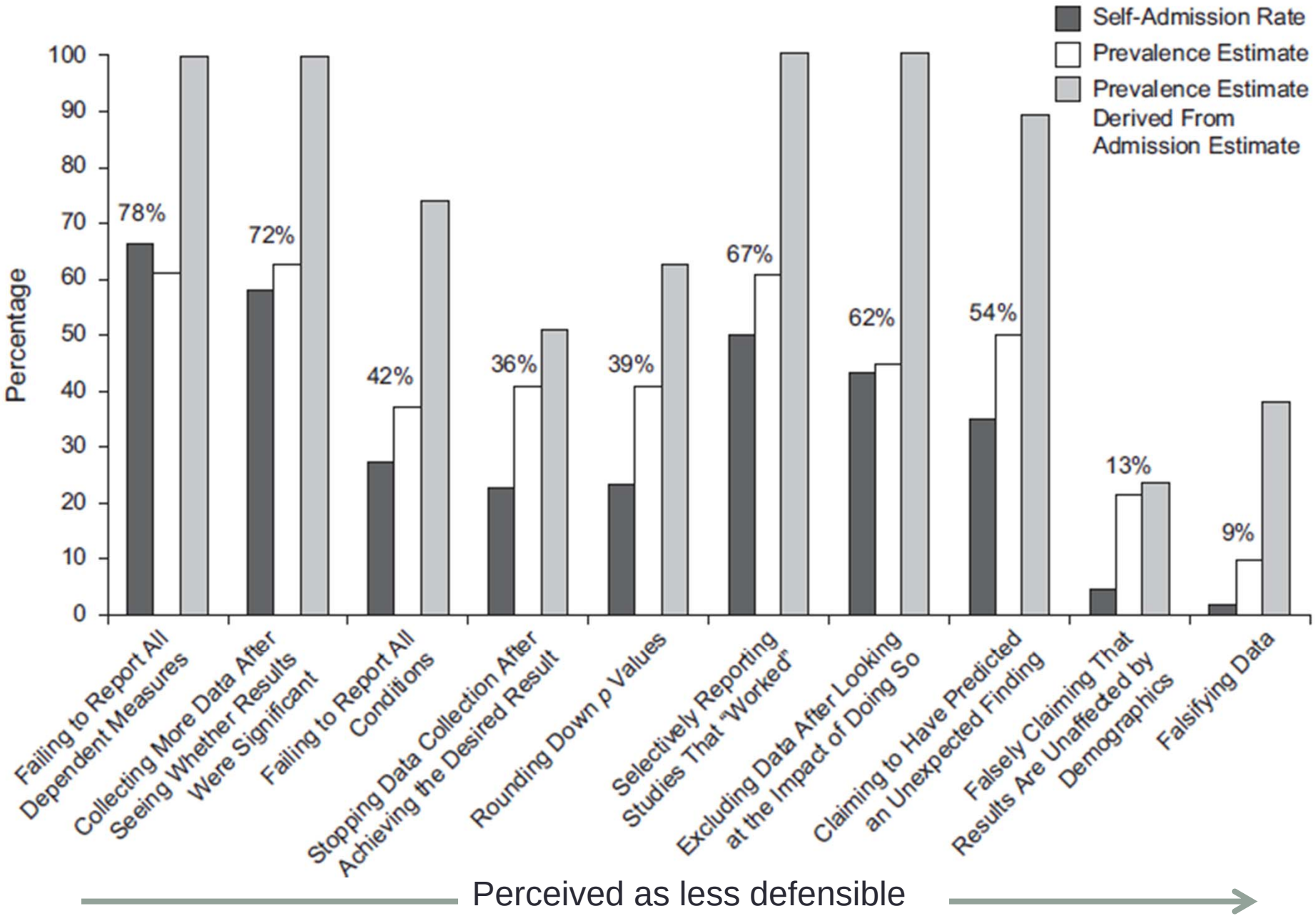
Presentation order randomized

Admission rates

	Control	BTS	Odds Ratio
Failing to report all DVs	63%	67%	1.1
Collecting more data after checking results	56%	58%	1.1
Failing to report all conditions	28%	27%	1.0
Stopping data collection after achieving desired result*	16%	23%	1.6
Rounding down p values	22%	23%	1.1
Selectively reporting studies that ‘worked’	46%	50%	1.2
Excluding data after looking at impact of doing so	38%	43%	1.2
Claiming to have predicted unexpected finding*	27%	35%	1.5
Falsely claiming results to be unaffected by demographics	3%	4%	1.5
Falsifying data	1%	2%	2.8

Admission rates

	Control	BTS	Defensible 0 = No 1 = Possibly 2 = Yes
Failing to report all DVs	63%	67%	1.8
Collecting more data after checking results	56%	58%	1.8
Failing to report all conditions	28%	27%	1.8
Stopping data collection after achieving desired result*	16%	23%	1.8
Rounding down p values	22%	23%	1.7
Selectively reporting studies that ‘worked’	46%	50%	1.7
Excluding data after looking at impact of doing so	38%	43%	1.6
Claiming to have predicted unexpected finding*	27%	35%	1.5
Falsely claiming results to be unaffected by demographics	3%	4%	1.3
Falsifying data	1%	2%	0.2


Self-Admission Rate
Prevalence Estimate
Prevalence Estimate Derived From Admission Estimate
Percentage
100
90
80
70
60
50
40
30
20
10
0
78%
72%
42%
36%
39%
67%
62%
54%
13%
9%
Failing to Report All Dependent Measures
Collecting More Data After Seeing Whether Results Were Significant
Failing to Report All Conditions
Stopping Data Collection After Achieving the Desired Result
Rounding Down p Values
Selectively Reporting Studies That "Worked"
Excluding Data After Looking at the Impact of Doing So
Claiming to Have Predicted an Unexpected Finding
Falsely Claiming That Results Are Unaffected by Demographics
Falsifying Data
Perceived as less defensible

Admission rates by sub-discipline

Discipline	Admission rate
Clinical	0.27*
Cognitive	0.37***
Developmental	0.31
Forensic	0.28
Health	0.30
Industrial Organizational	0.31
Neuro	0.35**
Personality	0.32
Social	0.40***

Significance codes:
*p<.05, **p<.01, ***p<.0005
For "Admission rate," significance codes are based on random effects logistic regression; for "Applicability" and "Defensibility", significance codes are based on random effects ordered probit regressions.

Admission rates by research type

Research type	Admission rate
Clinical	0.30
Behavioral	0.34*
Laboratory	0.37***
Field	0.31
Experiments	0.36***
Modelling	0.34

Significance codes:
*p<.05, **p<.01, ***p<.0005
For "Admission rate," significance codes are based on random effects logistic regression; for "Applicability" and "Defensibility", significance codes are based on random effects ordered probit regressions.

Sub-group differences

- applicability of the items?
- willingness to admit?
- publication pressures?
- perceived defensibility of the items?
- research integrity?

Follow-up survey sent to subset of original respondents:

- Ss presented with same 10 QRPs from initial study; rate:
  1. Applicability to their research methodology
     (never applicable / sometimes / often / always)
  2. General defensibility
     (indefensible / possibly defensible / defensible)
- Response rate = 35% (504 out of 1,440)

Findings

1. Subgroup differences in applicability and defensibility ratings coincided with prevalence estimate findings
   - But did not account for all variance in prevalence estimates

2. Across subgroups, the practices were deemed to be ***indefensible***

PsychDisclosure.org

(LeBel, Borsboom, Giner-Sorella, Hasselman, Peters, Ratliff, Tucker Smith, forthcoming, *Perspectives on Psychological Science*)

- Contacted 50% of authors of recent top Psych journals; asked them to disclose criteria recommended by Simmons, Nelson & Simonsohn (2012)

Disclosure categories:

1. **Exclusions:** Disclosed total number of observations excluded and criterion for doing so.
2. **Conditions:** Disclosed all tested experimental conditions, including failed manipulations.
3. **Measures:** Disclosed all administered measures and items.
4. **Sample size:** Disclosed (a) *basis* for chosen sample sizes and (b) *basis* for stopping data collection.

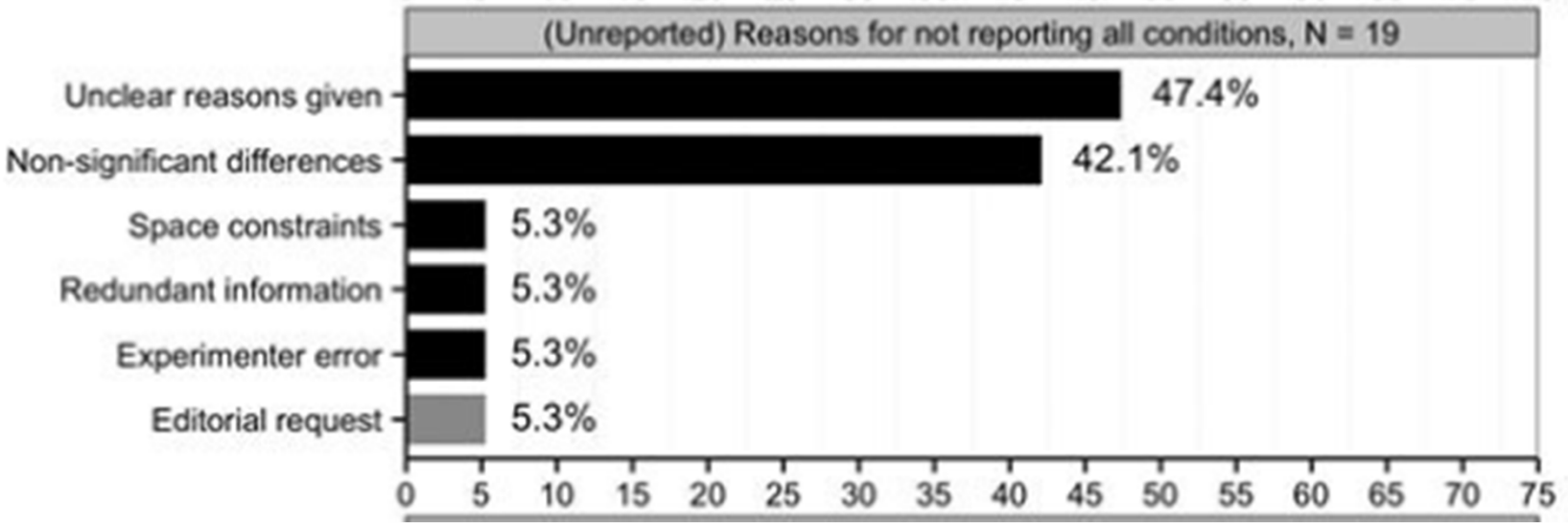
- Compliance rate: 50%

PsychDisclosure

(LeBel et al., forthcoming, Perspectives on Psychological Science)

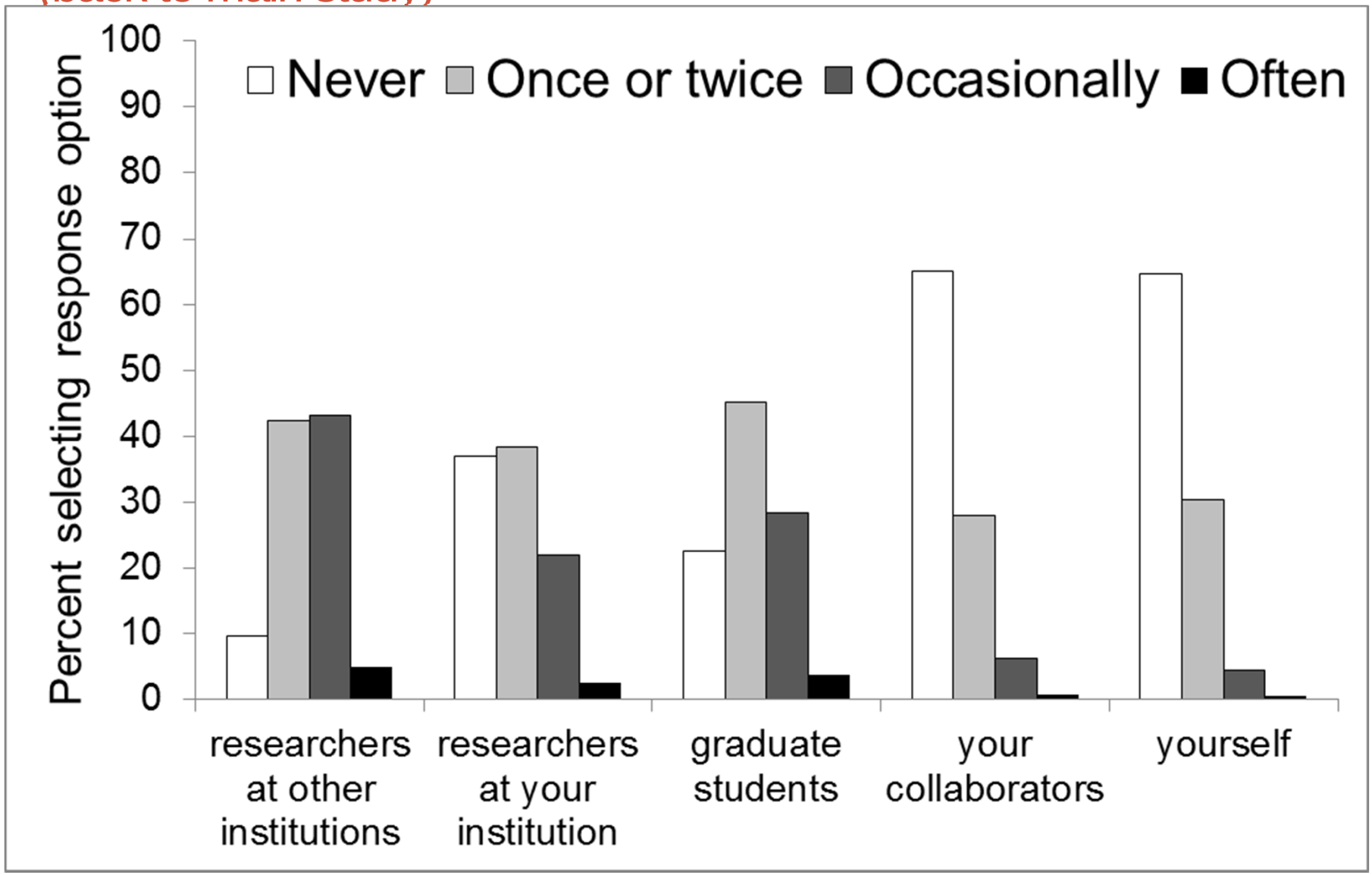
89% indicated that they had reported all conditions

Reasons for not having reported all conditions:



Doubts about research integrity

(back to main study)



Concluding comments

Summary:

1. 3 measures provide converging evidence of prevalence of QRPs
2. Incentive-compatible elicitation generates slightly higher estimates

We assume that researchers are sincerely motivated to conduct sound research, but…

- inherent ambiguity + incentives + motivated reasoning (Kunda, 1990) combine to raise prevalence

Thank you

Contact: ljohn@hbs.edu