Promoting Transparency in Social Science Research

Edward Miguel
University of California, Berkeley

Summer Institute
June 2014

BITSS
Berkeley Initiative for Transparency in the Social Sciences

Motivation

- **Data mining and selective presentation** of results are long standing concerns in Economics:
- – Leamer (1983, "Let's take the con out of econometrics")
- – Card and Krueger (1995) find publication bias in the labor economics minimum wage literature
- – Ashenfelter et al. (1999), DeLong and Lang (1992), etc.

- There has been growing interest in research transparency in economics and across social science disciplines, driven by a widespread perception that **many influential findings are fragile** (at best)

Motivation

- These concerns are not limited to Economics:
- **Medical trials**, Ioannidis (2005, "Why most published research findings are false")
- **Psychology**, Simmons et al. (2011, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant")
- **Political science**, Humphreys et al. (2012, "Fishing")

- The stakes are high since policy decisions based on social science research affect millions of people.

This talk: promoting research transparency

- This talk is based on two recently published papers

1. Miguel et al. (2014, *Science*); co-authors are economists, psychologists, political scientists, and bio-statisticians.
2. Casey, Glennerster & Miguel (2012, *QJE*) on local institutional reform in Sierra Leone.

- In this talk briefly discuss:
- **Emerging practices** in research transparency, including parallel efforts across social sciences;
- **Open questions** about how widely these practices could be adopted, and how to change research norms;
- **Set the stage** for the talks to come during the week.

Social science research practices

- In the last two decades, field experiments, lab experiments and other studies featuring original data collection and rigorous research designs (e.g., IV, RD, etc.) have become widespread across the social sciences

Social science research practices

- In the last two decades, field experiments, lab experiments and other studies featuring original data collection and rigorous research designs (e.g., IV, RD, etc.) have become widespread across the social sciences

- While this is a big step forward, the use of **rigorous experimental research designs alone may not be enough** to ensure credible bodies of scientific evidence.

Social science research practices

- Why? The norms, incentives and institutions governing social science research have not changed, e.g.:

- Statistically significant, novel and theoretically “tidy” results are published more easily than null, replication, and perplexing results, even conditional on the quality of the research design → an incomplete body of evidence.

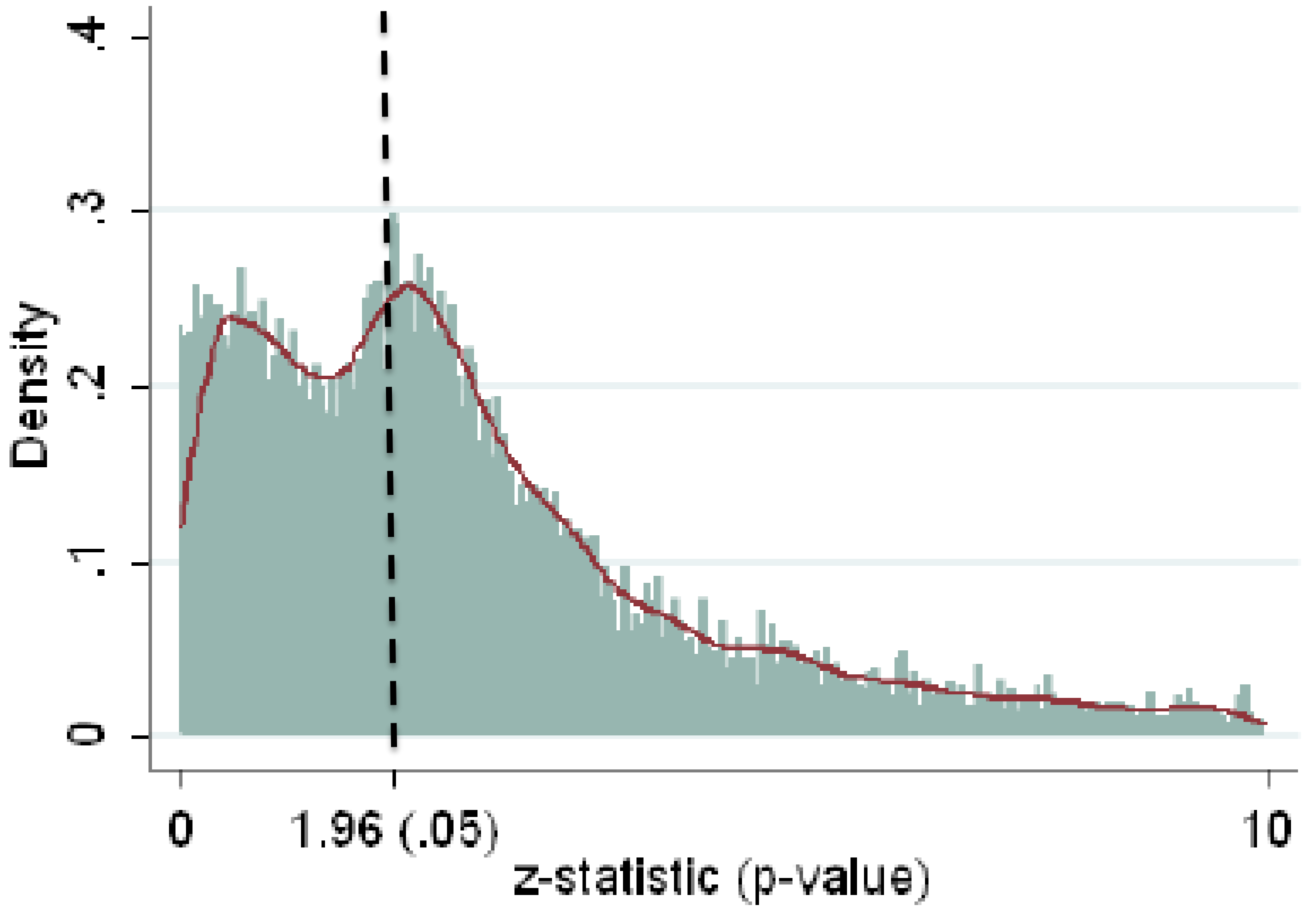
- Ample evidence of **publication bias** in all fields (i.e., large number of studies with p-values just below 0.05).

Publication bias and data mining

(b) Unrounded distribution of z-statistics.

AER, JPE, QJE (2005-2011)

Brodeur et al 2012



Publication bias and data mining

Gerber and Malhotra 2008

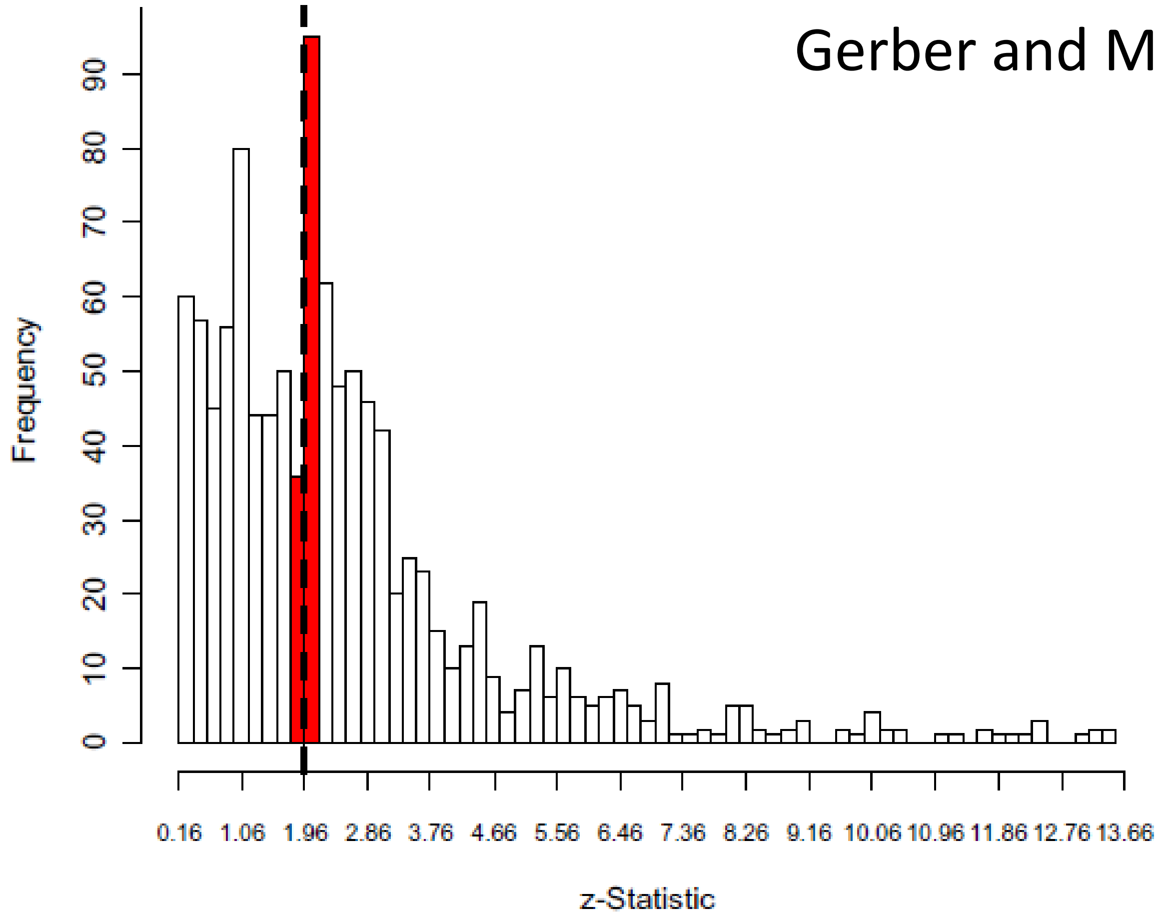
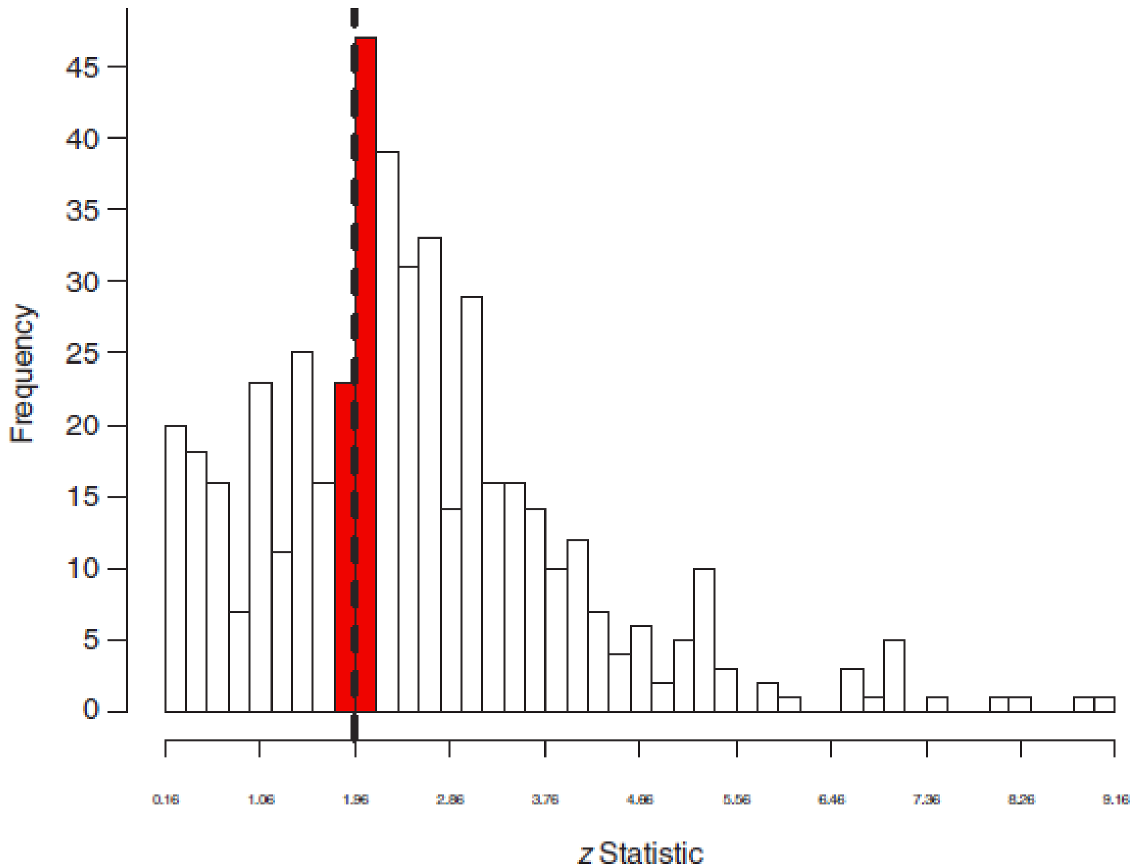


Figure 1(a). Histogram of z-statistics, *APSR & AJPS* (Two-Tailed). Width of bars (0.20) approximately represents 10% caliper. Dotted line represents critical z-statistic (1.96) associated with $p = 0.05$ significance level for one-tailed tests.

Publication bias and data mining

Gerber and Malhotra 2008

Figure 1
Histogram of *z* Statistics From the *American Sociological Review*, the *American Journal of Sociology*, and *The Sociological Quarterly* (Two-Tailed)



Emerging practices in research transparency

- Bottom-up innovation around **three sets of practices**:

Emerging practices in research transparency

- Bottom-up innovation around **three sets of practices**:

(1) *Disclosure*: require researchers to report all measures, manipulations, and data exclusions.

→ Political science, psychology journals have new standards

Fig. 1. Three mechanisms for increasing transparency in scientific reporting

Demonstrated with a research question: “Do shorter summer breaks improve educational outcomes?” *n.s.* denotes $P > 0.05$.

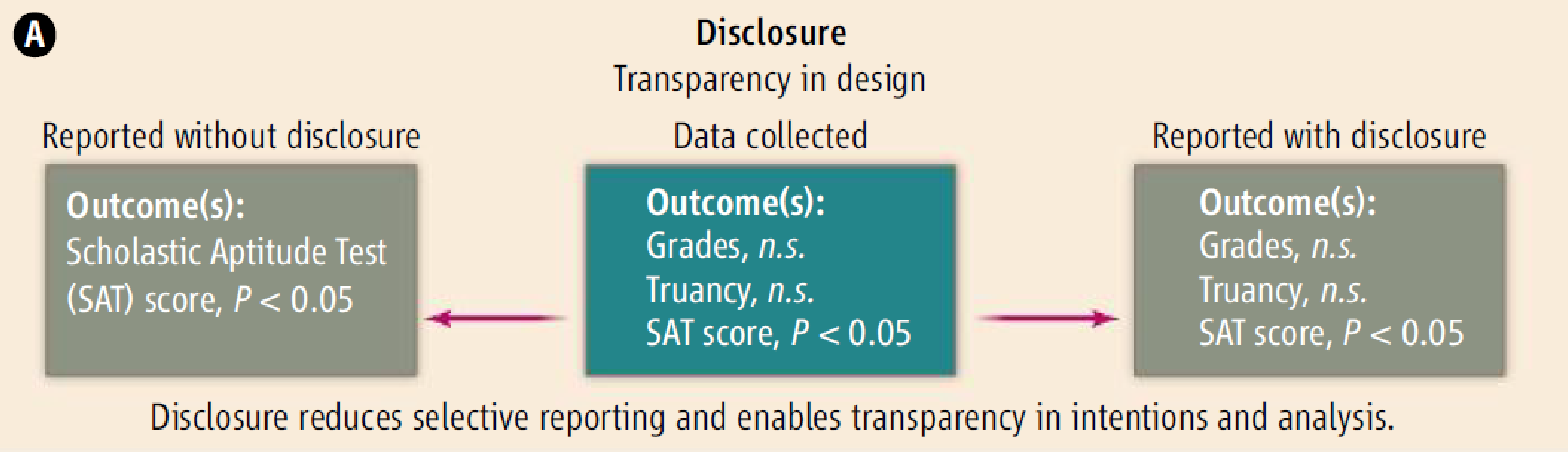


Fig. 1. Three mechanisms for increasing transparency in scientific reporting

- Bottom-up innovation around **three sets of practices**:

(1) *Disclosure*: require researchers to report all measures, manipulations, and data exclusions.

→ Political science, psychology journals have new standards

(2) *Open data and materials*: sharing of data, code, surveys

→ Open Science Framework (OSF) platform, and others

Fig. 1. Three mechanisms for increasing transparency in scientific reporting

Demonstrated with a research question: “Do shorter summer breaks improve educational outcomes?” *n.s.* denotes $P > 0.05$.

C

Open data and materials

Transparency in analysis

Summer break	Grades	Truancy	SAT score
Short	2.95	2%	1020
Short	3.30	0%	1360
Long	2.32	4%	9.80 ?
Long	3.87	0%	1450

Open data reduce errors and fraud
and facilitate replication and extension.

Fig. 1. Three mechanisms for increasing transparency in scientific reporting

- Bottom-up innovation around **three sets of practices**:

(1) *Disclosure*: require researchers to report all measures, manipulations, and data exclusions.

→ Political science, psychology journals have new standards

(2) *Open data and materials*: sharing of data, code, surveys

→ Open Science Framework (OSF) platform, and others

(3) *Registration and pre-analysis plans (PAP)*: prospectively register hypotheses in a public database.

→ New (2013) AEA registry (socialscienceregistry.org)

Fig. 1. Three mechanisms for increasing transparency in scientific reporting

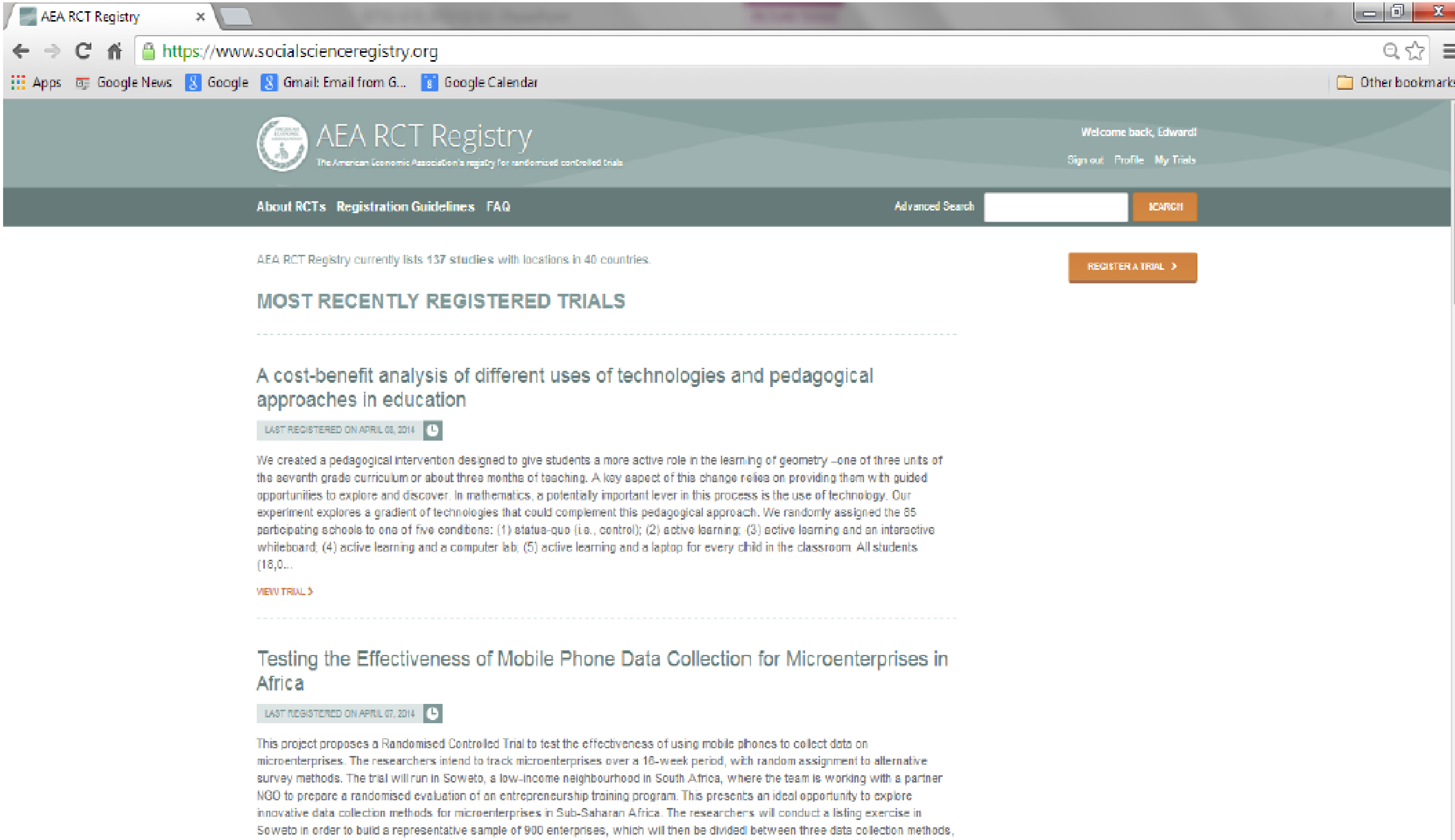
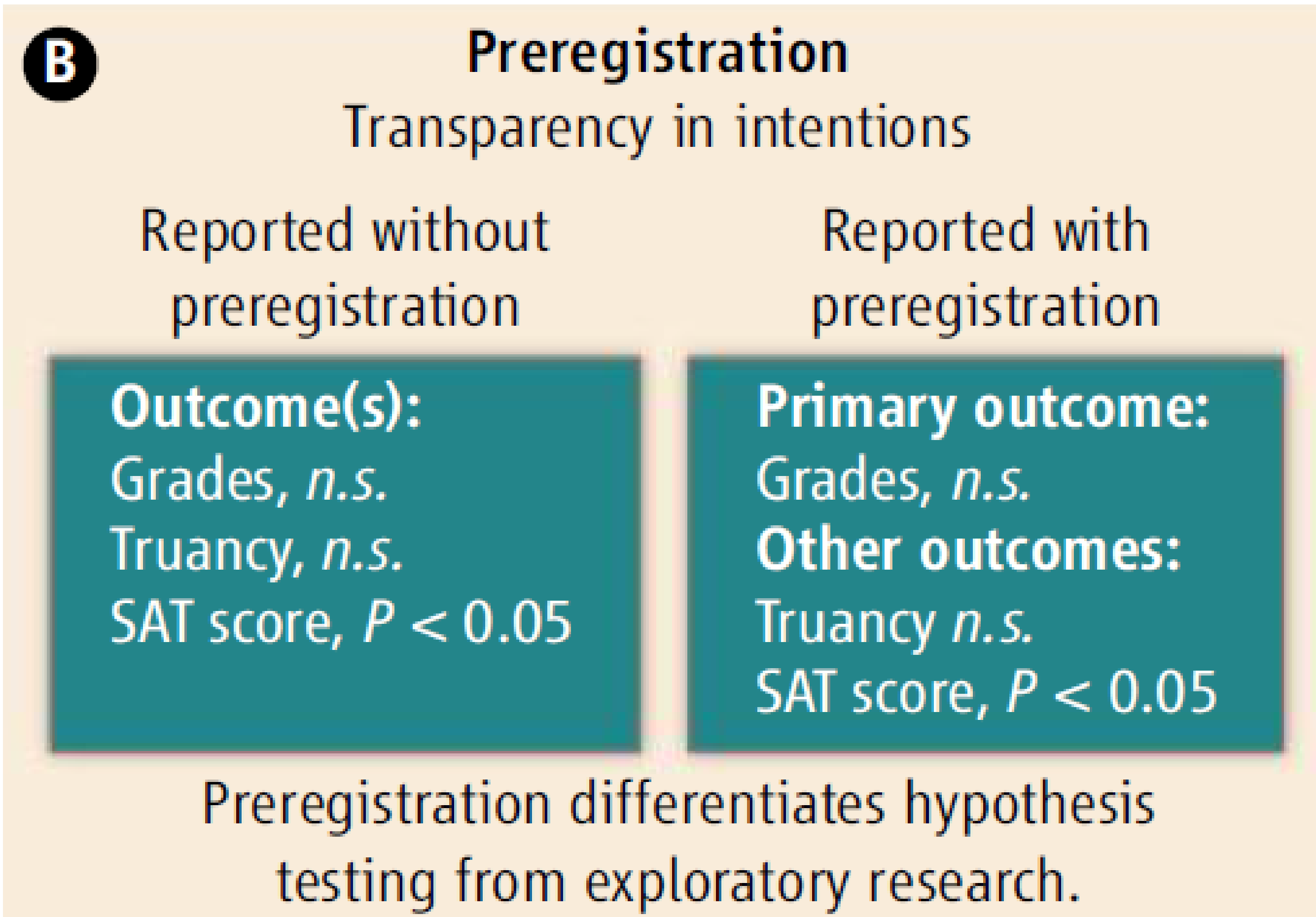


Fig. 1. Three mechanisms for increasing transparency in scientific reporting

Demonstrated with a research question: “Do shorter summer breaks improve educational outcomes?” *n.s.* denotes $P > 0.05$.



Registration of pre-analysis plans

- Why might **registration** of pre-analysis plans be useful?

Registration of pre-analysis plans

- Why might **registration** of pre-analysis plans be useful?

1. Rounds out the body of evidence by creating a “paper trail” of unpublished studies in an area, potentially helping to reduce **publication bias**

Registration of pre-analysis plans

- Why might **registration** of pre-analysis plans be useful?

1. Rounds out the body of evidence by creating a “paper trail” of unpublished studies in an area, potentially helping to reduce **publication bias**
2. Reduces the risk of **data mining** and other tendentious presentation of empirical results (“cherry-picking”)
3. Generates **correctly sized statistical tests**, bolstering the credibility of statistical significance levels

Registration of pre-analysis plans

- Why might **registration** of pre-analysis plans be useful?

1. Rounds out the body of evidence by creating a "paper trail" of unpublished studies in an area, potentially helping to reduce **publication bias**
2. Reduces the risk of **data mining** and other tendentious presentation of empirical results ("cherry-picking")
3. Generates **correctly sized statistical tests**, bolstering the credibility of statistical significance levels
4. As a side benefit, forces researchers to more carefully **think through their hypotheses** beforehand, improving the quality of research design, data collection

Registration of pre-analysis plans

- A leading concern: will pre-registration of plans **stifle creativity** and limit discoveries made through exploratory research?
- Many, if not most, important scientific findings undoubtedly originated as unexpected discoveries…

Registration of pre-analysis plans

- A leading concern: will pre-registration of plans **stifle creativity** and limit discoveries made through exploratory research?
- Many, if not most, important scientific findings undoubtedly originated as unexpected discoveries…

- But findings from such work are inherently more tentative because of the greater flexibility of tests and the greater opportunity for the outcome to obtain by chance.

→ Pre-specification is not intended to disparage exploratory analysis, but rather to free it from **the tradition of being portrayed as formal hypothesis testing**.

An example from development economics

- Casey, Glennerster and Miguel (2012, *QJE*)
- A **randomized experiment in Sierra Leone** to study the impact of a local institutional reform ("community driven development", called the "GoBifo" program) on local politics, collective action, inclusion of marginalized groups (i.e., women, youth), and social capital across 236 villages.

- PAP registered in the MIT Jameel Poverty Action Lab registry in 2009

An example from development economics

- A **large and diverse set of outcomes** derived from surveys and field observations, N=155 in total

- Even in the context of a field experiment, this immediately raises questions about data mining – not only in terms of econometric specification (Leamer's concern) but by focusing on a subset of outcomes

An example from development economics

- A **large and diverse set of outcomes** derived from surveys and field observations, N=155 in total

- Even in the context of a field experiment, this immediately raises questions about data mining – not only in terms of econometric specification (Leamer's concern) but by focusing on a subset of outcomes

- E.g., if a researcher has access to N=30 independent outcomes, none of which is related to the "treatment", the likelihood that at least one is "significant" at 95% confidence is $1 – (0.95)^{30} = 0.8$.

An example from development economics

- Main result: what was the impact of GoBifo on local politics, collective action, inclusion of marginalized groups (i.e., women, youth), and social capital?

- We examined nine distinct "groups" of outcomes using a mean effects approach, with a family-wise error rate (FWER) multiple testing adjustment (Westfall-Young 1993), and also combined them in a single index.

An example from development economics

- Main result: what was the impact of GoBifo on local politics, collective action, inclusion of marginalized groups (i.e., women, youth), and social capital?

- We examined nine distinct "groups" of outcomes using a mean effects approach, with a family-wise error rate (FWER) multiple testing adjustment (Westfall-Young 1993), and also combined them in a single index.

→ None of the nine hypotheses is rejected at traditional confidence levels, and the overall mean effect is **a precisely estimated zero**: 0.028 (s.e. 0.020) in sd units.

An example from development economics

- Why was the PAP useful here?

An example from development economics

- Why was the PAP useful here?
- It allowed us to **limit pressure from government officials** and donors who wanted us to show "success" by focusing on particular outcomes.
- In other cases, pressure from journal editors, referees or colleagues to confirm existing findings – or reaffirm central "tenets" of the discipline (e.g., Card and Krueger's minimum wage research).

An example from development economics

- Why was the PAP useful here?
- It allowed us to **limit pressure from government officials** and donors who wanted us to show "success" by focusing on particular outcomes.
- In other cases, pressure from journal editors, referees or colleagues to confirm existing findings – or reaffirm central "tenets" of the discipline (e.g., Card and Krueger's minimum wage research).

- **What might have been**: given the large number of outcomes, we show that cherry-picking could easily have led us to two completely divergent – and equally erroneous – interpretations of the evidence.

An example from development economics

TABLE VI

ERRONEOUS INTERPRETATIONS UNDER "CHERRY PICKING"

Negative impacts

Outcome variable	(1) Mean for controls	(2) Treatment effect	(3) Robust std. err.
Panel A: GoBifo "weakened" institutions			
Attended meeting to decide what to do with the tarp	0.81	-0.04^{+}	(0.02)
Everybody had equal say in deciding how to use the tarp	0.51	-0.11^{+}	(0.06)
Community used the tarp (verified by physical assessment)	0.90	-0.08^{+}	(0.04)
Community can show research team the tarp	0.84	-0.12^{*}	(0.05)
Respondent would like to be a member of the VDC	0.36	-0.04^{*}	(0.02)
Respondent voted in the local government election (2008)	0.85	-0.04^{*}	(0.02)

An example from development economics

TABLE VI

ERRONEOUS INTERPRETATIONS UNDER "CHERRY PICKING"

Outcome variable	(1) Mean for controls	(2) Treatment effect	(3) Robust std. err.
Panel B: GoBifo "strengthened" institutions			
Community teachers have been trained	0.47	0.12[+]	(0.07)
Respondent is a member of a women's group	0.24	0.06**	(0.02)
Someone took minutes at the most recent community meeting	0.30	0.14*	(0.06)
Building materials stored in a public place when not in use	0.13	0.25*	(0.10)
Chiefdom official did not have the most influence over tarp use	0.54	0.06*	(0.03)
Respondent agrees with "Responsible young people can be good leaders" and not "Only older people are mature enough to be leaders"	0.76	0.04*	(0.02)
Correctly able to name the year of the next general elections	0.19	0.04*	(0.02)

Positive impacts

Using pre-analysis plans

- How should PAPs be used?

Using pre-analysis plans

- How should PAPs be used?

1. In Casey et al (2012), we emphasize that it is important to allow for **flexibility with complete transparency**

- I.e., we registered 11 main hypotheses but forgot a very basic hypothesis (to test whether the program was implemented properly). We thus present 12 main hypotheses in our paper but flag for the reader the one that was added ex post.
- It is important that the adoption of PAPs does not stifle corrections or exploratory data analysis. The reader just needs to know what is pre-specified and what is not.

Using pre-analysis plans

- How should PAPs be used?

2. Finkelstein et al (2012, *QJE*) use a pre-registered plan to study impacts of Medicaid expansion on health expenditures and outcomes in Oregon.

- U.S. health care reform is another highly politicized setting where maintaining the highest standards of rigor and transparency is important for credibility.
- An innovation: **pre-specified that they would first use data from the control group only** to determine appropriate outcome variables (i.e., dropping those with no variation) and regression specifications.

Open questions

- A key question going forward: **how widely** should pre-registration of research plans be applied?

Open questions

- A key question going forward: how widely should pre-registration of research plans be applied?

1. **Laboratory experiments**: PAPs could be particularly fruitful, given the relatively low cost of running additional experiments and never publishing the data

- Registration gaining traction in psychology:
  - "Registered reports", studies accepted for publication based on their research design rather than results, are being introduced in several leading journals.
  - Crowd-sourced replication projects of major findings.

Open questions

2. How should registration and analysis plans be applied to **observational (non-experimental) data**?

Open questions

2. How should registration and analysis plans be applied to **observational (non-experimental) data**?

- One promising area is the registration of prospective non-experimental research, including **studies of anticipated policy changes**.
  - The first pre-analysis plan in Economics (to our knowledge) was Neumark's (2001) plan to study a future minimum wage increase on employment.

Open questions

2. How should registration and analysis plans be applied to **observational (non-experimental) data**?

- One promising area is the registration of prospective non-experimental research, including **studies of anticipated policy changes**.
  - The first pre-analysis plan in Economics (to our knowledge) was Neumark's (2001) plan to study a future minimum wage increase on employment.

- Pre-registration can be used when **new "rounds" of data** are being released (e.g., a new PSID wave, Census round), or where access to existing data is restricted and thus where data mining is impossible ex ante.

Open questions

3. Applications **beyond applied micro** studies:

- To reduce concerns about "specification search", researchers could also pre-register:
  - the parameters to be used in **macro calibrations,**
  - the models used in structural estimation (i.e., in IO),
  - prior distributions in Bayesian analysis.

Open questions

3. Applications **beyond applied micro** studies:

- To reduce concerns about "specification search", researchers could also pre-register:
  – the parameters to be used in **macro calibrations,**
  – the models used in structural estimation (i.e., in IO),
  – prior distributions in Bayesian analysis.

- Plans are archived but only published with a time lag (on the AEA registry or OSF site) to make sure researchers with creative ideas are not "scooped" by others.

Changing Research Norms

- How can we hasten the adoption of new norms?

Changing Research Norms

- How can we hasten the adoption of new norms?

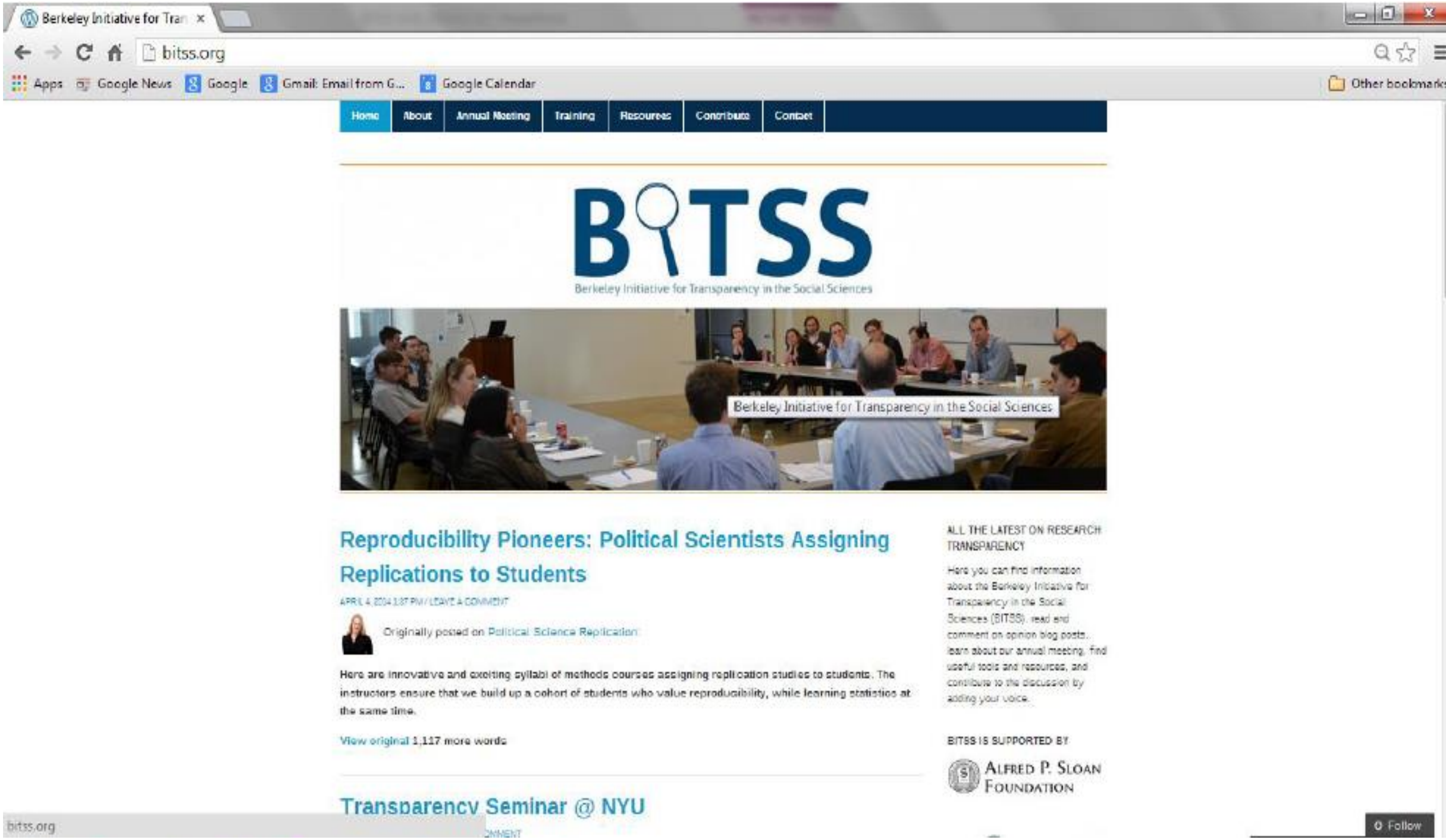
- The adoption of pre-registration norms in medical trial research was quite **“top down”**, with requirements from the government (FDA), journals, and funders starting around 2000.

Changing Research Norms

- A bottom-up approach: the **Berkeley Initiative for Transparency in the Social Sciences (BITSS)** is a network launched last year

- Multiple activities aimed at promoting dialogue, informing, and training – including this week's course

Changing Research Norms



Changing Research Norms

- **Social norms** are starting to shift on their own: 240 (!) registered plans on the AEA registry and the EGAP political science registry in the last year.

- The establishment of the AEA registry – which is free and open to non-AEA members – was an important milestone, a signal that PAP's are now "mainstream".

Changing Research Norms

- **Social norms** are starting to shift on their own: 240 (!) registered plans on the AEA registry and the EGAP political science registry in the last year.

- The establishment of the AEA registry – which is free and open to non-AEA members – was an important milestone, a signal that PAP's are now "mainstream".

- New transparency practices promise to hasten research progress, improve the quality of evidence we as a research community can provide policymakers, and **realign scholarly practice with scholarly values**.

EXTRA SLIDES

Changing Research Norms

- Planned future directions:

1. Meeting with leading journal editors, including editors-in chief of *Science* and *Nature*, and senior leadership at NIH and NSF (November 2014)
2. Possible inter-disciplinary graduate course on transparent social science research methods at U.C. Berkeley (Spring 2015).
3. Possible inclusion of relevant content in core econometrics courses (e.g., multiple testing adjustment, meta-analysis) and applied micro courses (e.g., replication exercises, pre-analysis plans, data sharing and management)

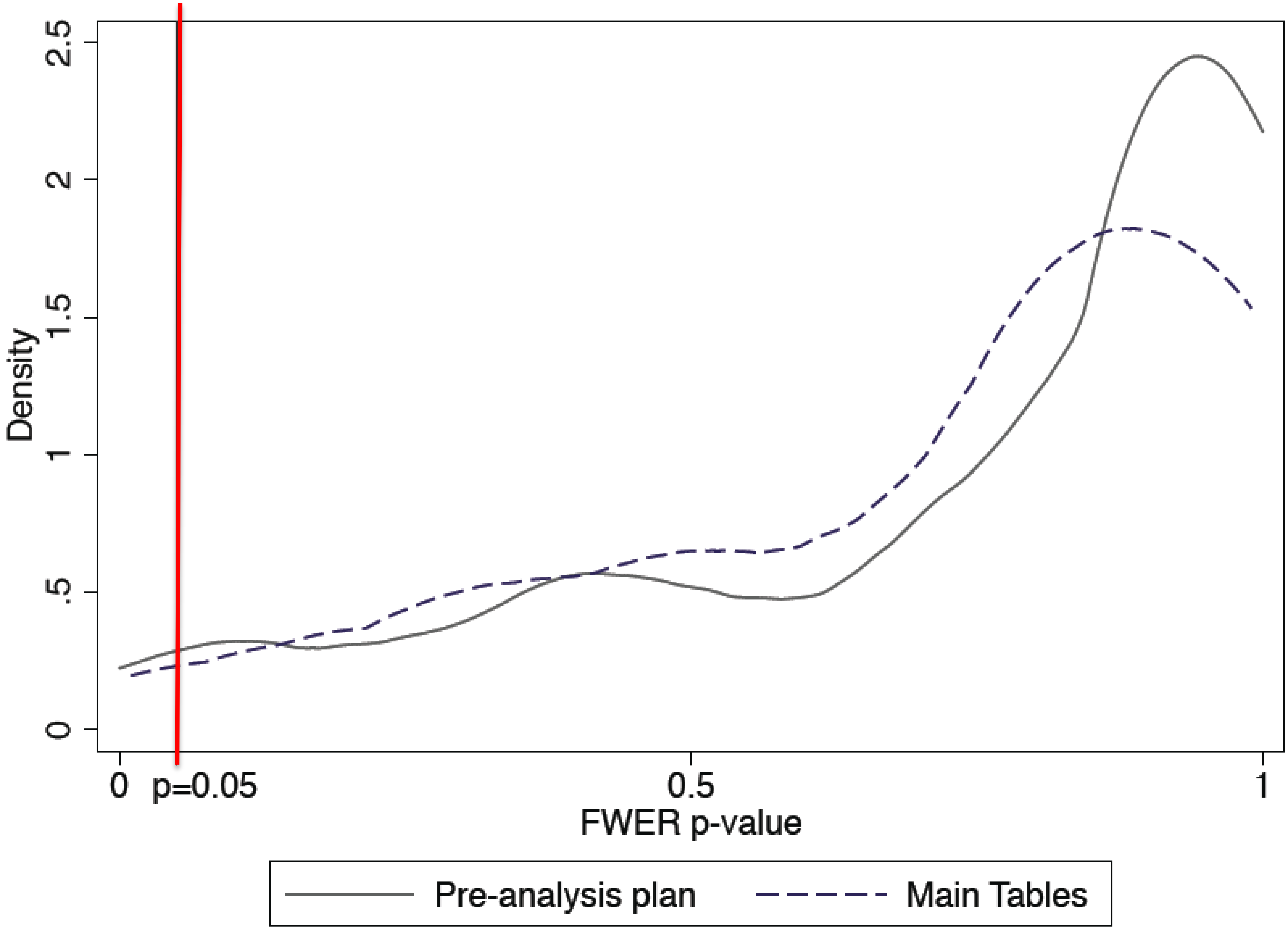
Using pre-analysis plans

- How should PAPs be used?

3. Berge et al (2014) use pre-analysis plans in a lab experiment studying of the role of election timing and ethnicity on cooperation in Kenya and Tanzania.

- Registered separate PAP’s before each of three lab rounds
- An innovation: compare the distribution of p-values presented in the final paper versus those pre-specified in the plan, **to assess if there is selective presentation of statistically significant results**.

Using pre-analysis plans



Changing Research Norms

- A bottom-up approach: the **Berkeley Initiative for Transparency in the Social Sciences (BITSS)** is a network launched last year
  - An active blog and forum (bitss.org)
  - Conferences and sessions at professional association meetings to facilitate discussions, build consensus – next meeting in Berkeley, December 2014
  - Training course on research transparency methods (for Ph.D. students, post-docs, others) in Berkeley June 2-6 2014

Title [Content slide]

- Content.
- Content.

Title [Table / Figure slide]