Credibility and reproducibility in economics & management

Introduction

- PhD candidate at Organization Studies
- Main interest: Behavioral strategy / Organizational Sociology
- Hobby: science on science / open science

Reproducibility: a hype to pass?

God blessed them and said to them, “Be fruitful and **<u>reproduce</u>**” (Genesis 1:28, 9:1)

Today's agenda

1. Times of crisis: replication, credibility, and reproducibility

2. Assessing credibility and reproducibility

3. Increase your own credibility and reproducibility

Definitions

"…I define "replication" as independent people going out and collecting new data and "reproducibility" as independent people analyzing the same data." (Peng, 2011)

Replication: Collect new data from similar population using similar methods (stimuli) and analytic techniques

Reproducibility: Reanalyze same data from conducted study

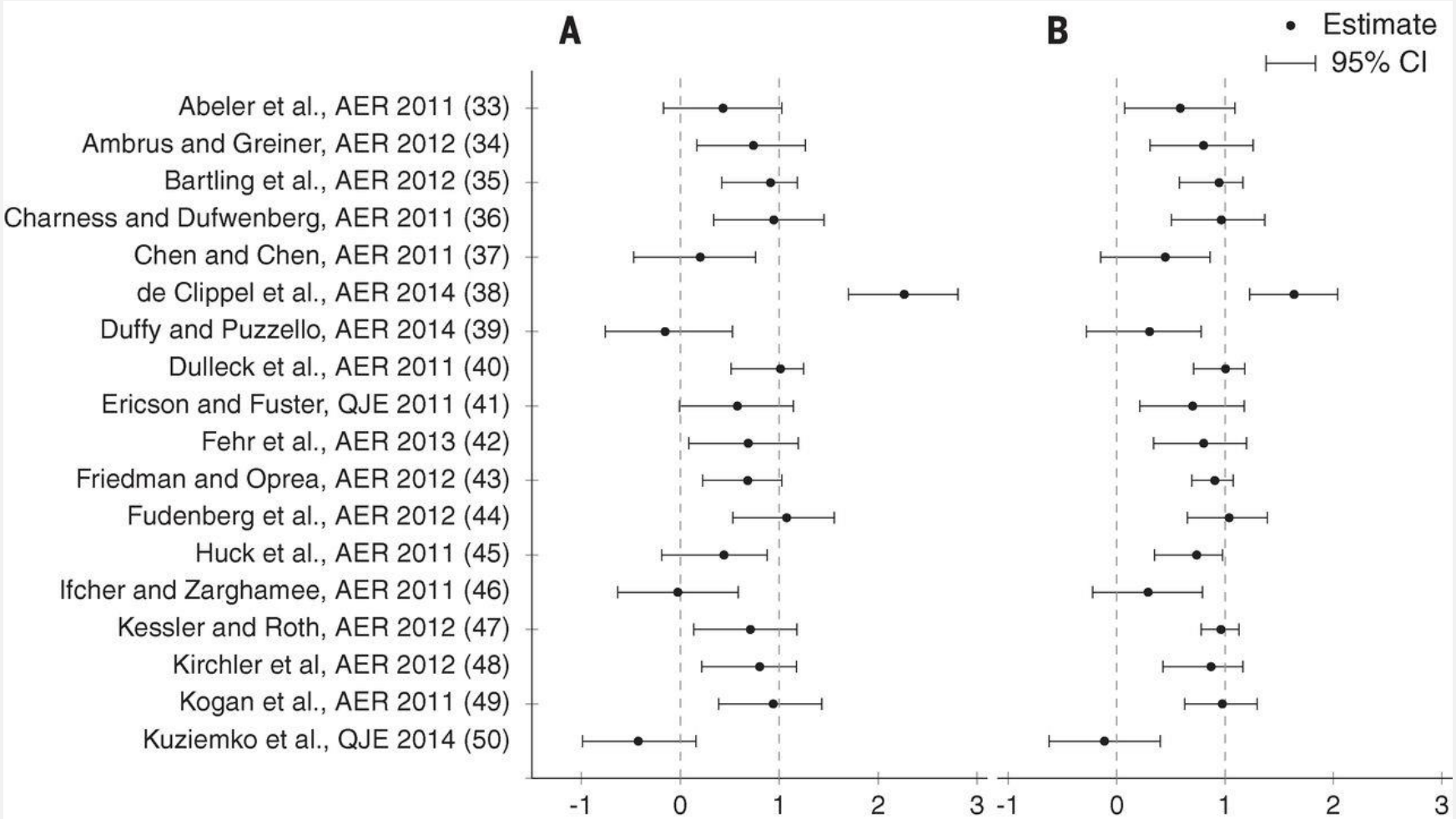
Credibility: The extent to which outcomes of research are trustworthy

Replicability

- Large, collaborative projects show that effects do not replicate
  - 36% (35/97) effects replicated (OSC, 2015)
  - 62% (13/21) effects in *Nature* & *Science* replicated (Camerer et al., 2018a)
  - Experimental economics: 61% (11/18) replicated (Camerer et al., 2018b)

Many findings in the literature do not seem to describe 'true' phenomena

Replication in Economics (Camerer et al., 2016)



Reproduciblity: Statistical inconsistencies

- Study on the reproducibility of findings in *Strategic Management Journal* by Bergh et al. (2017)

- Reproduction based on descriptive statistics (M, SD, N, *r*) using corr2data in STATA

Reproduciblity: Statistical inconsistencies

Table 2. Reproducibility of ordinary linear regression hypothesis findings: reported and reproduced statistical significance levels.

Study identifier	Reported *p* value	Reproduced *p* value by Stata	Reproduced *p* value by IBM SPSS
1	<0.01	0.087	0.087
2	<0.05	0.704	0.704
3	<0.01	0.244	0.244
4	<0.01	0.109	0.109
4	<0.001	0.179	0.179
4	<0.05	0.241	0.241
4	<0.05	0.386	0.386
4	<0.001	0.172	0.172
4	<0.01	0.093	0.093
4	<0.001	0.115	0.115
4	<0.05	0.909	0.909
5	<0.05	0.053	0.053
6	<0.05	0.174	0.174
6	<0.05	0.213	0.213

p, observed probability for the null hypothesis that the coefficient is zero in the population. Reported *p* values are those reported in the published studies and reproduced results are those obtained using the reproducibility procedures described in text.

Reproduciblity: Statistical inconsistencies

- Study on the reproducibility of findings in *Strategic Management Journal* by Bergh et al. (2017)

- Reproduction based on descriptive statistics (M, SD, N, *r*) using corr2data in STATA

- "**one-third reported hypotheses as statistically significant which were no longer so** and far more significant results were found to be non-significant in the reproductions than in the opposite direction." (Bergh et al., 2017)

Reproduciblity: Statistical inconsistencies

- Statistical inconsistencies: reported test statistics are *incongruent*
  - *F-value, df, and p-value*
  - *Beta, SE, and p-value*

- In psychology: 1 in 8 papers contains a *gross inconsistency* (Nuijten et al., 2016)
- In innovation research: 1 in 2 papers contains a *gross inconsistency* (Bruns et al., 2019)
- Business/management: similar results (van Zelst & Smeets, wip)

Credibility

- Credibility: The extent to which outcomes of research are trustworthy

- Do you automatically trust research because it is published in a peer-reviewed journal?

<u>Based on the aforementioned evidence:</u>
<u>You should stop doing that!</u>

Credibility

- Credibility: The extent to which outcomes of research are trustworthy
- Do you automatically trust research because it is published in a peer-reviewed journal?

<u>Based on the aforementioned evidence:</u>
<u>You should stop doing that!</u>

- You can quantify the credibility of published results
  - Example: Excess significance

Excess significance

- Credibility: The extent to which outcomes of research are trustworthy

- Power: the probability of finding a significant result given that the alternative hypothesis is true

- Independent-samples t-test, d = 0.3, alpha = 5%, power = 80%
- Required sample = 352

- Independent-samples t-test, d = 0.5, alpha = 5%, sample size = 60
- Power = 48%

Excess significance

- Credibility: The extent to which outcomes of research are trustworthy

  - % of significant results: **80%**
    - When testing true hypothesis 100% of the time
    - With alpha = 5%
    - And power = 80%

  - % of significant results: **42,5%**
    - When testing true hypothesis 50% of the time
    - With alpha = 5%
    - And power = 80%

Excess significance

- Credibility: The extent to which outcomes of research are trustworthy
  - % of significant results**80%**
    - When testing true hypothesis 100% of the time
    - With alpha = 5%
    - And power = 80%
  - % of significant results**42,5%**
    - When testing true hypothesis 50% of the time
    - With alpha = 5%
    - And power = 80%
- Typical power in psychology: ~50% (Fraley & Vazire, 2014)
- Percentage of tests that are significant: **~90%** (Sterling, 1995)

Excess significance

- Credibility: The extent to which outcomes of research are trustworthy
- Median statistical power in empirical economics: **~18%** (Ioannidis, Stanley, & Doucouliagos, 2017)

Table 1

Proportions of Empirical Economic Results with Adequate Statistical Power (n = 159 Research Areas)

	WLS-FE	Top 10%	Top 1	PET-PEESE
	(1)	(2)	(3)	(4)
Median proportion (%)	10.5	6.5	1.9	5.8
Mean proportion (%)	21.9	20.1	22.1	20.1

Excess significance

- Example: only significant results
  - 7 statistical tests, all significant
  - 100 participants, d = 0.5

1) Calculate power: 70%

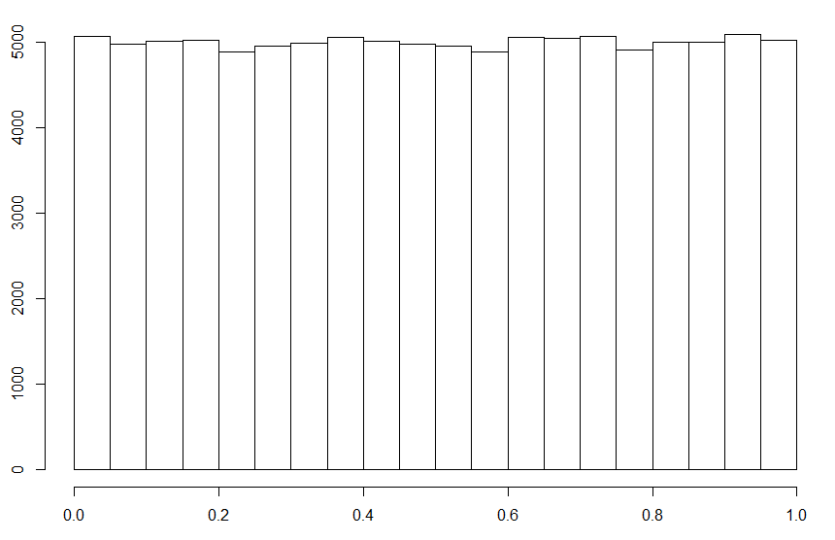
2) Calculate likelihood of finding 7 significant findings:

0.7 x 0.7 x 0.7 x 0.7 x 0.7 x 0.7 x 0.7 = **0.082**

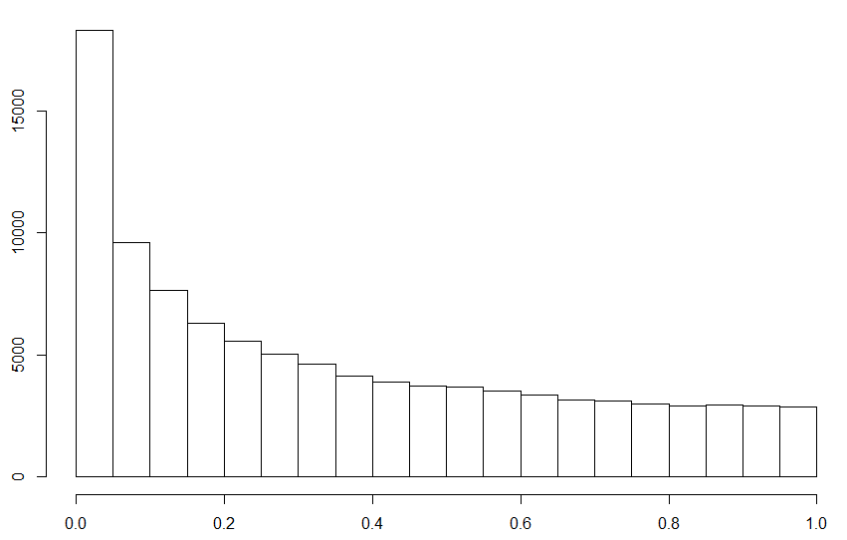
8.2% chance of finding only significant effects across 7 studies with 70% power

Distribution of p-values

d = 0



d = 0.15



Distribution of p-values

Higher fWHR for:

- DOWJones CEOS vs. controls $p = .014$
- DAX CEOs vs. controls $p = .094$
- NGO CEOS: vs. controls $p = .006$
- Popes vs. controls $p = .023$

Correlations:

- DOWJones CEO fWHR employee satisfaction $p = .015$
- DOWJones CEO fWHR CEO approval $p = .042$
- DOWJones CEO fWHR charitable donations $p = .033$
- DOWJones CEO fWHR sustainability index $p = .064$

Distribution of p-values

- 8 findings with p values between .006 and .094
- ~d = 0.8
- ~n = 30

- Likelihood of **1** p value in observed range: 28.57%
- Likelihood of **8** p values in observed range: 0.054%

0.054% chance of observing this set of p values given the effect is true

P-hacking (psychology)

Chronological rejuvenation study

- 20 university students
- Listen to 'When I'm Sixty-Four' or 'Kalimba'
- Indicated their birthday and their father's age (which was included to control for baseline age differences)
- 'people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted M = 20.1 years) rather than to "Kalimba" (adjusted M = 21.5 years), F(1, 17) = 4.92, **p = .040**.'

Recruited 34 participants, but some were excluded

A third condition was dropped

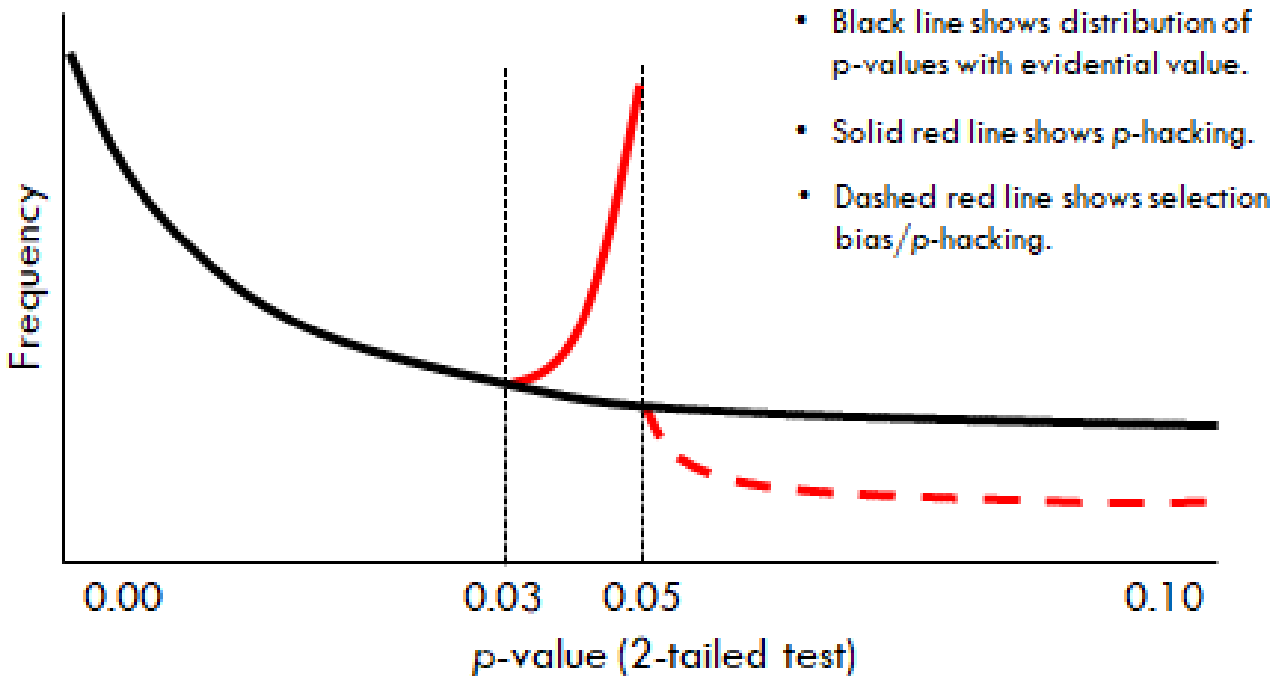
11 more variables (such as mother's age) were also measured

Not significant without the covariate

Simmons, Nelson, & Simonsohn (2011)

P-hacking (management)

Figure 1. Theoretical distribution of *p*-values

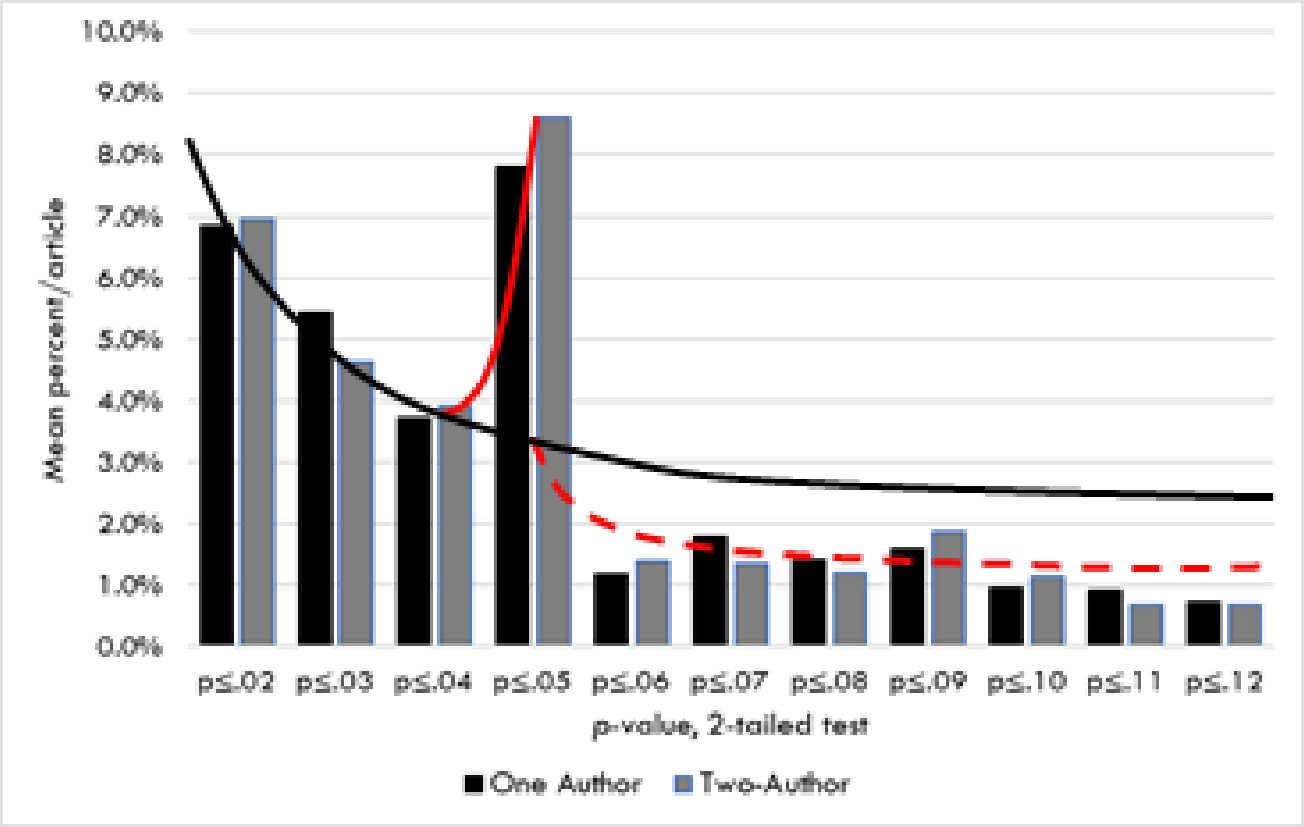


- Black line shows distribution of p-values with evidential value.
- Solid red line shows *p*-hacking.
- Dashed red line shows selection bias/*p*-hacking.

Adapted from Head et al. (2015)

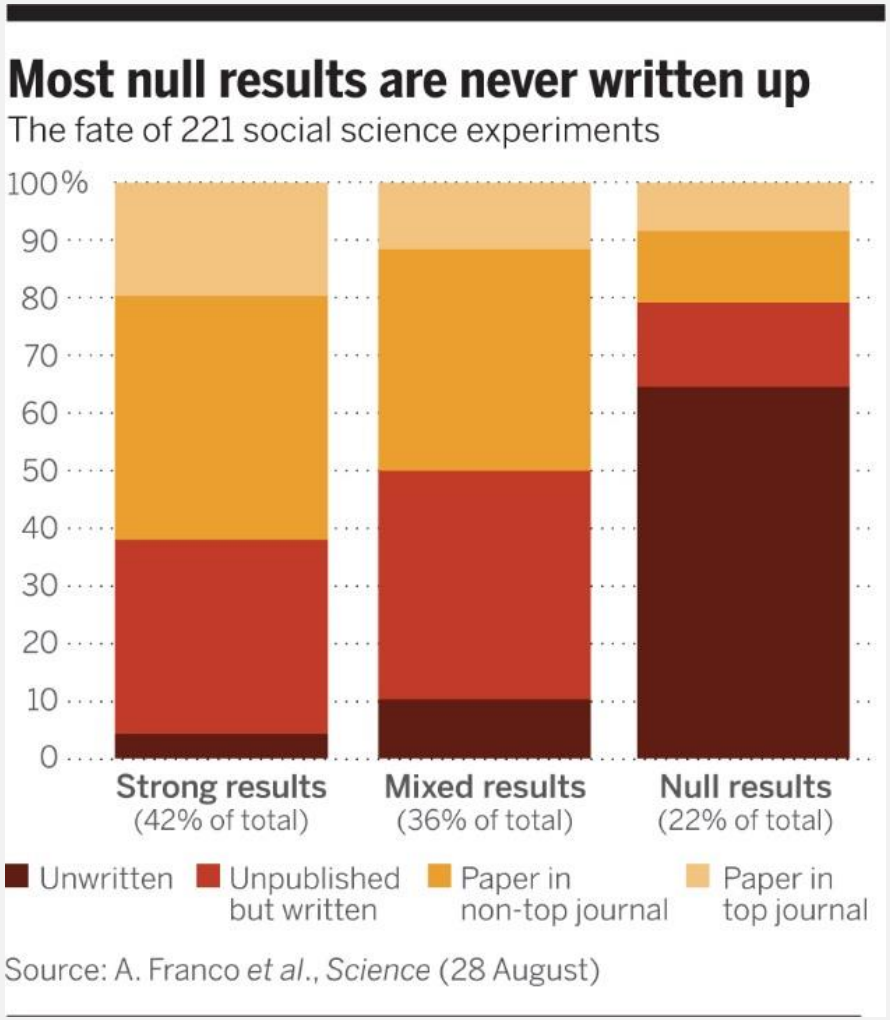
P-hacking (management)

Figure 2. Empirical distribution of *p*-values

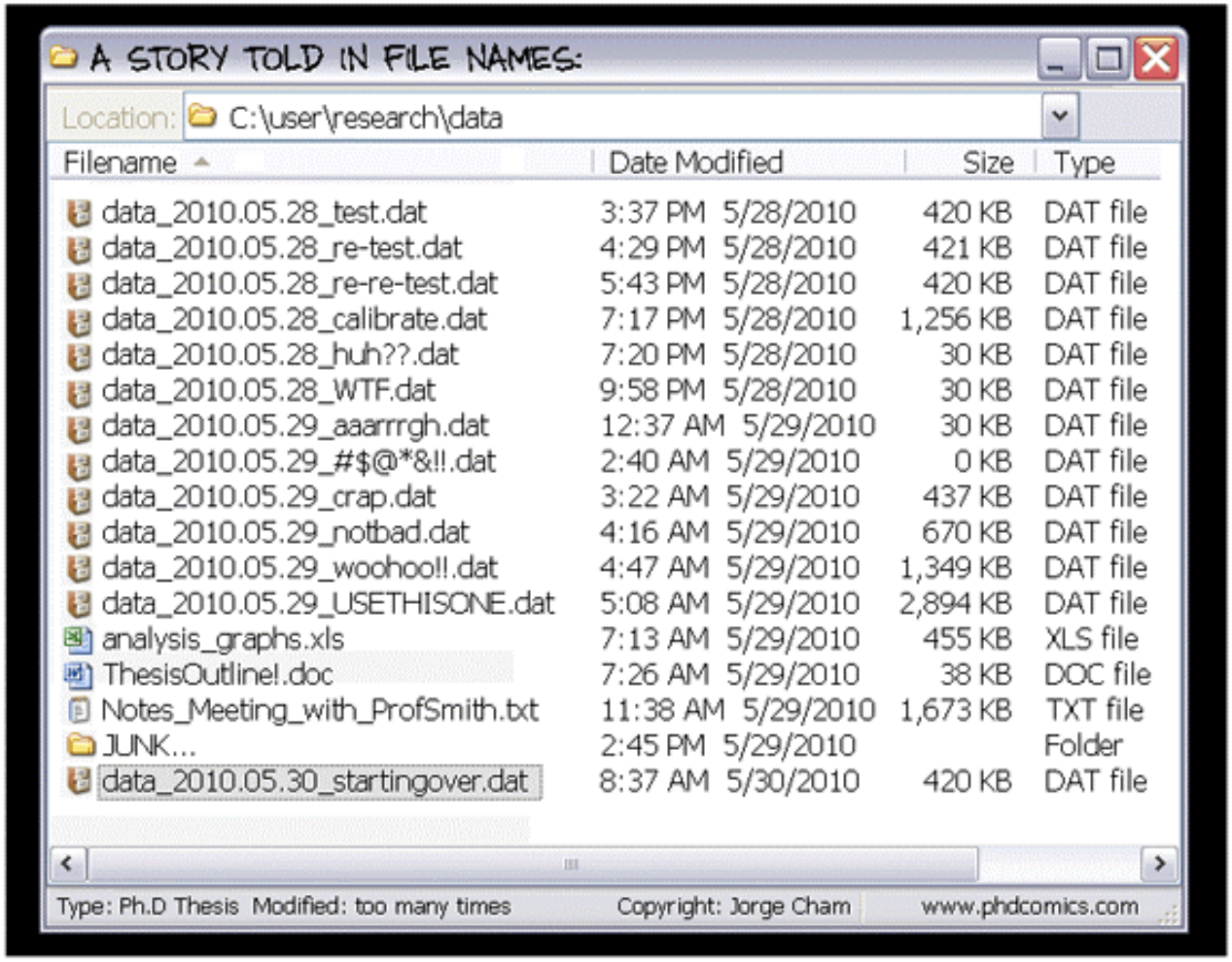


Adapted from Baum and Bromiley (2019)

Publication bias in social sciences (Mervis, 2014)



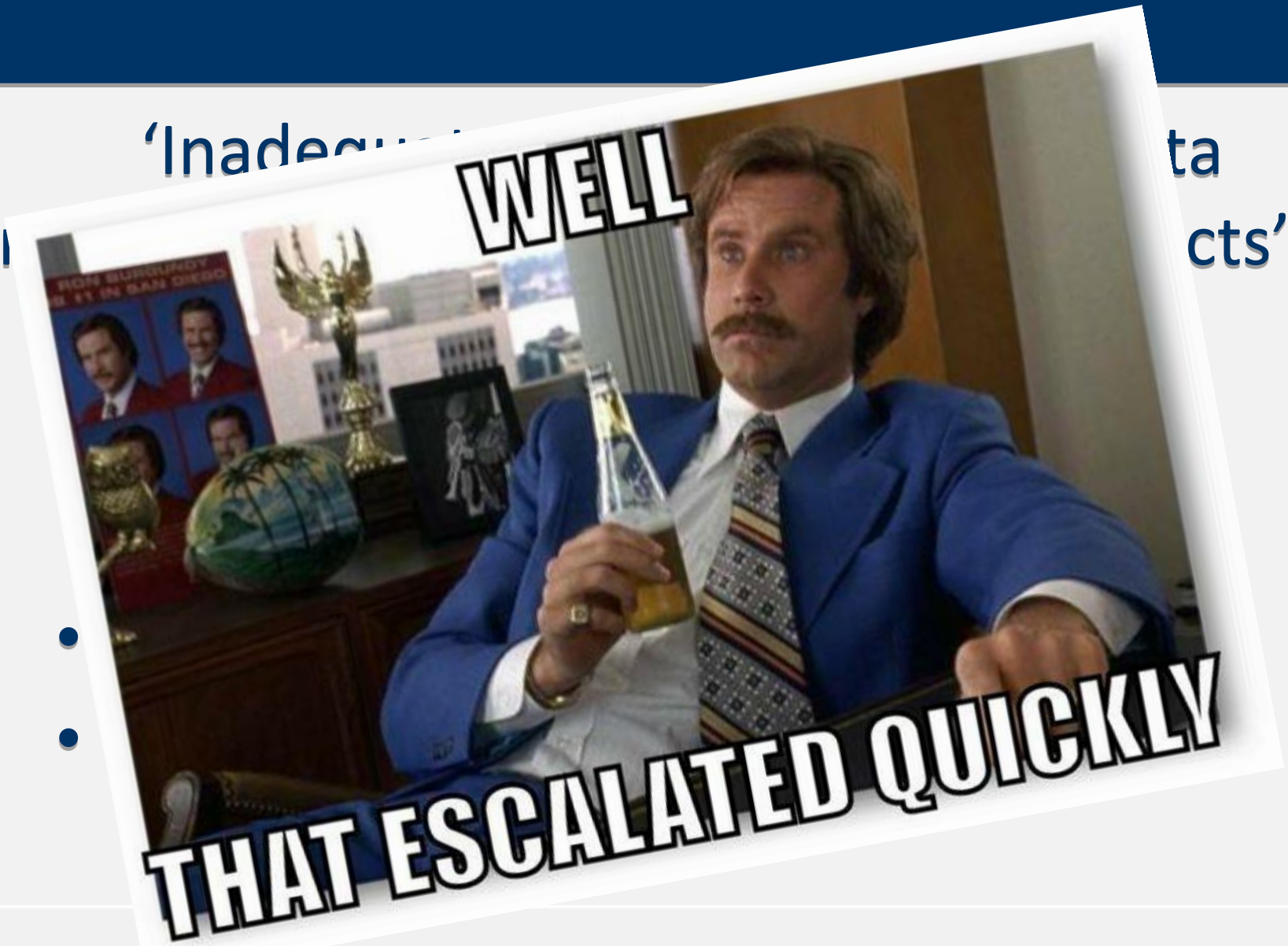
Finally, one practical problem



‘Inadequate record keeping or data management related to research projects’

- 27,5% (Martinson et al., 2005)
- 48%! (Godecharle et al., 2017)

‘Inadequ ta
cts’



So we know it’s bad..

How about some solutions?

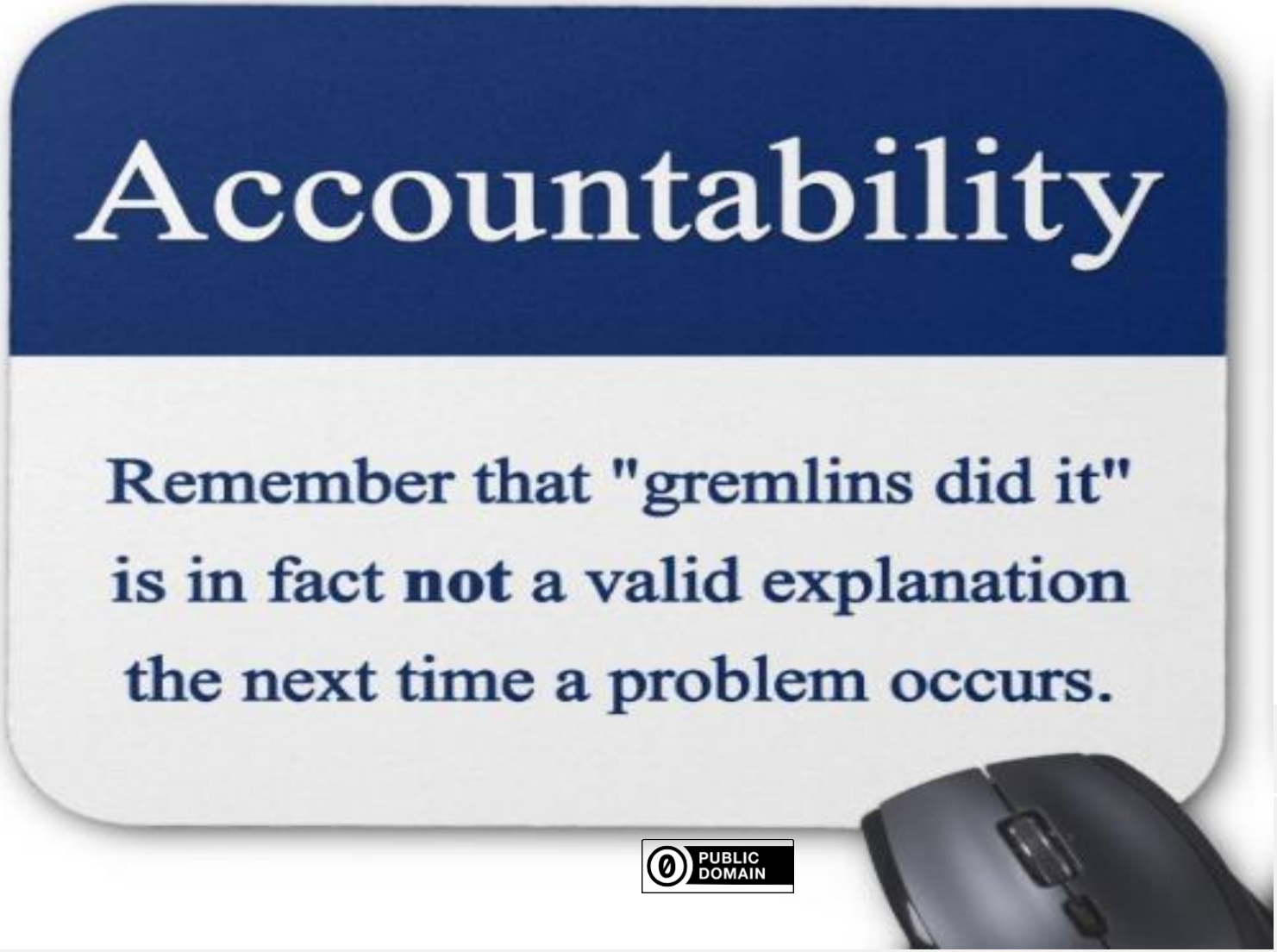
Transparency & Accountability

Responsible research conduct

High transparency

No transparency

Responsible research conduct



Some personal advantages of being open

- OA publications have higher citation rates (Tennant et al., 2016; Piwowar et al., 2018)
- "Economics journals with reproducibility policies are cited more often than others." (Höffler, 2017, AER)
- "I appreciate the fact that you registered your hypotheses and data in a public database" (personal comm. with SMJ editor during R&R)
- Scooping paradox: when open, you have proof you were the first

Say that you wake up one morning with full amnesia related to your current project, but you are still a very capable scientist. Are you able to understand what you did the day before?

How does one achieve T&A?

Today, we'll focus on three things:

- Version control
- Preregistration
- Reproducibility

A user-friendly platform: OSF

Open Science Framework

- Online environment
- Version control
- Preregistration
- Can make files public but not mandatory
- Watch out: US servers (privacy!)
- More advanced? Github/Bitbucket

How does one achieve T&A?

Version control

=

Track changes for files

‘Normal’ version control

manuscript.pdf

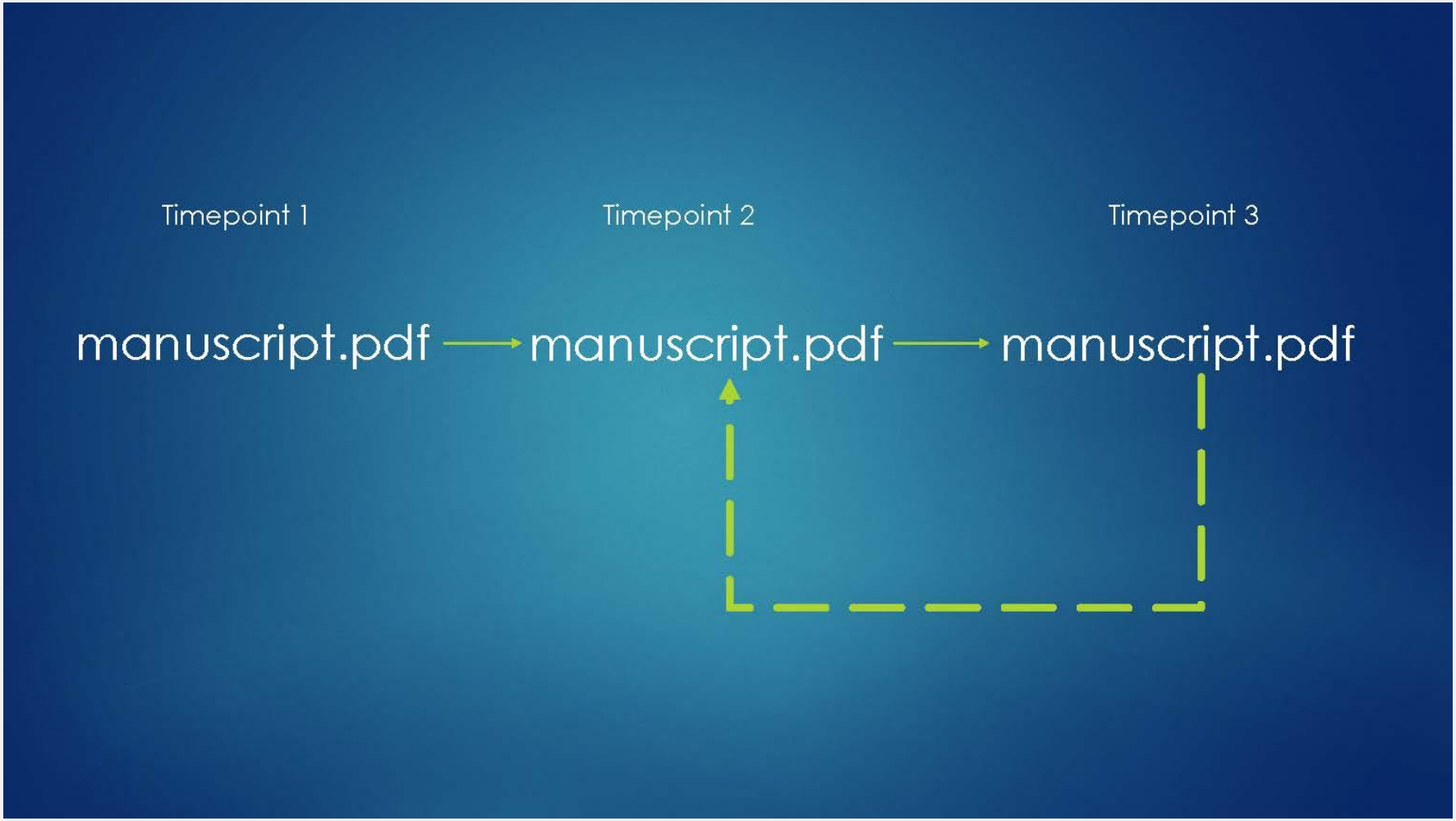
manuscript_final.pdf

manuscript_final2.pdf

manuscript_final3.pdf

Adapted from Hartgerink (2016): http://bit.ly/chrisPM101

Actual version control



Adapted from Hartgerink (2016): http://bit.ly/chrisPM101

Recent Activity

Marino van Zelst added file S5. Meta-regression operationalization.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S1. Papers by journal.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S3. Results of HOMA.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst removed file S1. Appendix Outlier Analyses.docx from OSF Storage in supplements
2016-05-10 06:02 PM

Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-10 06:02 PM

Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-09 08:26 PM

< 1 ... 4 5 6 ... 75 >

Recent Activity

Marino van Zelst added file S5. Meta-regression operationalization.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S1. Papers by journal.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S3. Results of HOMA.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst removed file S1. Appendix Outlier Analyses.docx from OSF Storage in supplements
2016-05-10 06:02 PM

Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-10 06:02 PM

Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-09 08:26 PM

< 1 ... 4 5 6 ... 75 >

Revisions

Version ID	Date	User	Download
13	2016-07-15 02:54 PM	Marino van Zelst	0
12	2016-05-27 09:13 PM	Marino van Zelst	2
11	2016-05-25 08:36 PM	Marino van Zelst	1
10	2016-05-10 06:09 PM	Marino van Zelst	3
9	2016-05-10 06:02 PM	Marino van Zelst	0
8	2016-05-09 08:26 PM	Marino van Zelst	0
7	2016-05-08 12:30 PM	Marino van Zelst	0
6	2016-05-04 09:18 PM	Marino van Zelst	0
5	2016-05-03 09:21 PM	Marino van Zelst	0
4	2016-05-03 11:29 AM	Marino van Zelst	0
3	2016-04-29 08:07 PM	Marino van Zelst	0
2	2016-04-25 07:38 PM	Marino van Zelst	0
1	2016-04-25 04:50 PM	Marino van Zelst	0

Preregistration

- Preregistration limits HARKing which is still practiced a lot (Banks et al., 2016; Bosco et al., 2016)

- Publicly register confirmatory and exploratory hypotheses/data collection/analyses/etc.

- "In a world of transparent reporting, I choose preregistration as a way to selfishly show off that I predicted the outcome of my study." (Datacolada.org, 2014)

OSF HOME My Projects Search Support Donate

preregister hypotheses Files Wiki Analytics Registrations Forks Contributors Settings

Registrations

New registration

Embargoed | preregister hypotheses | Registered: 2016-02-22 13:58 UTC

van Zelst, Oerlemans & Mannak

6 contributions

preregister hypotheses Files Wiki Analytics Forks Contributors Settings

This registration is a frozen, non-editable version of this project

This registration is currently embargoed. It will remain private until its embargo end date, Saturday, Feb 01, 2020.

TILBURG UNIVERSITY

How does one achieve transparency?

Ask ourselves continuously:

Can my future-self or anyone else reproduce this result within a reasonable amount of time?

Reproducibility

Data package requires you to:

- Be able to reproduce all your results in reasonable amounts of time

Reproducibility

Data package requires you to:

- Be able to reproduce all your results in reasonable amounts of time

- Use syntax for everything!
- Log WHY you use specific syntax

Reproducibility

```
dat <- read.csv("masterdata.csv",header=TRUE)
# Splined + separates #
dat <- dat[ which(dat$perf_mix < 4 & dat$spline==1 & dat$spline_correct ==1),]
dat$var <- 1/(dat$sample_size_firm-3)

# Preparing moderators to be included in models. Model-specific moderators are prepared within separate sections. #
# Rescale median sample year by subtracting minimum year in sample for ease of interpretation #
dat$medyear <- ((dat$sample_start+dat$sample_end)/2-(min(dat$sample_start,na.rm=TRUE)))
dat$medyear <- dat$medyear - min(dat$medyear,na.rm=TRUE)
dat$medyearsq <- dat$medyear^2
```

```r
## Meta-regression for prior ties. Moderator analysis with tie purpose, median sample year, and full risk set. #
# Tests for difference in ES between R&D ties and investment ties (btt) #
dat$priorrtoz <- 0.5*log((1+dat$prior_form_r)/(1-dat$prior_form_r))
# Model 1. Includes omnibus-test (H0:B1=B2=B3=0) for R&D, manufacturing and investment ties #
prior_wls <- rma(priorrtoz,var,mods=~ resanddev + investment + manufacturing + medyear +fullrisk,method="DL",
  data=dat,btt=c(6,7))
summary(prior_wls,digits=3)

prior2_wls <- rma(priorrtoz,var,mods=~ priordicho + priorcount + published,method="DL",data=dat)
summary(prior2_wls,digits=3)
anova(prior2_wls,L=c(0,1,-1,0))

# Chisquare-test for difference between coefficients of manufacturing (2) and investment (3) ties #
anova(prior_wls,L=c(0,1,-1,0,0,0)) #Omnibus: R&D vs. Manufacturing
anova(prior_wls,L=c(0,0,1,-1,0,0)) #Omnibus: Manufacturing vs. Investment
anova(prior_wls,L=c(0,1,0,-1,0,0)) #Omnibus: Manufacturing vs. R&D
```

The future: dynamic documenting

- knitR, Rmarkdown

- Code: Responses to historical performance feedback are heterogeneous, as the Q-statistic is *`r round(all.hpfb_dv$QE,3)` (*p*-value = `r round(all.hpfb_dv$QEp,3)`)*
- Text: Responses to historical performance feedback are heterogeneous, as the Q-statistic is 392.468 (p-value = 0.003)

1) Will you be capable of reproducing your own results in half a year from now?

1) Will you be capable of reproducing your own results in half a year from now?

2) Will your co-authors be able to do this?

1) Will you be capable of reproducing your own results in half a year from now?

2) Will your co-authors be able to do this?

3) Will a colleague that is not a co-author?

1) Will you be capable of reproducing your own results in half a year from now?

2) Will your co-authors be able to do this?

3) Will a colleague that is not a co-author?

4) An independent researcher who's
in your field of expertise?

'Open science' is just 'science'

“Open science describes the practice of carrying out scientific research in a completely transparent manner, and making the results of that research available to everyone. Isn’t that just ‘science’?”

Mick Watson, Genome Biology 2015, 16: 101 doi:10.1186/s13059-015-0669-2

Adapted from McKiernan (2015)

Data audit guidelines (applicable for all TiU schools)
https://www.tilburguniversity.edu/research/social-and-behavioral-sciences/download-guideline-datapackage-tsb/

Find me if you want help with making your work (more) transparent and/or reproducible!

- J.m.vanzelst@uvt.nl
- @mzelst
- S6.04