increasing transparency through a multiverse analysis (and a few other things)

francis tuerlinckx, wolf vanpaemel, sara steegen, & andrew gelman

replication day, vvs-or, 2019

utrecht

KU LEUVEN

what makes you trust a finding?

a finding

Psychological Science
24(6) 1007–1016

DOI: 10.1177/0956797612466416
pss.sagepub.com
SAGE

The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

Kristina M. Durante[1], Ashley Rae[1], and Vladas Griskevicius[2]
[1]College of Business, University of Texas, San Antonio, and [2]Carlson School of Management, University of Minnesota

Abstract
Each month, many women experience an ovulatory cycle that regulates fertility. Although research has found that this cycle influences women's mating preferences, we proposed that it might also change women's political and religious views. Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single women and women in committed relationships. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney. In addition, ovulation-induced changes in political orientation mediated women's voting behavior. Overall, the ovulatory cycle not only influences women's politics but also appears to do so differently for single women than for women in relationships.

Keywords
evolutionary psychology, fertility, relationships, political attitudes, religiosity, religious beliefs

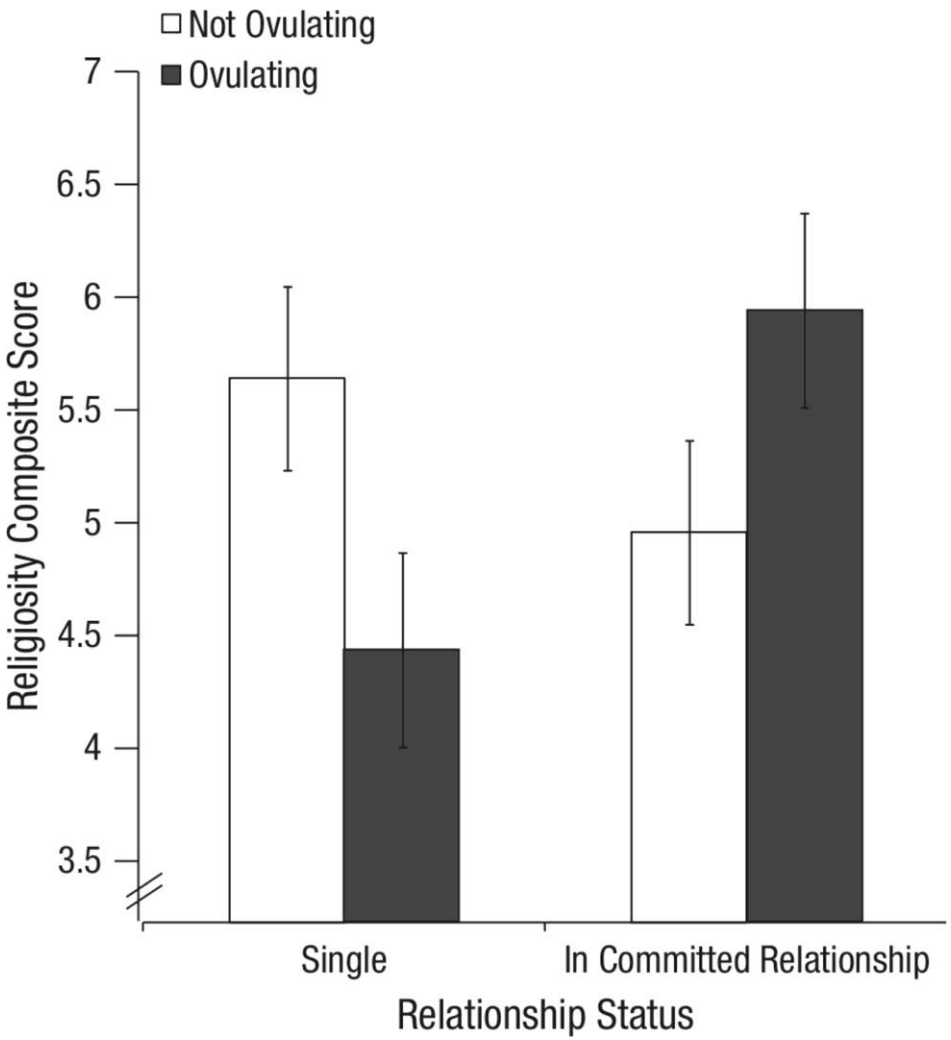
Received 7/30/12; Revision accepted 10/5/12

a finding

- we focus on religiosity in study 1 only

- analyses are based on the following data
  - relationship status (single vs committed)
  - fertility status (high vs low)
  - religiosity score

a finding



women's religiosity as a function of fertility and relationship status

fertility x relationship status interaction, $F(1,159)=6.46$, $p=.012$

can we trust this finding?

some basic checks

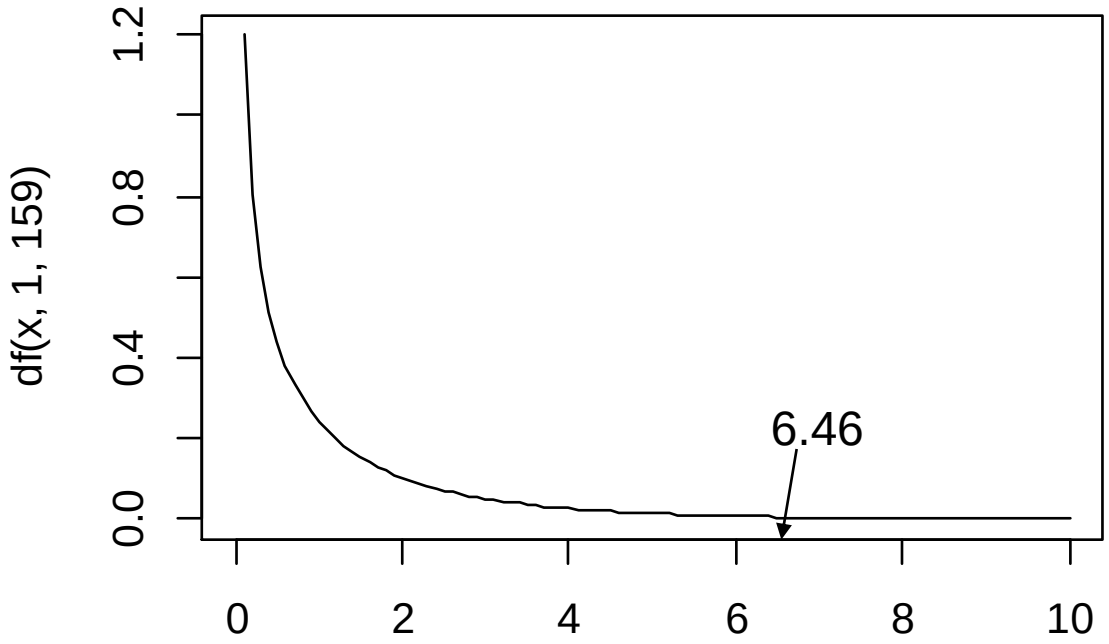
- has it been *peer-reviewed*?
  - let's check: yes
  - important because: a little
- has it been published in a *high-impact journal*?
  - let's check: yes (4.940)
  - important because: not
- has it been *cited* a lot?
  - let's check: quite a bit (102 on google scholar)
  - important because: not
- did it appear in the *media*?
  - let's check: hell, yes
  - important because: not

- are the analyses correct and correctly reported?
  - important because: duh!

- are the analyses correct and correctly reported?
  - let's check 0:
    - was there a co-pilot?
      - a person who independently analyzed the data
      - preferably using another language (R, python, SPSS, SAS, etc)
    - in this case: not mentioned, so probably not

- are the analyses correct and correctly reported?
  - let's check 1:
    - check degrees of freedom
    - n=81 (single) +82 (committed) =163
    - df interaction term: (2-1)x(2-1)=1
    - df error term: 163-2x2=159
    - F(1,159)

- are the analyses correct and correctly reported?
  - let's check 2a:
    - re-compute p-values based on summary statistics and degrees of freedom by hand



- are the analyses correct and correctly reported?
  - let's check 2a:
    - re-compute p-values based on summary statistics and degrees of freedom by hand
    - in R pf: given an x value, it returns the probability of having a value lower than x

```
1-pf(6.46,1,159)
0.01198962
```

    - p=.012

- are the analyses correct and correctly reported?
  - let's check 2b:
    - re-compute p-values based on summary statistics and degrees of freedom automatically
    - statcheck.io
    - it flags two (less important) p-values as being wrong
      - probably typos, that don't change any conclusions

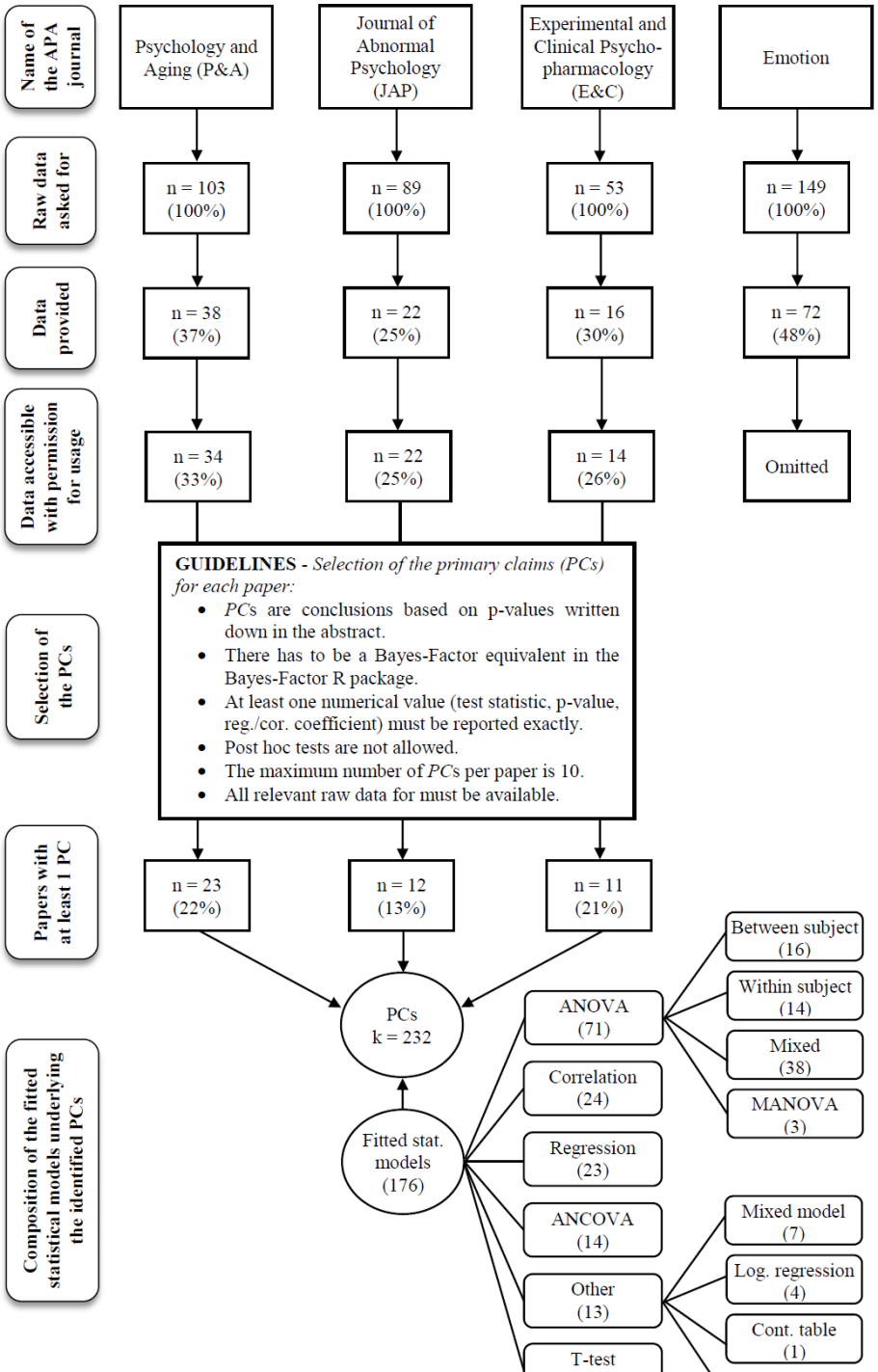
- are the analyses correct and correctly reported?
  - let's check 3:
    - redo the analyses based on the original raw data
      - aka check the *reproducibility*
    - the data are publically available (https://osf.io/hj9gr/)
    - redoing the analyses in R yields the same main results
      - at least, after correcting a few typos
        - impossible dates, …

(thanks to Kristina Durante for sharing the data)

- are the analyses correct and correctly reported?
  - if you can't reproduce a result, it's not definitely wrong
    - there might be software differences
      - this doesn't speak to the trustworthiness of the result
    - you might have done something wrong
      - this probably indicates the authors didn't provide enough detail about their analyses

digression

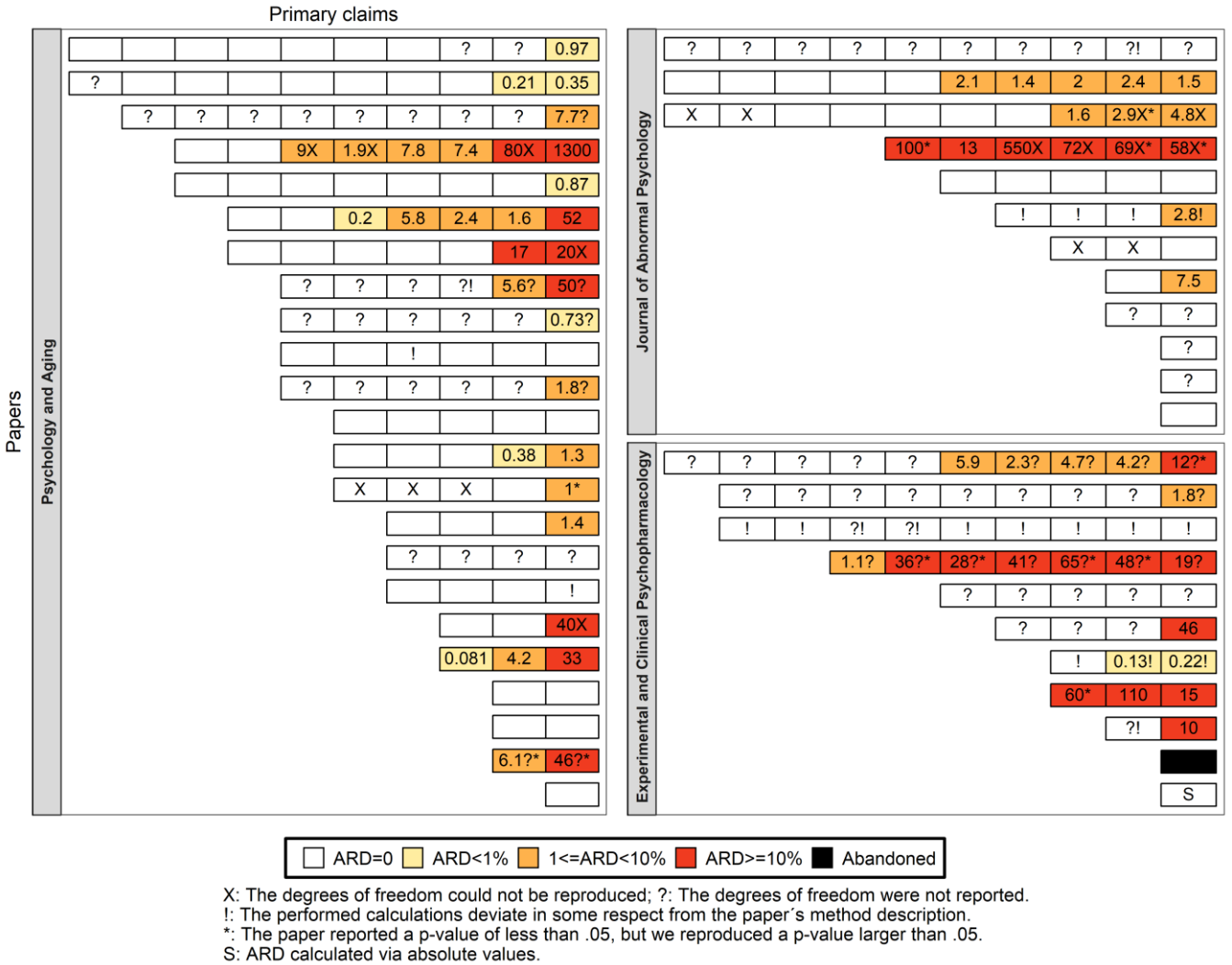
systematic reproducibility study (artner et al., 2019)



digression

artner et al.

(2019)



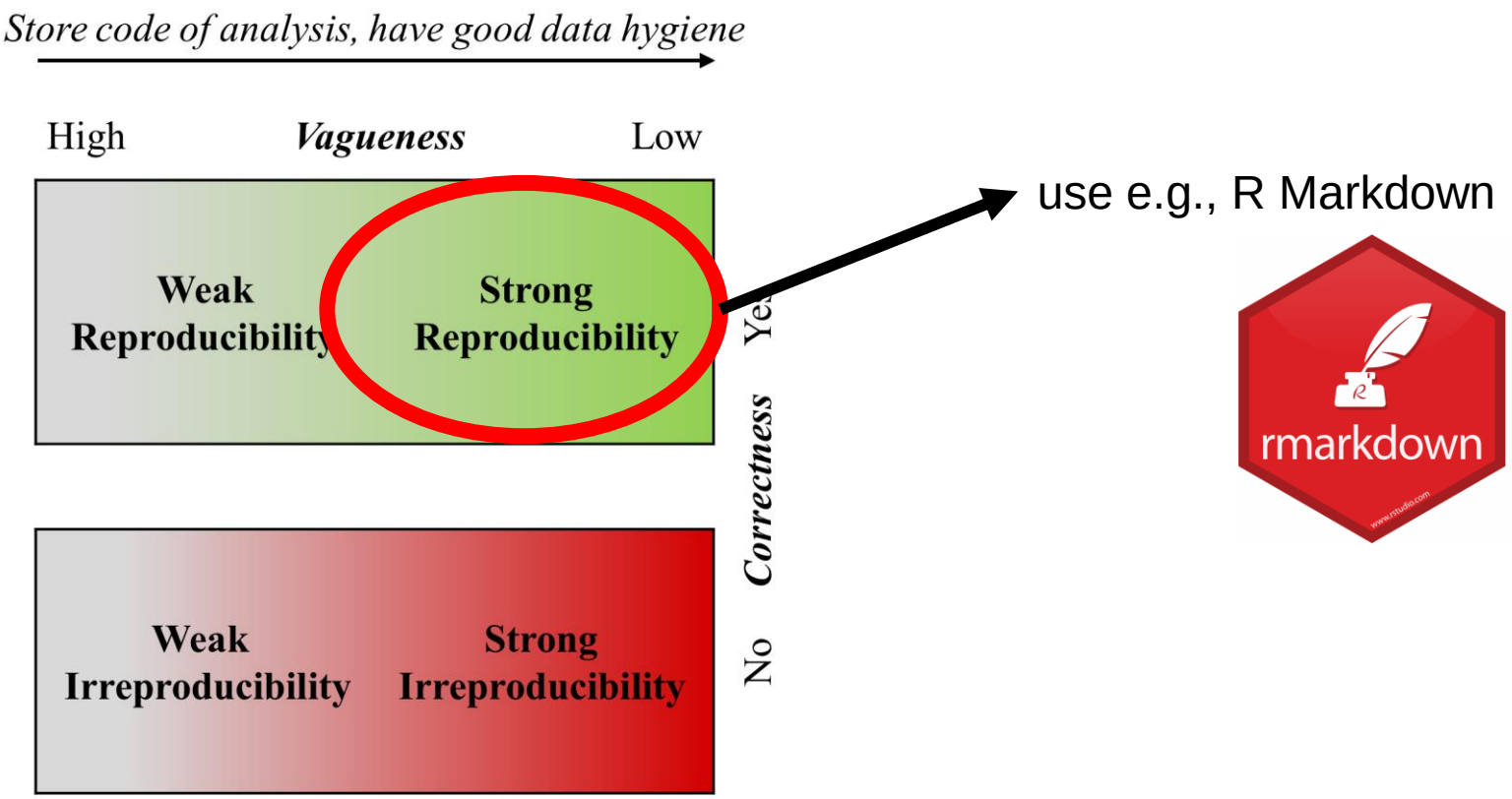
digression

some reasons for errors :

- rounding rounded results (T = 3.41461880... → T = 3.415 → T = 3.42)
- related: calculating with rounded numbers
- incorrect selection of variables/cases (what is reported ≠ what is done)
- incorrect labeling of variables or numerical results
- typos
- copy-paste errors

but the main underlying issue is ...

digression


Store code of analysis, have good data hygiene
High
Vagueness
Low
Weak Reproducibility
Strong Reproducibility
Weak Irreproducibility
Strong Irreproducibility
Correctness
No
use e.g., R Markdown
rmarkdown

what makes you trust a finding?

- has it been *peer-reviewed*?
- has it been published in a *high-impact journal*?
- has it been *cited* a lot?
- did it appear in the *media*?
- are the analyses correct and correctly reported?

- are the statistical conclusions robust against arbitrary data-processing and data-analytical decisions?
  - important because: often, there is a lot of arbitrariness in data processing, which is inherited by the statistical result
  - if your data are arbitrary, so is your statistical result
  - let's check:

- analyses are based on the following 'observed data'
  - relationship status (single vs committed)
  - fertility status (high vs low)
  - religiosity score

- but these are not the data actually observed

- the observed, raw data include
  - answer to three statements on religiosity
  - answer to several fertility related questions
    - the start of the last period
    - the start date of the period before the last period
    - the typical cycle length
    - the start of the next period
    - how sure are you about the start of the last period
    - how sure are you the start date of the period before the last period
  - answer to "what is your current romantic relationship status?"
    - (1) not dating/romantically involved with anyone
    - (2) dating or involved with only one partner
    - (3) engaged or living with my partner
    - (4) married

fertility status?

answer to fertility related questions

- the start of the last period
- the start date of the period before the last period
- the typical cycle length
- the start of the next period
- how sure are you about the start of the last period
- how sure are you the start date of the period before the last period

cycle length →
next menstrual
onset → cycle day

high in fertility when cycle day is between 7 and 14

low in fertility when cycle day is between 17 and 25

relationship status?

answer to “what is your current romantic relationship status?”

- (1) *not dating/romantically involved with anyone* — single
- (2) *dating or involved with only one partner* — single
- (3) *engaged or living with my partner* — committed
- (4) *married* — committed

translating the observed, raw data to the processed data ready for analysis involved several choices

the observed data are more **constructed** rather than observed

the original data construction choices seem reasonable-ish

but other data construction choices are reasonable too

fertility status?

answer to fertility related questions

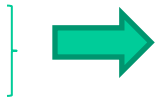
- the start of the last period
- the start date of the period before the last period
- the typical cycle length
- the start of the next period
- how sure are you about the start of the last period
- how sure are you the start date of the period before the last period

cycle length → next menstrual onset → cycle day

fertility status?

answer to fertility related questions

- the start of the last period
- the start date of the period before the last period
- the typical cycle length
- the start of the next period
- how sure are you about the start of the last period
- how sure are you the start date of the period before the last period



next menstrual onset → cycle day

This app shows how different choices in constructing the data leads to different analysis results.

Dependent Variable ?

Religiosity Study 1

Next menstrual onset ?

NMO1

Relationship status ?

R1

Fertility ?

F1

Exclusion of women based on cycle length ?

ECL1

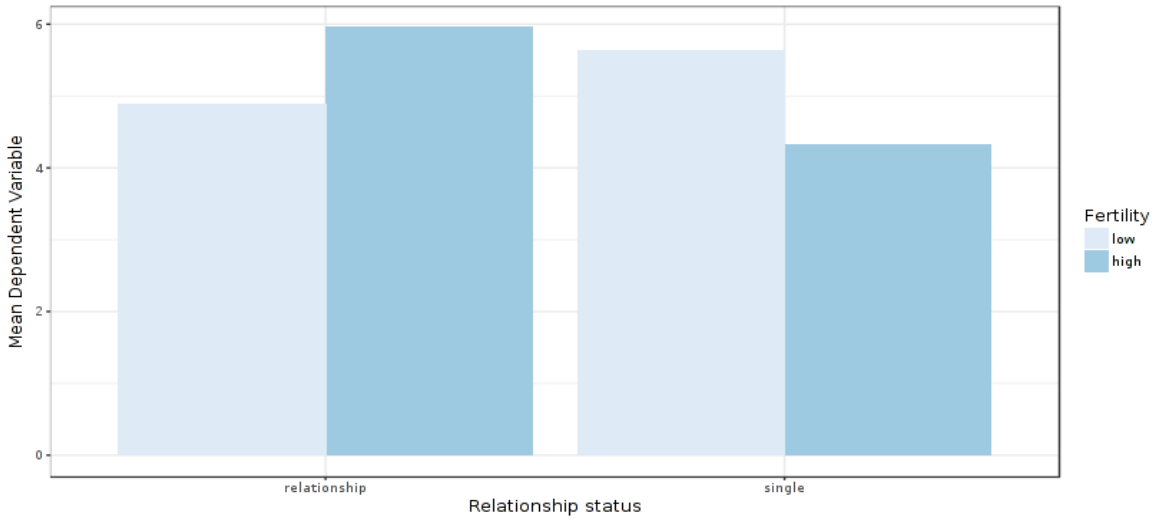
Exclusion of women based on certainty ratings ?

EC1

General info | Single data set | Single data set analysis | Multiverse

Graphical summary

The barplot represent the dependency between the chosen dependent variable and the relationship status and fertility status (n= 163)



ANOVA results

Effects	F-value	DF	SS	MS	p-value
Relationship	0.746	1	6.633	6.633	0.389
Fertility	0.061	1	0.543	0.543	0.805
Interaction	6.46	1	57.408	57.408	0.012

This app shows how different choices in constructing the data leads to different analysis results.

Dependent Variable ?

Religiosity Study 1

Next menstrual onset ?

NMO2

Relationship status ?

R1

Fertility ?

F1

Exclusion of women based on cycle length ?

ECL1

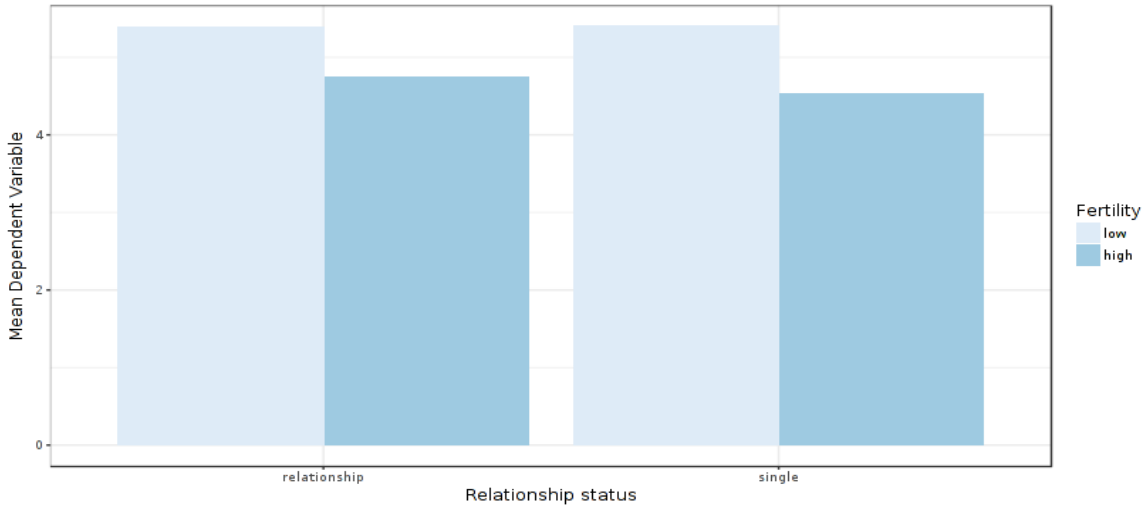
Exclusion of women based on certainty ratings ?

EC1

General info | Single data set | Single data set analysis | Multiverse

Graphical summary

The barplot represent the dependency between the chosen dependent variable and the relationship status and fertility status (n= 131)



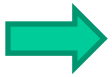
ANOVA results

Effects	F-value	DF	SS	MS	p-value
Relationship	0.044	1	0.399	0.399	0.835
Fertility	2.116	1	19.362	19.362	0.148
Interaction	0.043	1	0.396	0.396	0.835

fertility status?

answer to fertility related questions

- the start of the last period
- the start date of the period before the last period
- the typical cycle length
- the start of the next period

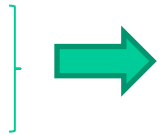
}  cycle day

- how sure are you about the start of the last period
- how sure are you the start date of the period before the last period

fertility status?

answer to fertility related questions

- the start of the last period
- the start date of the period before the last period
- the typical cycle length
- the start of the next period
- how sure are you about the start of the last period
- how sure are you the start date of the period before the last period



cycle length → next menstrual onset → cycle day

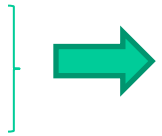
high in fertility when cycle day is between 7 and 14

low in fertility when cycle day is between 17 and 25

fertility status?

answer to fertility related questions

- the start of the last period
- the start date of the period before the last period
- the typical cycle length
- the start of the next period
- how sure are you about the start of the last period
- how sure are you the start date of the period before the last period



cycle length → next menstrual onset → cycle day

high in fertility when cycle day is between 6 and 14

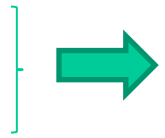
low in fertility when cycle day is between 17 and 27

durante et al., 2011

fertility status?

answer to fertility related questions

- the start of the last period
- the start date of the period before the last period
- the typical cycle length
- the start of the next period
- how sure are you about the start of the last period
- how sure are you the start date of the period before the last period



cycle length → next menstrual onset → cycle day

high in fertility when cycle day is between 9 and 17

low in fertility when cycle day is between 18 and 25

durante et al., 2012

relationship status?

answer to “what is your current romantic relationship status?”

(1) *not dating/romantically involved with anyone*
(2) *dating or involved with only one partner*

single

(3) *engaged or living with my partner*
(4) *married*

committed

relationship status?

answer to “what is your current romantic relationship status?”

- (1) *not dating/romantically involved with anyone* — single
- (2) *dating or involved with only one partner* — committed
- (3) *engaged or living with my partner* — committed
- (4) *married* — committed

relationship status?

answer to “what is your current romantic relationship status?”

- (1) *not dating/romantically involved with anyone* } single
- (2) *dating or involved with only one partner*
- (3) *engaged or living with my partner* } committed
- (4) *married*

who to include?

only women who are reasonably sure about their start dates

only women who have regular cycle lengths

- the estimated cycle length
- the reported cycle length

relationship status assessment (3 choice options)

fertility assessment (5 choice options)

cycle day assessment (3 choice options)

exclusion criteria based on certainty (2 choice options)

exclusion criteria based on cycle length (3 choice options)

all choices have been used in other studies and seem reasonable

each combination of choices gives rise to a separate data set

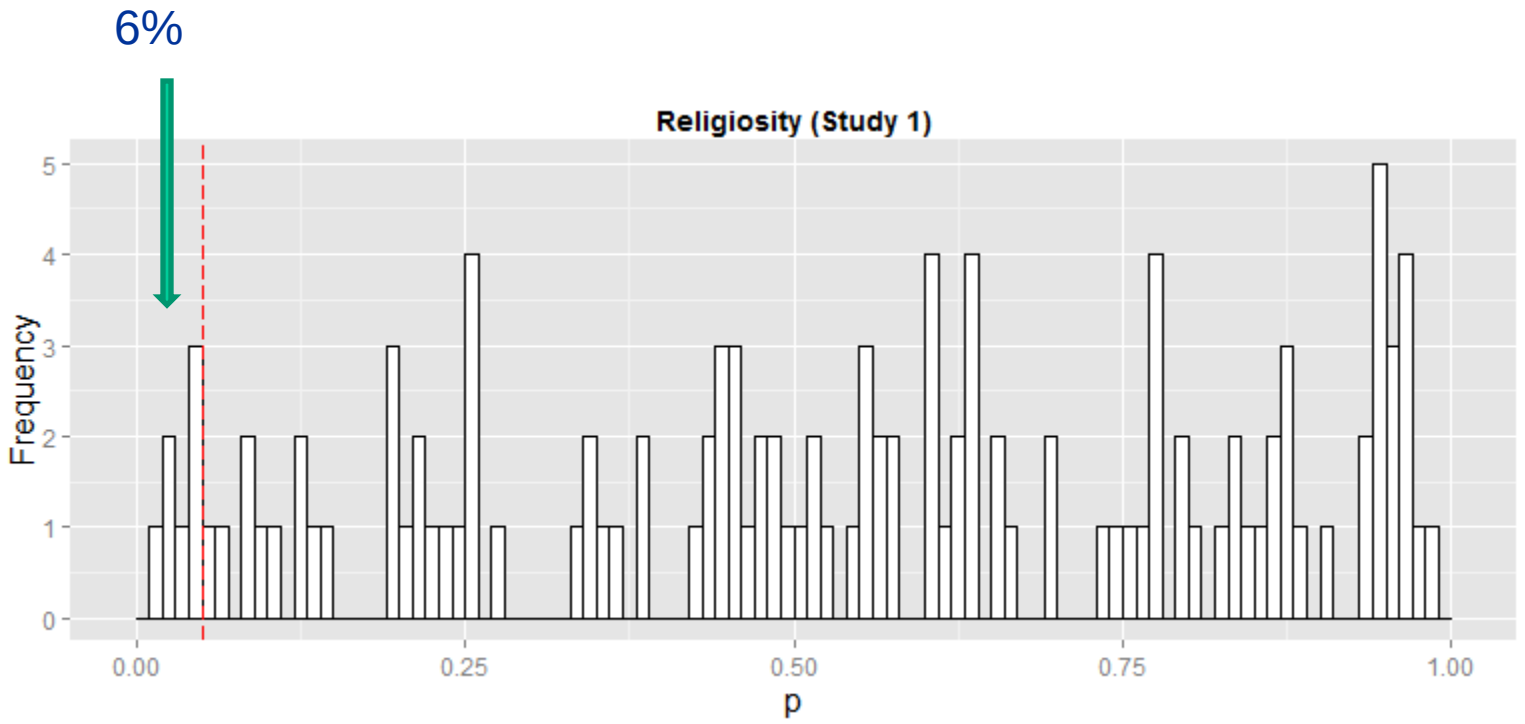
→ a multiverse of > 100 reasonable data sets

→ a multiverse of statistical results

if there are no good reasons to prefer a data processing choice over another one, there is no good reason to prefer a data set, and a statistical result over another one

let's look at all reasonable results

effect is too fragile to be taken seriously



Steegen, Tuerlinckx, Gelman, & Vanpaemel (2016).
see https://r.tquant.eu/KULeuven/Multiverse/

digression

- arbitrariness shows up at several levels

  -design of the study

  -preprocessing the data

  -analysis method

digression

- arbitrariness shows up at several levels

  -design of the study

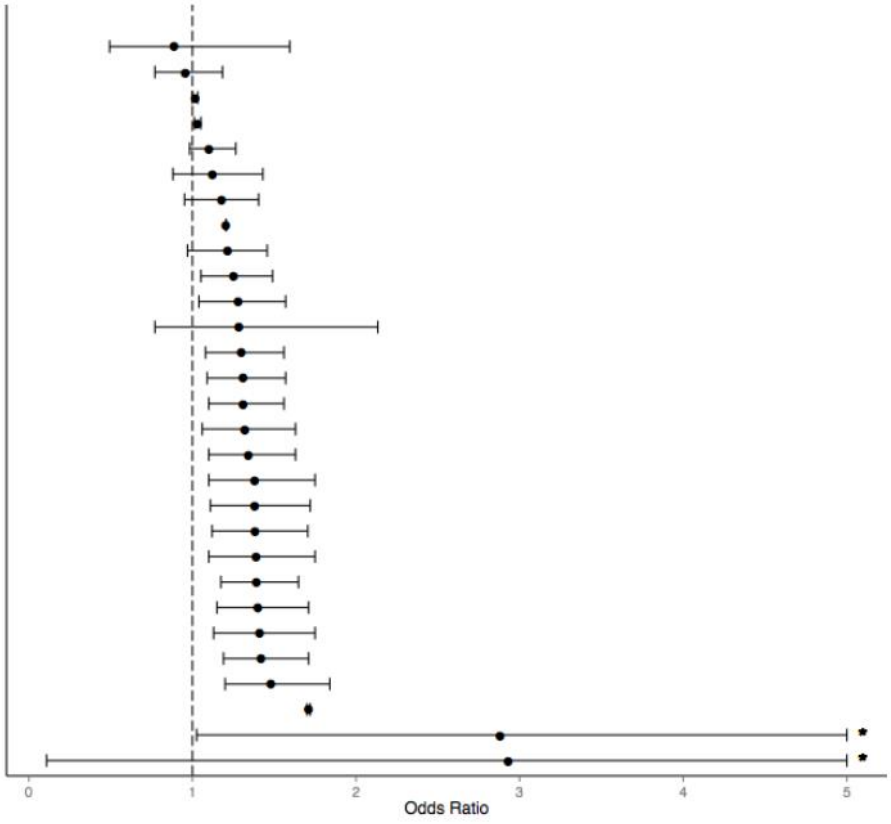
  -preprocessing the data

  -analysis method

digression

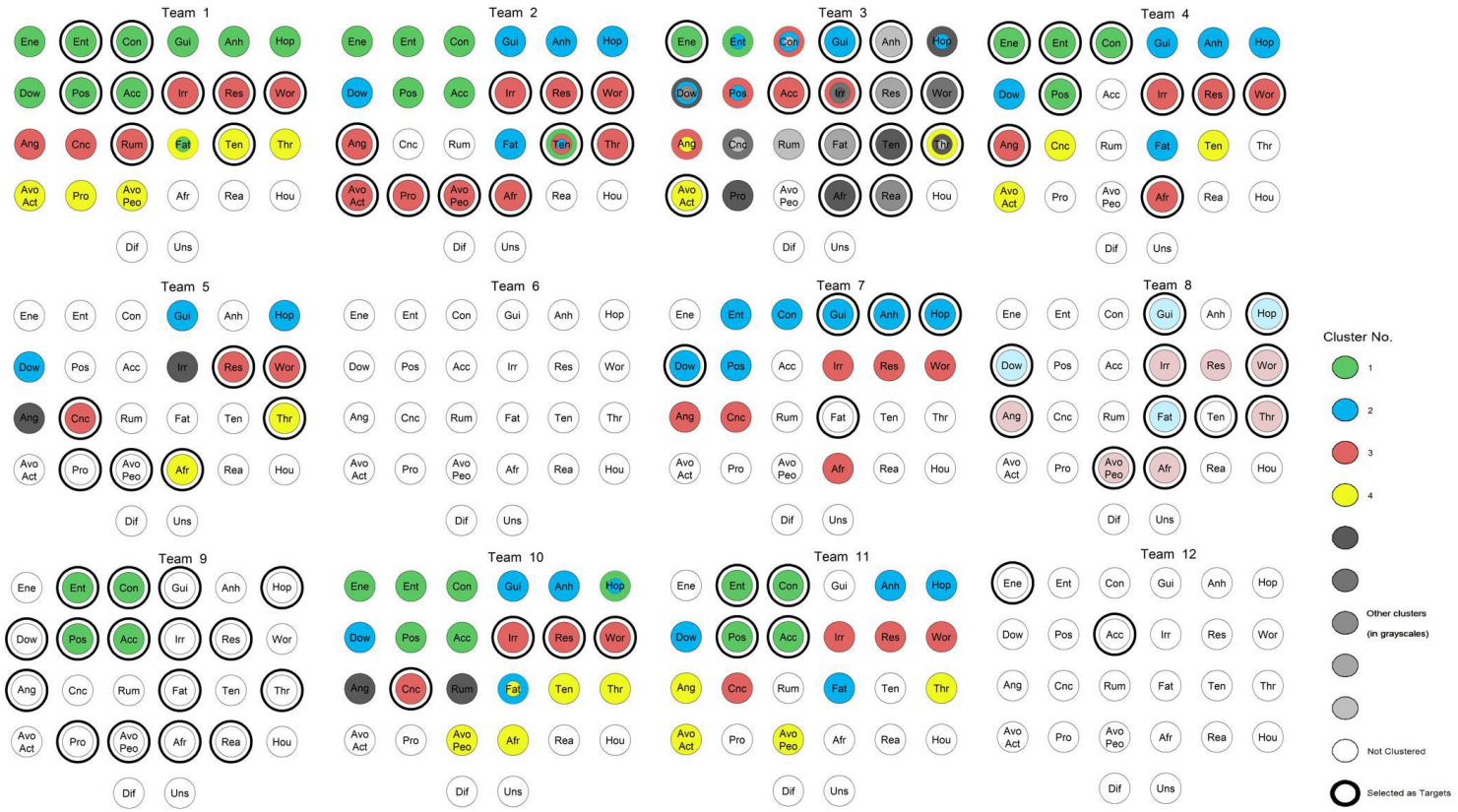
Many analysts, one dataset: are soccer referees more likely to give red cards to dark skin toned players than light skin toned players? (Silberzahn et al., 2018)

Team	Analytic Approach	OR
12	Zero-inflated Poisson regression	0.89
17	Bayesian logistic regression	0.96
15	Hierarchical log-linear modeling	1.02
10	Multilevel regression and logistic regression	1.03
18	Hierarchical Bayes model	1.10
31	Logistic regression	1.12
1	Ordinary least squares with robust standard errors, logistic regression	1.18
4	Spearman correlation	1.21
14	Weighted least squares regression with referee fixed-effects and clustered standard errors	1.21
11	Multiple linear regression	1.25
30	Clustered robust binomial logistic regression	1.28
6	Linear Probability Model	1.28
26	Three-level hierarchical generalized linear modeling with Poisson sampling	1.30
3	Multilevel Binomial Logistic Regression using bayesian inference	1.31
23	Mixed model logistic regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear probability model, logistic regression	1.34
5	Generalized linear mixed models	1.38
24	Multilevel logistic regression	1.38
28	Mixed effects logistic regression	1.38
32	Generalized linear models for binary data	1.39
8	Negative binomial regression with a log link analysis	1.39
20	Cross-classified multilevel negative binomial model	1.40
13	Poisson Multi-level modeling	1.41
25	Multilevel logistic binomial regression	1.42
9	Generalized linear mixed effects models with a logit link function	1.48
7	Dirichlet process Bayesian clustering	1.71
21	Tobit regression	2.88
27	Poisson regression	2.93



digression

the impact of researchers' choices on the selection of treatment targets using the experience sampling methodology (Bastiaansen et al., 2019)



what makes you trust a finding?

- has it been *peer-reviewed*?
- has it been published in a *high-impact journal*?
- has it been *cited* a lot?
- did it appear in the *media*?
- are the analyses correct and correctly reported?
- are the statistical conclusions robust against arbitrary data-processing and data-analytical decisions?

- is the study transparent about researchers degrees of freedom?
  - maybe some "bad" participants were excluded
    - outlying data
    - didn't follow instructions
    - etc
  - maybe there was a second measure for religiosity, for which the effect was not found (selective reporting)
  - maybe the effect was not found after the initial data collection (e.g., 100 women), and more data were collected until the desired effect was found (data peeking; optional stopping)
  - …

- is the study transparent about researchers degrees of freedom?
  - *important because*: exploiting *researchers degrees of freedom* increase the false positive rate (incorrect rejections of the null hypothesis)

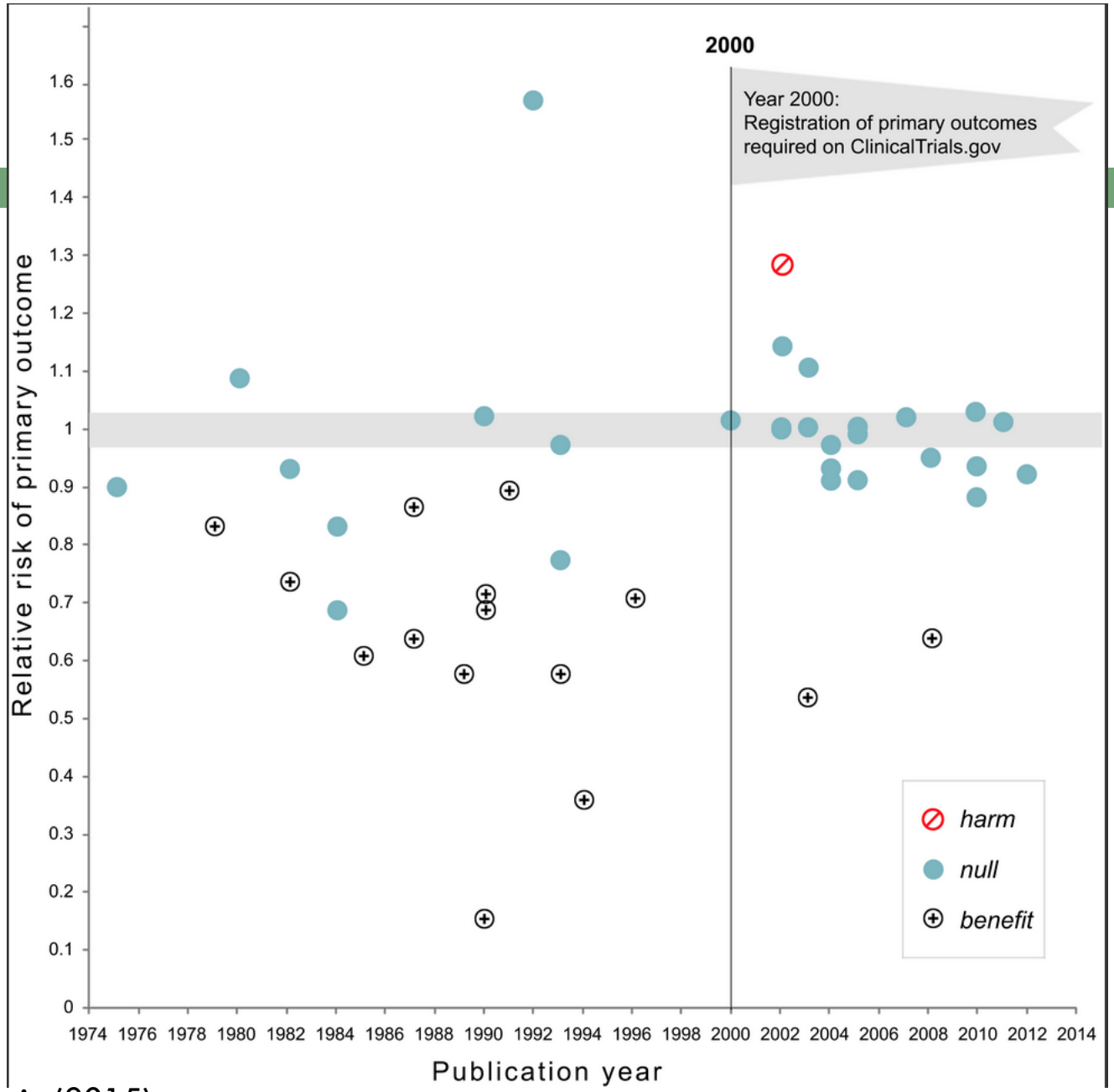
Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three *t* tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one *t* test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a *t* test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender × Condition interaction was significant. Results for Situation D were obtained by conducting *t* tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = −1, medium = 0, high = 1).

Joseph P. Simmons; Leif D. Nelson; Uri Simonsohn; *Psychological Science* 22, 1359-1366.
DOI: 10.1177/0956797611417632

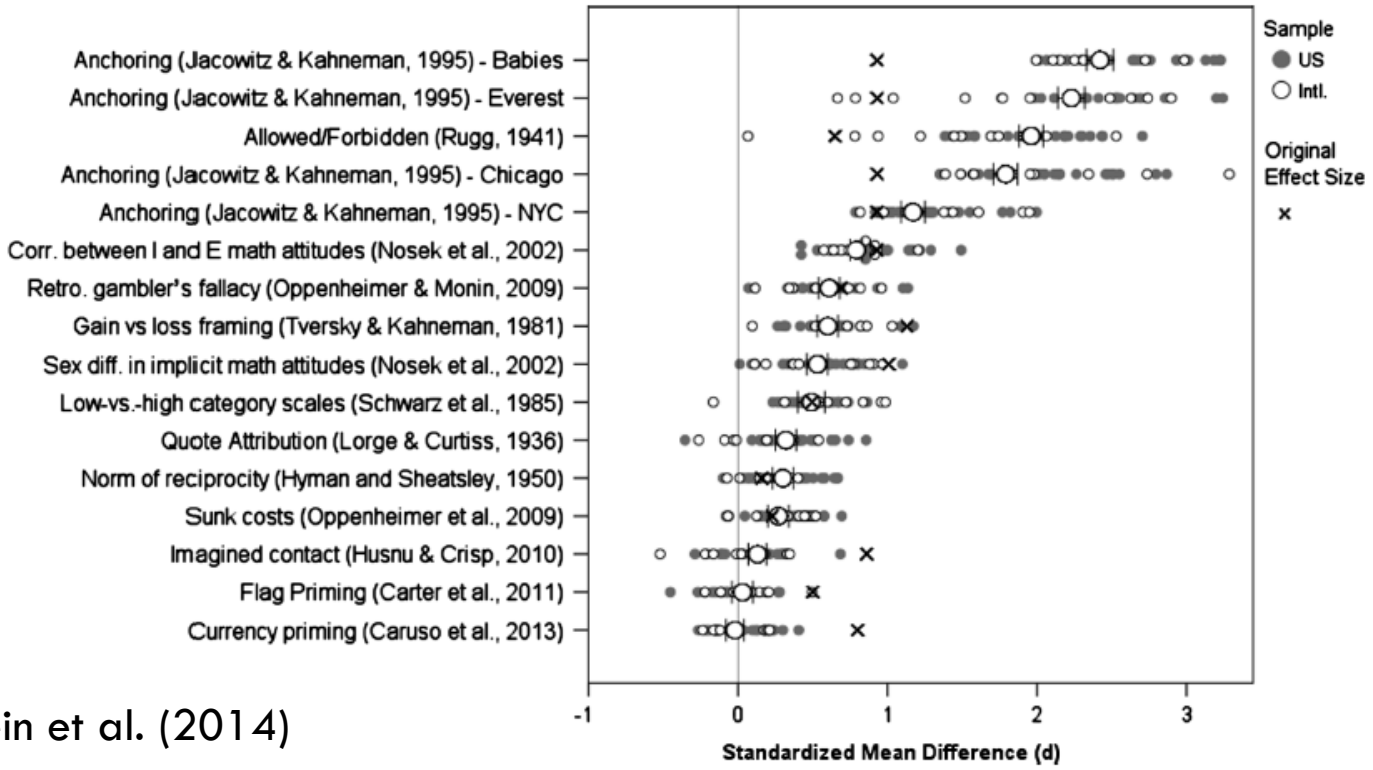
- is the study transparent about researchers degrees of freedom?
  - let's check: no mention of preregistration
    - a publically available, uneditable, time-stamped description of the hypotheses and analyses *before* data collection
    - to be fair, pre-registration was rare to non-existent at the time
    - since 2014, papers in *Psychological Science* with at least one pre-regsitered study receive a badge



Kaplan & Irvin (2015)

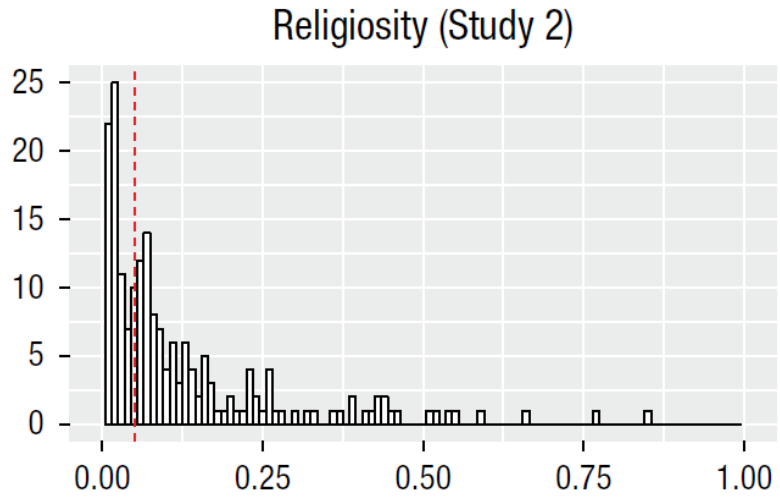
- is the study transparent about researchers degrees of freedom?
  - note that being preregistered doesn't mean that researchers degrees of freedom were not exploited
    - maybe the preregistration protocol was not concisely followed

- has the finding been replicated?
  - important because: no single study is conclusive on its own



Klein et al. (2014)

- has the finding been replicated?
  - let's check:
    - admirably, in-paper replication (study 2) and also the multiverse analysis looks better



    - but failed replication in harris, chabot and mickes (2014)
    - but replicated again in durante, et al. (2014)

- does the finding make theoretical sense?
  - important because: good theory is a filter for nonsense
  - let's check:
    - in the original paper's introduction:

    "The driving theory behind this research is that ovulation should lead women to prioritize the securement of genetic benefits from a mate who possesses indicators of genetic fitness"

      - *"Given that …, ovulation may lead women to become less religious"*
      - *"Because …, ovulation might lead married women to become more religious"*

- does the finding make theoretical sense?
  - let's check:
    - in a later reply to a commentary:

"Fertility had the predicted effect [ovulation may lead women to become less religious] for single women, but to our surprise had the opposite effect for women in committed relationships."

    - the intro is a clear case of HARKing (Hypothesizing After the Results are Known (Kerr, 1998)

- does the finding make theoretical sense?
  - let's check:
    - also: within vs between participants

what makes you trust a finding?

- has it been *peer-reviewed*?
- has it been published in a *high-impact journal*?
- has it been *cited* a lot?
- did it appear in the *media*?
- are the analyses correct and correctly reported?
- are the statistical conclusions robust against arbitrary data-processing and data-analytical decisions?
- is the study transparent about researchers degrees of freedom?
- has the finding been replicated?
- does the finding make theoretical sense?

discussion

- our starting question was

  “what makes you trust a finding?”

- a finding = published finding by others

conclusion

- arbitrary choices at several levels
  1. design of the study
  2. preprocessing the data
  3. analysis method

discussion

- our starting question was

  “what makes you trust a finding?”

- a finding = published finding by others

discussion

- a more important question

  “what makes you trust your own finding?”
  “what makes others trust your finding?”

- robustness and its limits of your finding can be assessed and shown through a multiverse analysis

the end