MUNI FSS Katedra psychologie

Přednáška 2:

Replikovatelnost výzkumu a metaanalýza

21. 9. 2021 | PSYn4790 | **Psychometrika: Měření v psychologii**

Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler & Vít Gabrhel (i.m.) | hynek.cigler@mail.muni.cz

Acknowledgement:
Děkuji Vítu Gabrhelovi!

„No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon.“

FISHER, 1971, S. 13

Fisher, R. (1971). *The Design of Experiments*. Hafner Publishing Company.

Metaanalýza

Meta-analýza

„Věda má kumulativní povahu, ke studiím však přistupujeme nikoli jako jedné z mnoha, ale izolovaně, stojícími o sobě.“

- (Chalmers, cit. dle Borenstein et al., 2009)

Tradiční přístup: Narativní review

- Expert shrne poznatky k danému tématu a dojde k závěru.
- Subjektivita
- Proces rozhodování není popsán (není replikovatelný).
- Omezení při velkém množství zdrojů.
- Nedostatečné narativní postižení variability velikostí účinku.
- Narativní review je typicky součástí úvodu k empirickým článkům, DP...

Meta-analýza

Od cca 90. let přechod k meta-analýze a systematické review.

- Proces systematického vyhledávání, hodnocení a následné syntézy dat z velkého počtu zdrojů.

Systematická review.

- Jasně definovaná kritéria pro volbu studií a transparentní popis.
  - Volba kritérií stále zahrnuje určitou míru subjektivity.
- Obvykle zahrnuje meta-analýzu.
- Výběr studií by měl být replikovatelný.

Meta-analýza.

- Statistická syntéza předchozího výzkumu (ale existuje i „qualitative meta-analysis“).
- Význam té které studie je dán podle vnějších (matematických) pravidel.
- Cílem je odhad „souhrnné velikosti efektu“.

Ve skutečnosti existuje **velké množství designů** souhrnně řazených pod „systematickou review“.

Label	Description
Critical review	Aims to demonstrate writer has extensively researched literature and critically evaluated its quality. Goes beyond mere description to include degree of analysis and conceptual innovation. Typically results in hypothesis or model
Literature review	Generic term: published materials that provide examination of recent or current literature. Can cover wide range of subjects at various levels of completeness and comprehensiveness. May include research findings
Mapping review/ systematic map	Map out and categorize existing literature from which to commission further reviews and/or primary research by identifying gaps in research literature
Meta-analysis	Technique that statistically combines the results of quantitative studies to provide a more precise effect of the results
Mixed studies review/mixed methods review	Refers to any combination of methods where one significant component is a literature review (usually systematic). Within a review context it refers to a combination of review approaches for example combining quantitative with qualitative research or outcome with process studies
Overview	Generic term: summary of the [medical] literature that attempts to survey the literature and describe its characteristics
Qualitative systematic review/ /qualitative evidence synthesis	Method for integrating or comparing the findings from qualitative studies. It looks for 'themes' or 'constructs' that lie in or across individual qualitative studies
Rapid review	Assessment of what is already known about a policy or practice issue, by using systematic review methods to search and critically appraise existing research
Scoping review	Preliminary assessment of potential size and scope of available research literature. Aims to identify nature and extent of research evidence (usually including ongoing research)
State-of-the-art review	Tend to address more current matters in contrast to other combined retrospective and current approaches. May offer new perspectives on issue or point out area for further research
Systematic review	Seeks to systematically search for, appraise and synthesis research evidence, often adhering to guidelines on the conduct of a review
Systematic search and review	Combines strengths of critical review with a comprehensive search process. Typically addresses broad questions to produce 'best evidence synthesis'
Systematized review	Attempt to include elements of systematic review process while stopping short of systematic review. Typically conducted as postgraduate student assignment
Umbrella review	Specifically refers to review compiling evidence from multiple reviews into one accessible and usable document. Focuses on broad condition or problem for which there are competing interventions and highlights reviews that address these interventions and their results

Grant, M.J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies.
Health Information and Libraries Journal, 26(2), 91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Meta-analýza: Pojmy

Velikost účinku (Effect size)

- **Souhrnný efekt (summary effect)** – vážený průměr velikostí účinku dle stanovených pravidel.
- Jde vlastně o odhad „skutečného efektu“ (true effect).

Přesnost souhrnného efektu: **celkové *N***.

Váha dílčích studií: ***n* dané studie**.

Homogenita/heterogenita: Míra konzistence napříč studiemi.

Signifikance souhrnného efektu: často i grafická interpretace.

- Zpravidla intervaly spolehlivosti.

Meta-analýza: Příklad

Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies

Luke E. Taylor, Amy L. Swerdfeger, Guy D. Eslick*

The Whiteley-Martin Research Centre, Discipline of Surgery, The University of Sydney, Nepean Hospital, Level 3, Clinical Building, PO Box 63, Penrith 2751, NSW, Australia

A B S T R A C T

There has been enormous debate regarding the possibility of a link between childhood vaccinations and the subsequent development of autism. This has in recent times become a major public health issue with vaccine preventable diseases increasing in the community due to the fear of a 'link' between vaccinations and autism. We performed a meta-analysis to summarise available evidence from case-control and cohort studies on this topic (MEDLINE, PubMed, EMBASE, Google Scholar up to April, 2014). Eligible studies assessed the relationship between vaccine administration and the subsequent development of autism or autism spectrum disorders (ASD). Two reviewers extracted data on study characteristics, methods, and outcomes. Disagreement was resolved by consensus with another author. Five cohort studies involving 1,256,407 children, and five case-control studies involving 9,920 children were included in this analysis. The cohort data revealed no relationship between vaccination and autism (OR: 0.99; 95% CI: 0.92 to 1.06) or ASD (OR: 0.91; 95% CI: 0.68 to 1.20), nor was there a relationship between autism and MMR (OR: 0.84; 95% CI: 0.70 to 1.01), or thimerosal (OR: 1.00; 95% CI: 0.77 to 1.31), or mercury (Hg) (OR: 1.00; 95% CI: 0.93 to 1.07). Similarly the case-control data found no evidence for increased risk of developing autism or ASD following MMR, Hg, or thimerosal exposure when grouped by condition (OR: 0.90, 95% CI: 0.83 to 0.98; $p = 0.02$) or grouped by exposure type (OR: 0.85, 95% CI: 0.76 to 0.95; $p = 0.01$). Findings of this meta-analysis suggest that vaccinations are not associated with the development of autism or autism spectrum disorder. Furthermore, the components of the vaccines (thimerosal or mercury) or multiple vaccines (MMR) are not associated with the development of autism or autism spectrum disorder.

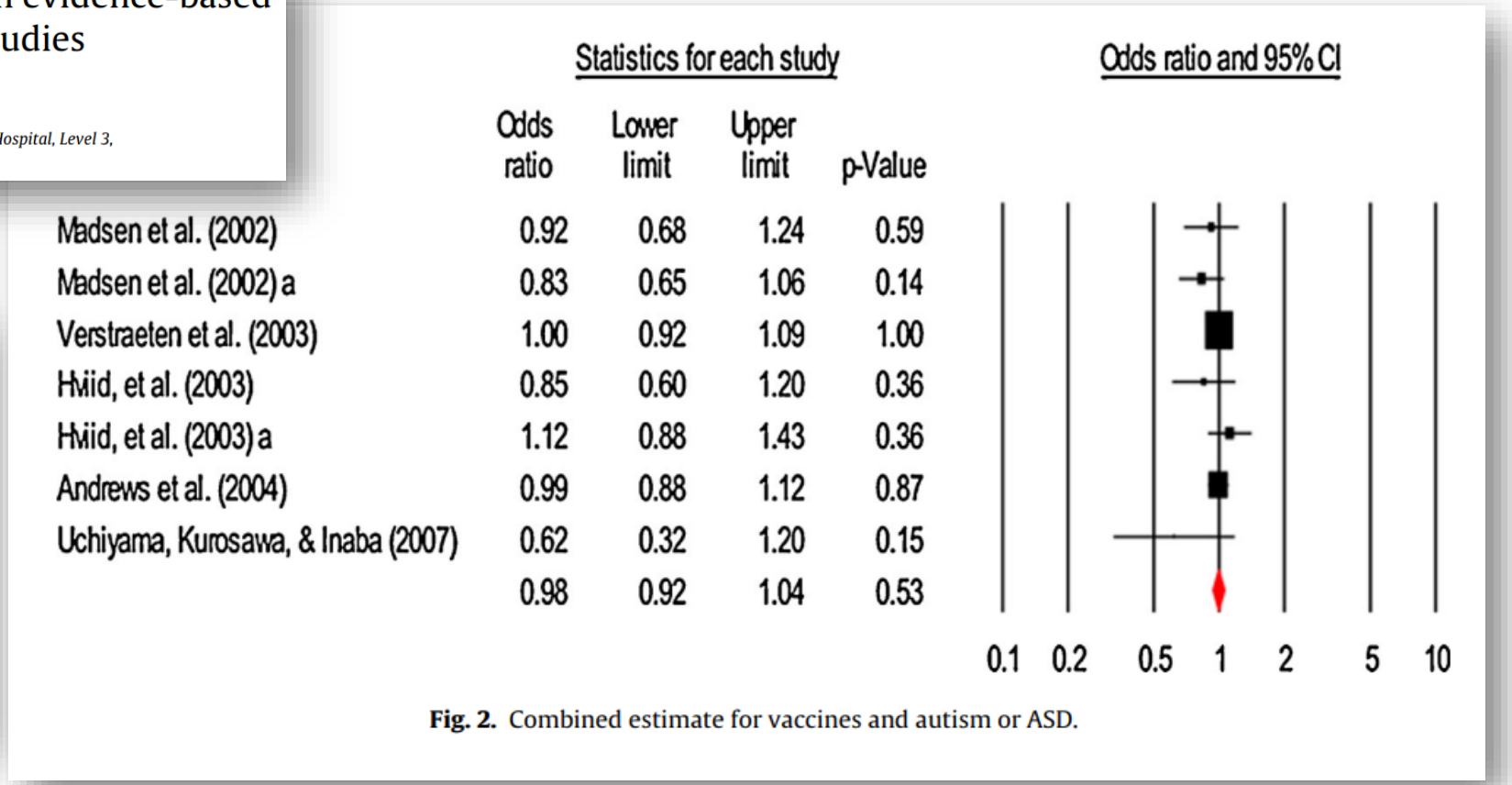


Fig. 2. Combined estimate for vaccines and autism or ASD.

Taylor, L.E., Swerdfeger, A.L., & Eslick, G.D. (2014). Vaccines are not associated with autism: an evidence-based meta-analysis of case-control and cohort studies. *Vaccine 32*(29), 3623-9. doi: 10.1016/j.vaccine.2014.04.085.

Meta-analýza: Příklad

Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and coho

Luke E. Taylor, Amy L. Swerdfeger, Guy D. Eslick*

The Whiteley-Martin Research Centre, Discipline of Surgery, The University of Sydney, Clinical Building, PO Box 63, Penrith 2751, NSW, Australia

ABSTRACT

There has been enormous debate regarding the possibility of a link between childhood vaccination the subsequent development of autism. This has in recent times become a major public health issue vaccine preventable diseases increasing in the community due to the fear of a 'link' between vaccina and autism. We performed a meta-analysis to summarise available evidence from case-control and c studies on this topic (MEDLINE, PubMed, EMBASE, Google Scholar up to April, 2014). Eligible st assessed the relationship between vaccine administration and the subsequent development of autis autism spectrum disorders (ASD). Two reviewers extracted data on study characteristics, methods outcomes. Disagreement was resolved by consensus with another author. Five cohort studies invo 1,256,407 children, and five case-control studies involving 9,920 children were included in this ana The cohort data revealed no relationship between vaccination and autism (OR: 0.99; 95% CI: 0.92 to or ASD (OR: 0.91; 95% CI: 0.68 to 1.20), nor was there a relationship between autism and MMR (OR: 95% CI: 0.70 to 1.01), or thimerosal (OR: 1.00; 95% CI: 0.77 to 1.31), or mercury (Hg) (OR: 1.00; 95 0.93 to 1.07). Similarly the case-control data found no evidence for increased risk of developing au or ASD following MMR, Hg, or thimerosal exposure when grouped by condition (OR: 0.90, 95% CI: to 0.98; $p = 0.02$) or grouped by exposure type (OR: 0.85, 95% CI: 0.76 to 0.95; $p = 0.01$). Findings o meta-analysis suggest that vaccinations are not associated with the development of autism or au spectrum disorder. Furthermore, the components of the vaccines (thimerosal or mercury) or mu vaccines (MMR) are not associated with the development of autism or autism spectrum disorder.

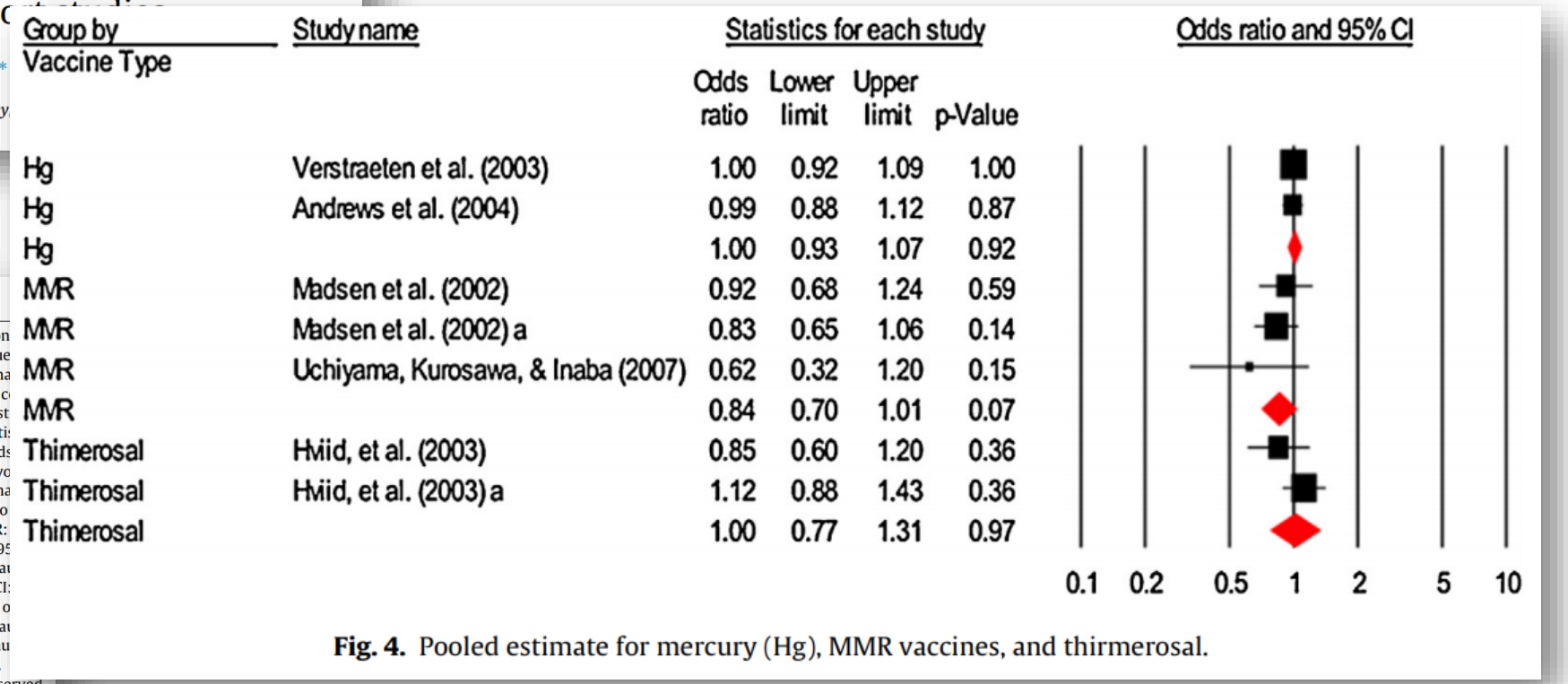


Fig. 4. Pooled estimate for mercury (Hg), MMR vaccines, and thirmerosal.

Taylor, L.E., Swerdfeger, A.L., & Eslick, G.D. (2014). Vaccines are not associated with autism: an evidence-based meta-analysis of case-control and cohort studies. *Vaccine 32*(29), 3623-9. doi: 10.1016/j.vaccine.2014.04.085.

Meta-analýza: Příklad

Někdy se dílčí detaily grafu liší

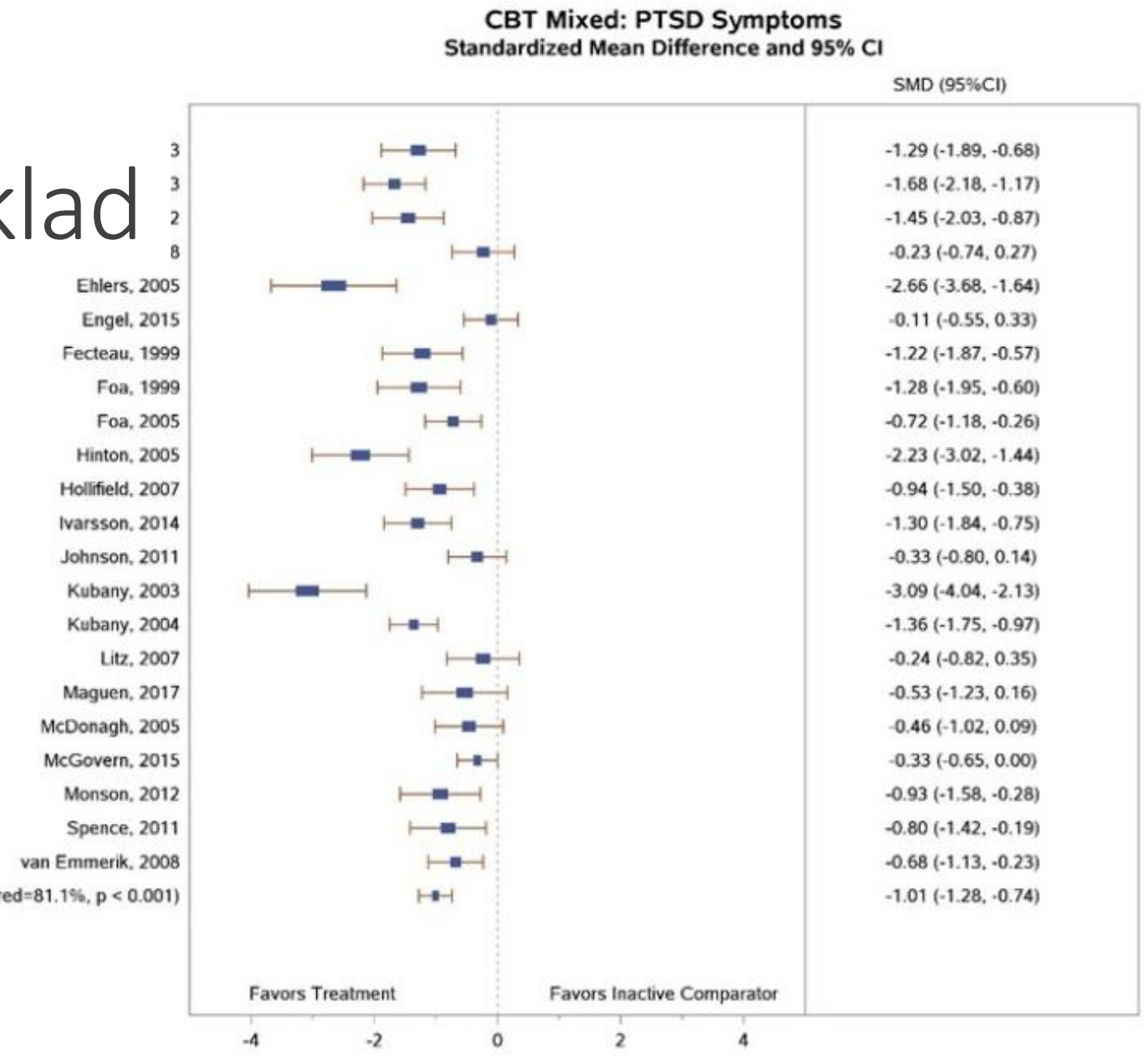
- Zde je souhrnný efekt znázorněný boxplotem a ne „diamantem“, není zdůrazněna velikost vzorků.

Někdy je graf doplněn o další informace.

- Zde např. heterogenita (I^2), viz dále.
- Jindy rozdělení efektů do skupin apod.

Je potřeba věnovat pozornost tomu, jaký byl použit ukazatel velikosti účinku.

- Zde standardizovaný rozdíl průměru (tedy Cohenovo d); žádný efekt → $d_0 = 0$.
- Na předchozích grafech šlo o poměr šancí (OR); žádný efekt → $OR_0 = 1$.



Forman-Hoffman V, Middleton JC, Feltner C, et al. Psychological and Pharmacological Treatments for Adults With Posttraumatic Stress Disorder: A Systematic Review Update [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018 May. (Comparative Effectiveness Review, No. 207.) Evidence Summary. Available from: https://www.ncbi.nlm.nih.gov/books/NBK525135/

Meta-analýza: potíže a řešení

Lze jedinou oblast výzkumnou oblast zastoupit jedním číslem?

- Zkoumáme jeden (fixed) efekt nebo populaci (random) efektů?

Zdrojové studie.

- Zkreslené původní studie, vynechání důležitých studií. Garbage in, garbage out.
- Srovnávání nesrovnatelného?
- Rozdílné velikost efektů a interpretace testů.

Úroveň realizovaných meta-analýz.

- Nedostatečná kontrola kvality původních studií a korekce na publikační zkreslení.

Analytická vs. explorační meta-analýza.

Meta-analýza je kvalitní do té míry, do jaké jsou kvalitní individuální studie.

Meta-analýza: potíže a řešení

Příklad A: ***Znáte*** skutečnou velikost efektu, d = 0,3.
Realizujete dvě studie, N_1 = 50 a N_2 = 500.

- Otázka A1: Ve které studii budete pravděpodobněji pozorovat statisticky významný efekt?
- Otázka A2: Ve které studii budete pravděpodobněji pozorovat větší velikost účinku?

Příklad B: ***Neznáte*** skutečnou velikost efektu.
Realizujete dvě studie, N_1 = 50 a N_2 = 500.

- Otázka B1: Ve které studii budete pravděpodobněji pozorovat statisticky významný efekt?
- Otázka B2: Ve které studii budete pravděpodobněji pozorovat větší velikost účinku?

Příklad C: ***Neznáte*** skutečnou velikost efektu.
V databázi naleznete dvě publikované studie, N_1 = 50 a N_2 = 500.

- Otázka C1: Ve které studii budete pravděpodobněji pozorovat statisticky významný efekt?
- Otázka C2: Ve které studii budete pravděpodobněji pozorovat větší velikost účinku?

Meta-analýza: Funnel-plot

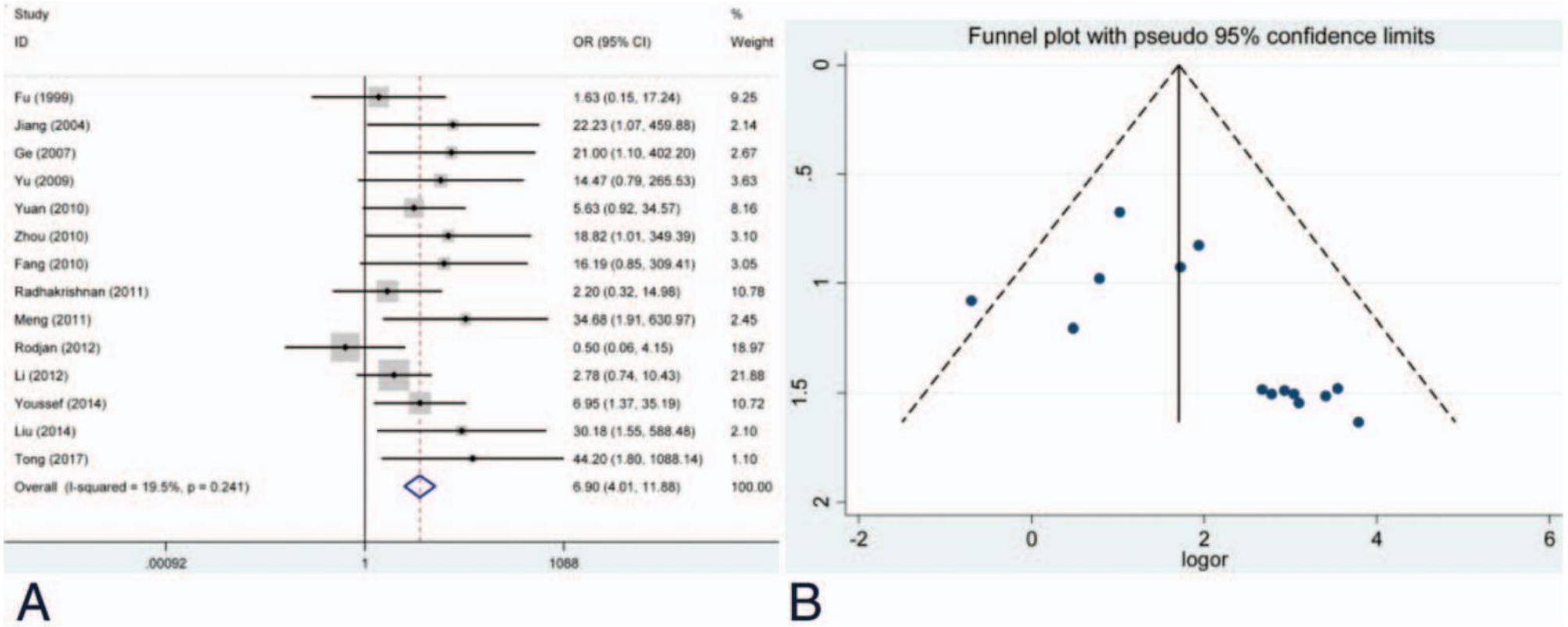


Figure 4. Forest plot and funnel plot of association between vascular endothelial growth factor protein expression and optic nerve involvement of retinoblastoma. (A) Forest plot and (B) funnel plot. CI = confidence interval, OR = odds ratio.

Meta-analýza: Funnel-plot

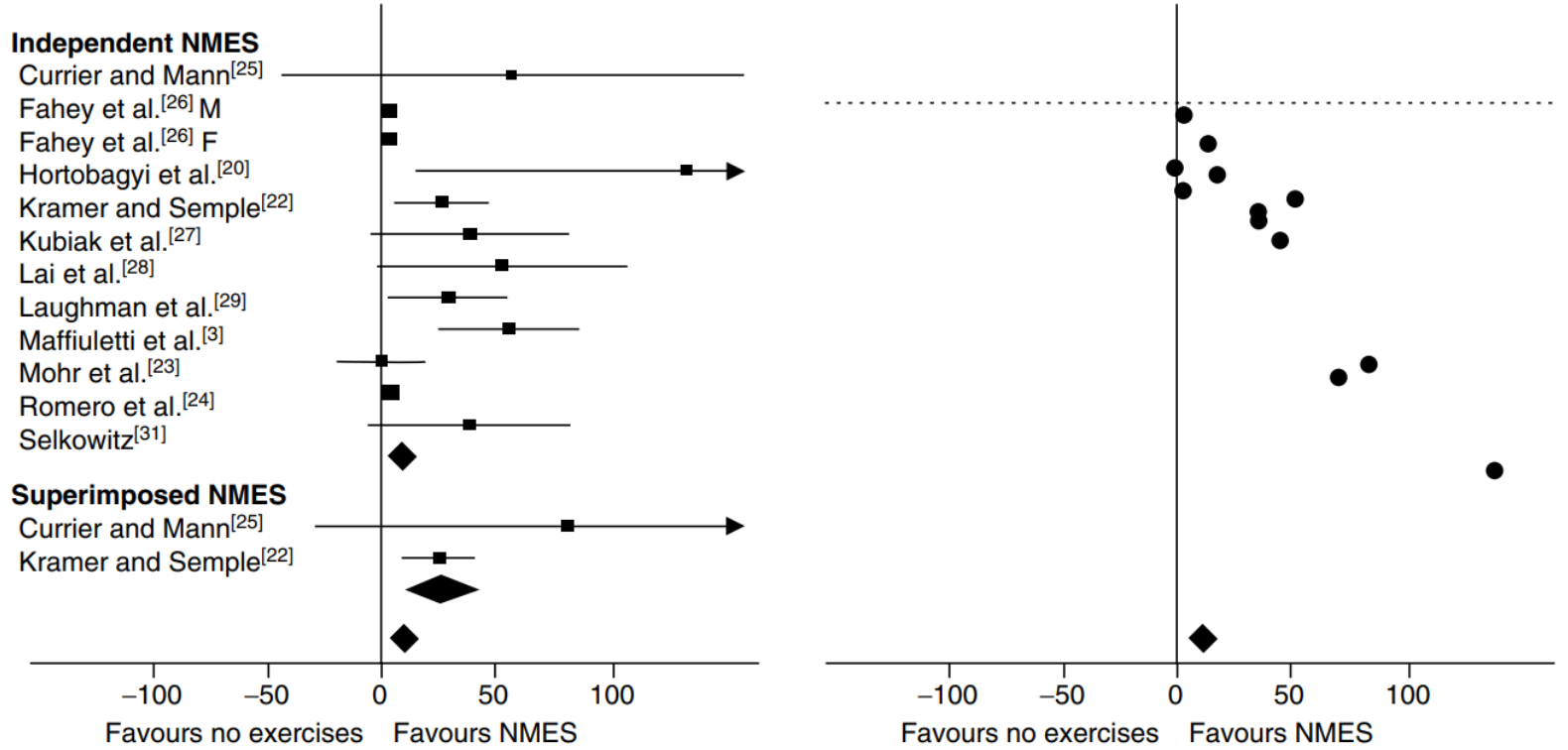
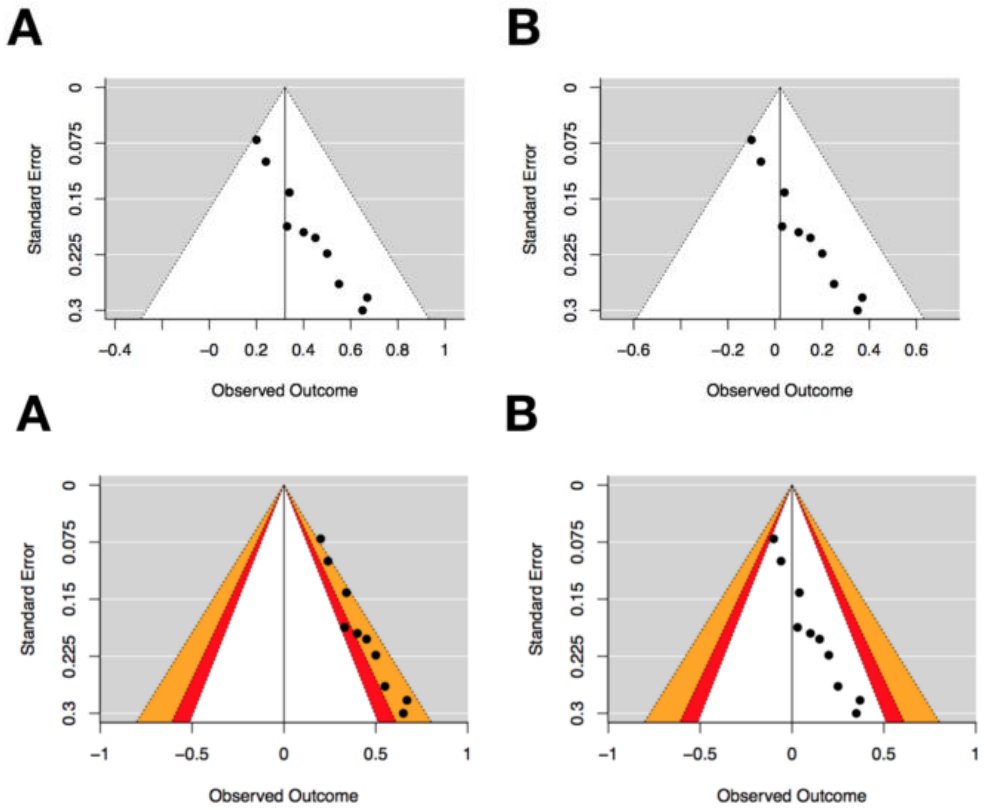
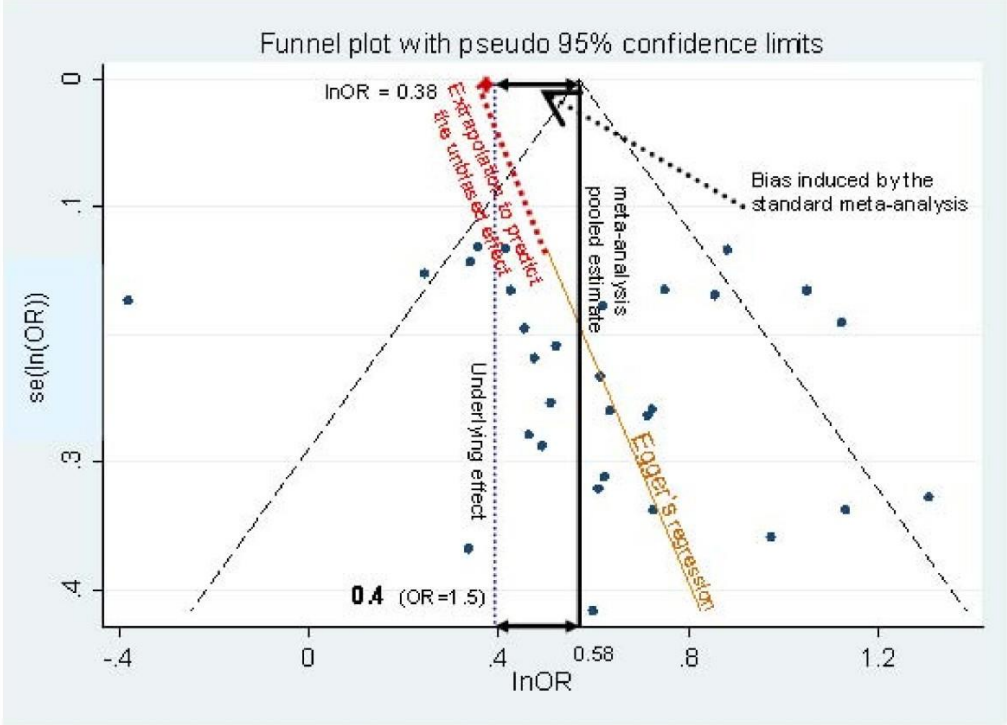


Fig. 3. Forest plot and funnel plot of neuromuscular electrical stimulation (NMES) versus no exercises – healthy quadriceps. The squares and circles represent the mean outcome of each study and the corresponding horizontal lines are 95% confidence intervals. The diamonds represent the pooled (subgroup) outcomes with the horizontal width corresponding to the outcome's 95% confidence interval.[3,20,22-29,31] **F** = females; **M** = males.

Meta-analýza: Funnel-plot



https://towardsdatascience.com/constructing-contour-enhanced-funnel-plots-for-meta-analysis-6434cc8e51d0



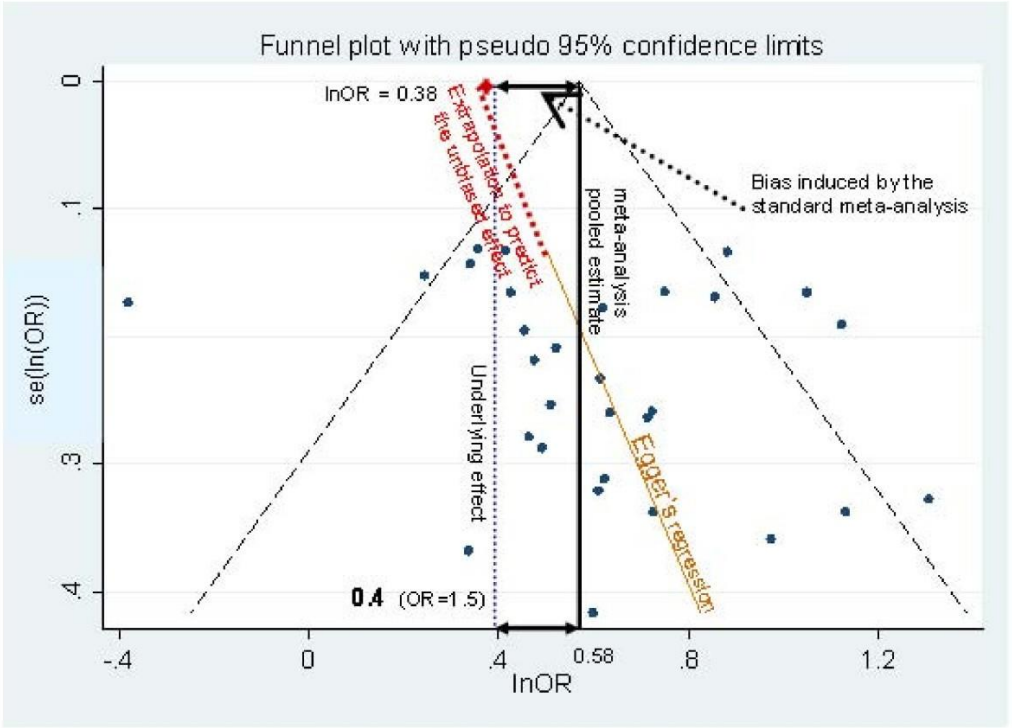
Moreno, S.G., Sutton, A.J., Ades, A., et al.(2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology 9*(2). https://doi.org/10.1186/1471-2288-9-2

Meta-analýza: Eggerův test

Existuje souvislost mezi pozorovanou velikostí účinku a standardní chybou jejího odhadu napříč studiemi?

- Signifikantní výsledek: podklad pro existenci publikačního zkreslení.
- Technicky jde o obyčejný Waldův z-test o signifikanci regresního koeficientu pojmenovaný po Eggerovi (1997), který toto použití navrhnul.
- Analogicky je někdy používaná Kendallova korelace velikosti vzorku a velikost efektu.

Eggerův test posloužil jako podklad pro tzv. *„bias corrected effect size estimates“.*



Moreno, S.G., Sutton, A.J., Ades, A., et al.(2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology 9*(2). https://doi.org/10.1186/1471-2288-9-2

Zdrojové studie?

Silná preference statisticky signifikantních výsledků.

- 92 % publikovaných výsledků v psychologii je statisticky signifikantních (Fanelli, 2010)
- Nárůst zejména v období mezi lety 1990 a 2007 (Fanelli, 2012).

→ Konfirmační zkreslení (confirmation bias in publication).

Bakker, Van Dijk, & Wicherts (2012): 13 meta-analýz s 281 studiemi.

- Medián $N = 40$; Statistická síla $1\text{-}\beta = 0{,}35$; $d = 0{,}5$.

Fraley & Marks (2007): Meta-analýza korelačních studií osobnosti

- Medián: $N = 120$, statistická síla $1\text{-}\beta = 0{,}65$, $r = 0{,}21$.

„Consequently, if all effects reported in published studies were true, only 35% would be replicable in similarly underpowered studies.“ (Asendorpf et al. 2013, s. 110)

Nic nového pod sluncem...

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145–153. doi:10.1037/h0045186

- Odhad replikovatelnosti: Statistická síla 50 %.
- Doporučení: Zvýšit sílu na 80 %.

A další...

Journal of Abnormal and Social Psychology
1962, Vol. 65, No. 3, 145–153

THE STATISTICAL POWER OF ABNORMAL–SOCIAL PSYCHOLOGICAL RESEARCH:

A REVIEW [1]

JACOB COHEN

New York University

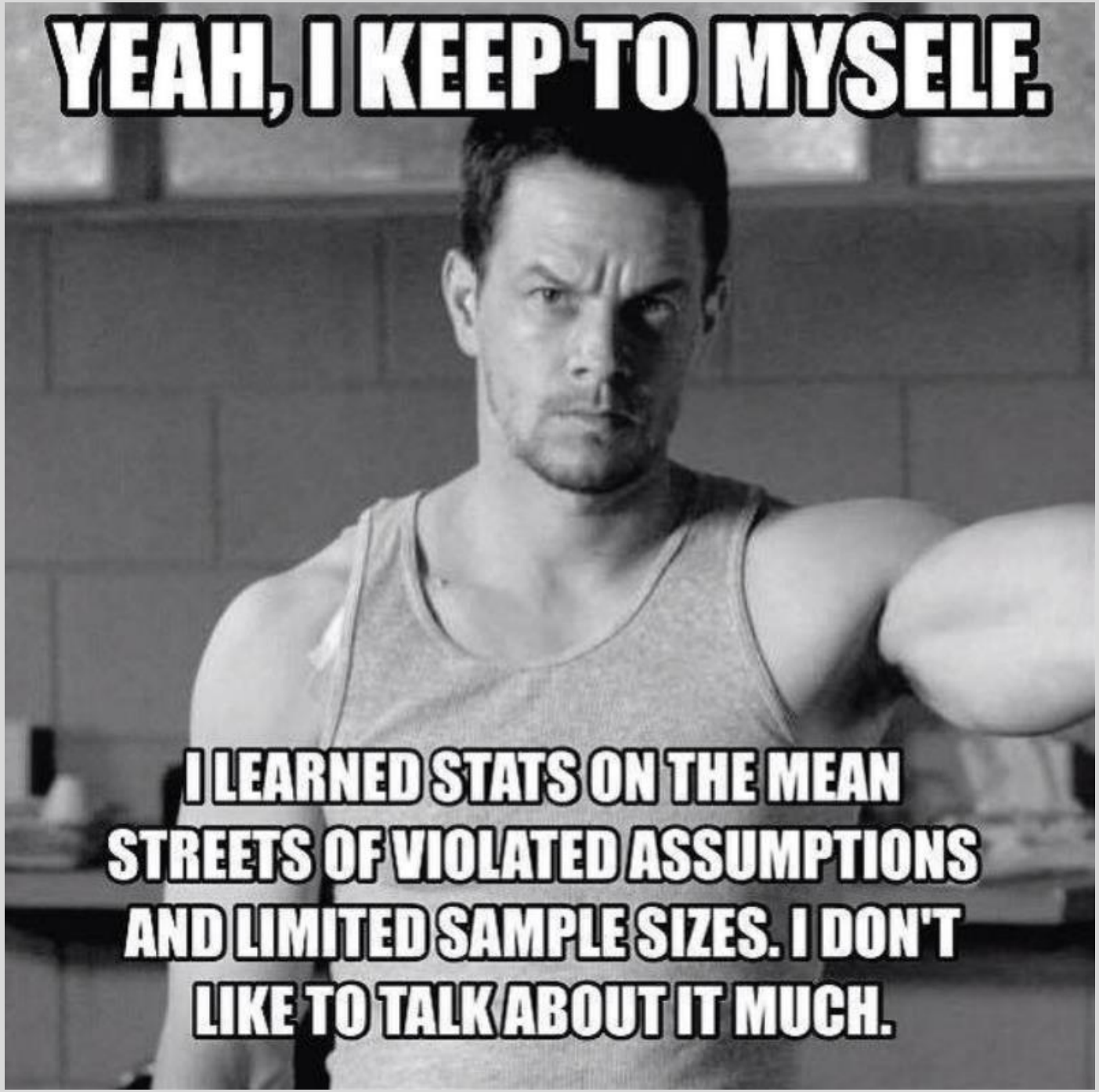
Given an experimental effect in a population, how likely is the null hypothesis to be rejected? Equivalently, what is the power of the statistical test? What is the expectation that the (false) null hypothesis will be sustained and thus a Type II error committed?

It is a remarkable phenomenon that the research which is reported by psychological investigators rarely refers to this issue, and even more rarely actually investigates it. On the other hand, issues concerning Type I error or "significance," i.e., the validity of the *rejection* of the null hypothesis, are more or less conscientiously attended to. This marked asymmetry of sophistication and attention to these two types of error is mirrored, and largely determined, by the exposition of these issues in the statistics textbooks used in the graduate training of the investigators. These texts are characterized by an early explanation of Type I and Type II errors, followed by a neglect of the latter throughout the remainder of the text. Thus, every statistical test is described with careful attention to

doctoral candidate and sponsor, or author and editor) and rarely on the basis of a Type II error analysis, which can always be performed *prior* to the collection of data. These nonrational bases for setting sample size must often result in investigations being undertaken which have little chance of success despite the actual falsity of the null hypothesis, and probably less often in the use of a far larger sample than is necessary. Either of these circumstances is wasteful of research effort.

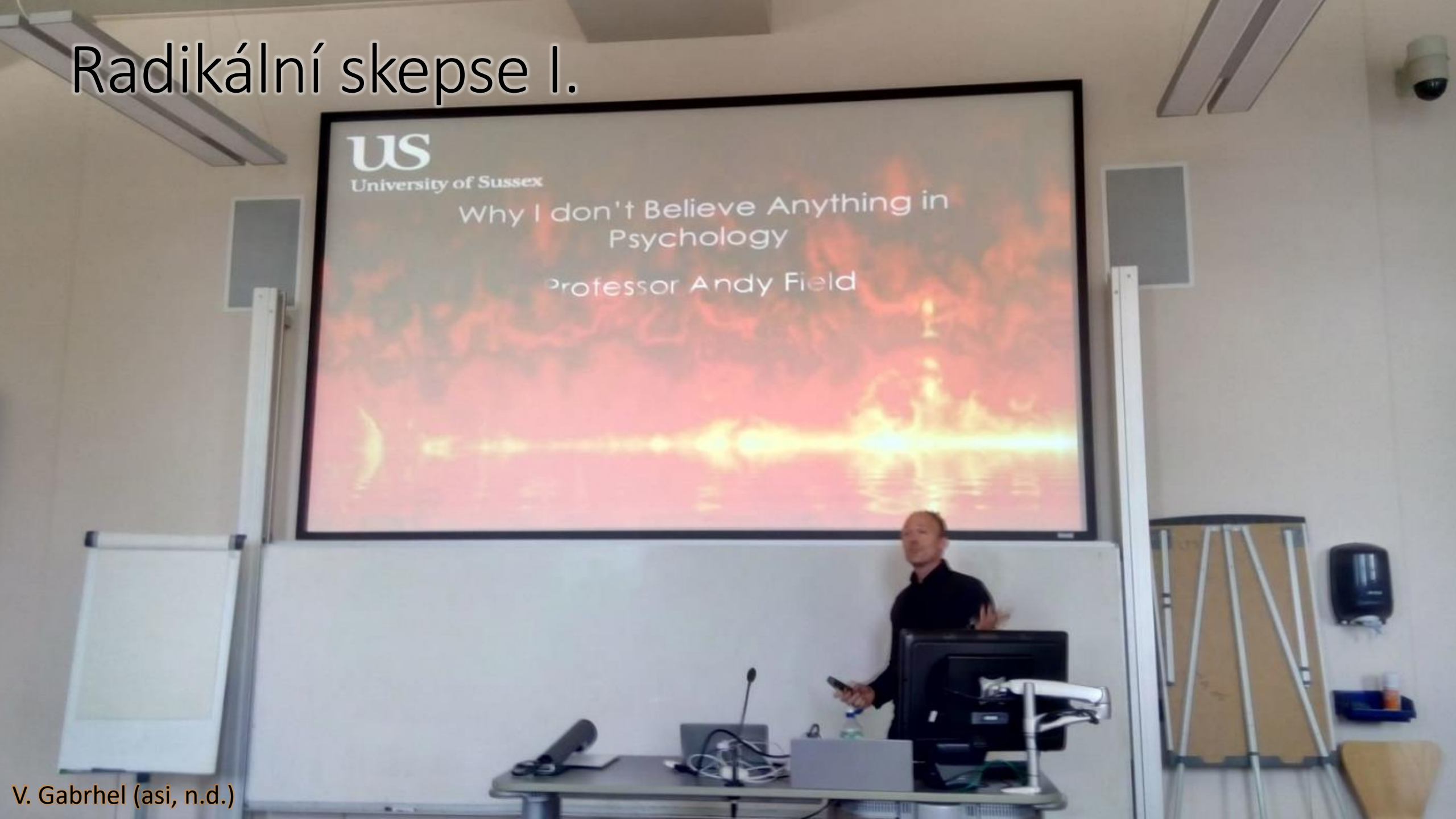
Stemming from these considerations, a program of investigation, computation, and reportage has been undertaken whose major aims are as follows:

1. To call these issues to the attention of investigators, consumers of research, and evaluators of research planned or completed (sponsors, agency panels, journal editors).
2. To provide tables and conventional standards which will facilitate the performance of power analyses for the most common statistical tests.
3. To conduct surveys of the psychological


YEAH, I KEEP TO MYSELF.
I LEARNED STATS ON THE MEAN STREETS OF VIOLATED ASSUMPTIONS AND LIMITED SAMPLE SIZES. I DON'T LIKE TO TALK ABOUT IT MUCH.

Replikovatelnost (psychologického) výzkumu

Radikální skepse I.



V. Gabrhel (asi, n.d.)

Začátek „krize“: 2011–2012

Daryl Bem:
Feeling the Future (2011)

Diederik Stapel
(58 retrakcí 2011–2019)

John Bargh
priming stářím (5.000 citací)

Radikální skepse II:
Estimating the reproducibility of psychological science

„We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.“

100 studií a výsledky jejich replikace

- Psychological Science
- Journal of Personality and Social Psychology
- Journal of Experimental Psychology: Learning, Memory, and Cognition

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Alexander A. Aarts, Joanna E. Anderson, Christopher J. Anderson, Peter R. Attridge, Angela Attwood, Jordan Axt, Molly Babel, Štěpán Bahník, Erica Baranski, Michael Barnett-Cowan,Elizabeth Bartmess, Jennifer Beer, Raoul Bell, Heather Bentley, Leah Beyan, Grace Binion, Denny Borsboom, Annick Bosch, Frank A. Bosco, Sara D. Bowman, Mark J. Brandt, Erin Braswell, Hilmar Brohmer, Benjamin T. Brown, Kristina Brown, Jovita Brüning, Ann Calhoun-Sauls, Shannon P. Callahan, Elizabeth Chagnon, Jesse Chandler, Christopher R. Chartier, Felix Cheung, Cody D. Christopherson, Linda Cillessen, Russ Clay, Hayley Cleary, Mark D. Cloud, Michael Cohn, Johanna Cohoon,Simon Columbus, Andreas Cordes, Giulio Costantini, Leslie D. Cramblet Alvarez, Ed Cremata, Jan Crusius, Jamie DeCoster, Michelle A. DeGaetano, Nicolás Della Penna, Bobby den Bezemer, Marie K. Deserno, Olivia Devitt, Laura Dewitte, David G. Dobolyi, Geneva T. Dodson, M. Brent Donnellan, Ryan Donohue, Rebecca A. Dore, Angela Dorrough, Anna Dreber, Michelle Dugas, Elizabeth W. Dunn, Kayleigh Easey, Sylvia Eboigbe, Casey Eggleston, Jo Embley, Sacha Epskamp, Timothy M. Errington, Vivien Estel, Frank J. Farach, Jenelle Feather, Anna Fedor, Belén Fernández-Castilla, Susann Fiedler, James G. Field, Stanka A. Fitneva, Taru Flagan, Amanda L. Forest, Eskil Forsell, Joshua D. Foster, Michael C. Frank, Rebecca S. Frazier, Heather Fuchs, Philip Gable, Jeff Galak,Elisa Maria Galliani, Anup Gampa, Sara Garcia, Douglas Gazarian, Elizabeth Gilbert, Roger Giner-Sorolla, Andreas Glöckner, Lars Goellner, Jin X. Goh, Rebecca Goldberg, Patrick T. Goodbourn, Shauna Gordon-McKeon, Bryan Gorges, Jessie Gorges, Justin Goss, Jesse Graham, James A. Grange, Jeremy Gray, Chris Hartgerink, Joshua Hartshorne, Fred Hasselman, Timothy Hayes, Emma Heikensten, Felix Henninger, John Hodsoll,Taylor Holubar, Gea Hoogendoorn, Denise J. Humphries, Cathy O.-Y. Hung, Nathali Immelman, Vanessa C. Irsik, Georg Jahn, Frank Jäkel, Marc Jekel, Magnus Johannesson, Larissa G. Johnson, David J. Johnson, Kate M. Johnson, William J. Johnston, Kai Jonas, Jennifer A. Joy-Gaba, Heather Barry Kappes, Kim Kelso, Mallory C. Kidwell, Seung Kyung Kim, Matthew Kirkhart, Bennett Kleinberg, Goran Knežević,Franziska Maria Kolorz, Jolanda J. Kossakowski, Robert Wilhelm Krause, Job Krijnen, Tim Kuhlmann, Yoram K. Kunkels, Megan M. Kyc, Calvin K. Lai, Aamir Laique, Daniël Lakens,Kristin A. Lane, Bethany Lassetter, Ljiljana B. Lazarević, Etienne P. LeBel, Key Jung Lee,Minha Lee, Kristi Lemm, Carmel A. Levitan, Melissa Lewis, Lin Lin, Stephanie Lin,Matthias Lippold, Darren Loureiro, Ilse Luteijn, Sean Mackinnon, Heather N. Mainard,Denise C. Marigold, Daniel P. Martin, Tylar Martinez, E.J. Masicampo, Josh Matacotta,Maya Mathur, Michael May, Nicole Mechin, Pranjal Mehta, Johannes Meixner, Alissa Melinger, Jeremy K. Miller, Mallorie Miller, Katherine Moore, Marcus Möschl, Matt Motyl, Stephanie M. Müller, Marcus Munafo, Koen I. Neijenhuijs, Taylor Nervi, Gandalf Nicolas, Gustav Nilsonne, Brian A. Nosek, Michèle B. Nuijten, Catherine Olsson,Colleen Osborne, Lutz Ostkamp, Misha Pavel, Ian S. Penton-Voak, Olivia Perna, Cyril Pernet, Marco Perugini, R. Nathan Pipitone, Michael Pitts, Franziska Plessow, Jason M. Prenoveau, Rima-Maria Rahal, Kate A. Ratliff, David Reinhard, Frank Renkewitz,Ashley A. Ricker, Anastasia Rigney, Andrew M. Rivers, Mark Roebke, Abraham M. Rutchick, Robert S. Ryan, Onur Sahin, Anondah Saide, Gillian M. Sandstrom, David Santos, Rebecca Saxe, René Schlegelmilch, Kathleen Schmidt, Sabine Scholz,Larissa Seibel, Dylan Faulkner Selterman, Samuel Shaki, William B. Simpson, H. Colleen Sinclair, Jeanine L. M. Skorinko, Agnieszka Slowik, Joel S. Snyder, Courtney Soderberg,Carina Sonnleitner, Nick Spencer, Jeffrey R. Spies, Sara Steegen, Stefan Stieger, Nina Strohminger, Gavin B. Sullivan, Thomas Talhelm, Megan Tapia, Anniek te Dorsthorst,Manuela Thomae, Sarah L. Thomas, Pia Tio, Frits Traets, Steve Tsang, Francis Tuerlinckx, Paul Turchan, Milan Valášek, Anna E. van 't Veer, Robbie Van Aert, Marcel van Assen, Riet van Bork, Mathijs van de Ven, Don van den Bergh, Marije van der Hulst,Roel van Dooren, Johnny van Doorn, Daan R. van Renswoude, Hedderik van Rijn, Wolf Vanpaemel, Alejandro Vásquez Echeverría, Melissa Vazquez, Natalia Velez, Marieke Vermue, Mark Verschoor, Michelangelo Vianello, Martin Voracek, Gina Vuu, Eric-Jan Wagenmakers, Joanneke Weerdmeester, Ashlee Welsh, Erin C. Westgate, Joeri Wissink,Michael Wood, Andy Woods, Emily Wright, Sining Wu, Marcel Zeelenberg, Kellylynn Zuni

Radikální skepse II:
Estimating the reproducibility of psychological science

„We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.“

100 studií a výsledky jejich replikace

- Psychological Science
- Journal of Personality and Social Psychology
- Journal of Experimental Psychology: Learning, Memory, and Cognition

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Alexander A. Aarts, Joanna E. Anderson, Christopher J. Anderson, Peter R. Attridge, Angela Attwood, Jordan Axt, Molly Babel, **Štěpán Bahník**, Erica Baranski, Michael Barnett-Cowan,Elizabeth Bartmess, Jennifer Beer, Raoul Bell, Heather Bentley, Leah Beyan, Grace Binion, Denny Borsboom, Annick Bosch, Frank A. Bosco, Sara D. Bowman, Mark J. Brandt, Erin Braswell, Hilmar Brohmer, Benjamin T. Brown, Kristina Brown, Jovita Brüning, Ann Calhoun-Sauls, Shannon P. Callahan, Elizabeth Chagnon, Jesse Chandler, Christopher R. Chartier, Felix Cheung, Cody D. Christopherson, Linda Cillessen, Russ Clay, Hayley Cleary, Mark D. Cloud, Michael Cohn, Johanna Cohoon,Simon Columbus, Andreas Cordes, Giulio Costantini, Leslie D. Cramblet Alvarez, Ed Cremata, Jan Crusius, Jamie DeCoster, Michelle A. DeGaetano, Nicolás Della Penna, Bobby den Bezemer, Marie K. Deserno, Olivia Devitt, Laura Dewitte, David G. Dobolyi, Geneva T. Dodson, M. Brent Donnellan, Ryan Donohue, Rebecca A. Dore, Angela Dorrough, Anna Dreber, Michelle Dugas, Elizabeth W. Dunn, Kayleigh Easey, Sylvia Eboigbe, Casey Eggleston, Jo Embley, Sacha Epskamp, Timothy M. Errington, Vivien Estel, Frank J. Farach, Jenelle Feather, Anna Fedor, Belén Fernández-Castilla, Susann Fiedler, James G. Field, Stanka A. Fitneva, Taru Flagan, Amanda L. Forest, Eskil Forsell, Joshua D. Foster, Michael C. Frank, Rebecca S. Frazier, Heather Fuchs, Philip Gable, Jeff Galak,Elisa Maria Galliani, Anup Gampa, Sara Garcia, Douglas Gazarian, Elizabeth Gilbert, Roger Giner-Sorolla, Andreas Glöckner, Lars Goellner, Jin X. Goh, Rebecca Goldberg, Patrick T. Goodbourn, Shauna Gordon-McKeon, Bryan Gorges, Jessie Gorges, Justin Goss, Jesse Graham, James A. Grange, Jeremy Gray, Chris Hartgerink, Joshua Hartshorne, Fred Hasselman, Timothy Hayes, Emma Heikensten, Felix Henninger, John Hodsoll,Taylor Holubar, Gea Hoogendoorn, Denise J. Humphries, Cathy O.-Y. Hung, Nathali Immelman, Vanessa C. Irsik, Georg Jahn, Frank Jäkel, Marc Jekel, Magnus Johannesson, Larissa G. Johnson, David J. Johnson, Kate M. Johnson, William J. Johnston, Kai Jonas, Jennifer A. Joy-Gaba, Heather Barry Kappes, Kim Kelso, Mallory C. Kidwell, Seung Kyung Kim, Matthew Kirkhart, Bennett Kleinberg, Goran Knežević,Franziska Maria Kolorz, Jolanda J. Kossakowski, Robert Wilhelm Krause, Job Krijnen, Tim Kuhlmann, Yoram K. Kunkels, Megan M. Kyc, Calvin K. Lai, Aamir Laique, Daniël Lakens,Kristin A. Lane, Bethany Lassetter, Ljiljana B. Lazarević, Etienne P. LeBel, Key Jung Lee,Minha Lee, Kristi Lemm, Carmel A. Levitan, Melissa Lewis, Lin Lin, Stephanie Lin,Matthias Lippold, Darren Loureiro, Ilse Luteijn, Sean Mackinnon, Heather N. Mainard,Denise C. Marigold, Daniel P. Martin, Tylar Martinez, E.J. Masicampo, Josh Matacotta,Maya Mathur, Michael May, Nicole Mechin, Pranjal Mehta, Johannes Meixner, Alissa Melinger, Jeremy K. Miller, Mallorie Miller, Katherine Moore, Marcus Möschl, Matt Motyl, Stephanie M. Müller, Marcus Munafo, Koen I. Neijenhuijs, Taylor Nervi, Gandalf Nicolas, Gustav Nilsonne, Brian A. Nosek, Michèle B. Nuijten, Catherine Olsson,Colleen Osborne, Lutz Ostkamp, Misha Pavel, Ian S. Penton-Voak, Olivia Perna, Cyril Pernet, Marco Perugini, R. Nathan Pipitone, Michael Pitts, Franziska Plessow, Jason M. Prenoveau, Rima-Maria Rahal, Kate A. Ratliff, David Reinhard, Frank Renkewitz,Ashley A. Ricker, Anastasia Rigney, Andrew M. Rivers, Mark Roebke, Abraham M. Rutchick, Robert S. Ryan, Onur Sahin, Anondah Saide, Gillian M. Sandstrom, David Santos, Rebecca Saxe, René Schlegelmilch, Kathleen Schmidt, Sabine Scholz,Larissa Seibel, Dylan Faulkner Selterman, Samuel Shaki, William B. Simpson, H. Colleen Sinclair, Jeanine L. M. Skorinko, Agnieszka Slowik, Joel S. Snyder, Courtney Soderberg,Carina Sonnleitner, Nick Spencer, Jeffrey R. Spies, Sara Steegen, Stefan Stieger, Nina Strohminger, Gavin B. Sullivan, Thomas Talhelm, Megan Tapia, Anniek te Dorsthorst,Manuela Thomae, Sarah L. Thomas, Pia Tio, Frits Traets, Steve Tsang, Francis Tuerlinckx, Paul Turchan, Milan Valášek, Anna E. van 't Veer, Robbie Van Aert, Marcel van Assen, Riet van Bork, Mathijs van de Ven, Don van den Bergh, Marije van der Hulst,Roel van Dooren, Johnny van Doorn, Daan R. van Renswoude, Hedderik van Rijn, Wolf Vanpaemel, Alejandro Vásquez Echeverría, Melissa Vazquez, Natalia Velez, Marieke Vermue, Mark Verschoor, Michelangelo Vianello, Martin Voracek, Gina Vuu, Eric-Jan Wagenmakers, Joanneke Weerdmeester, Ashlee Welsh, Erin C. Westgate, Joeri Wissink,Michael Wood, Andy Woods, Emily Wright, Sining Wu, Marcel Zeelenberg, Kellylynn Zuni

Radikální skepse II:
Estimating the reproducibility of psychological science

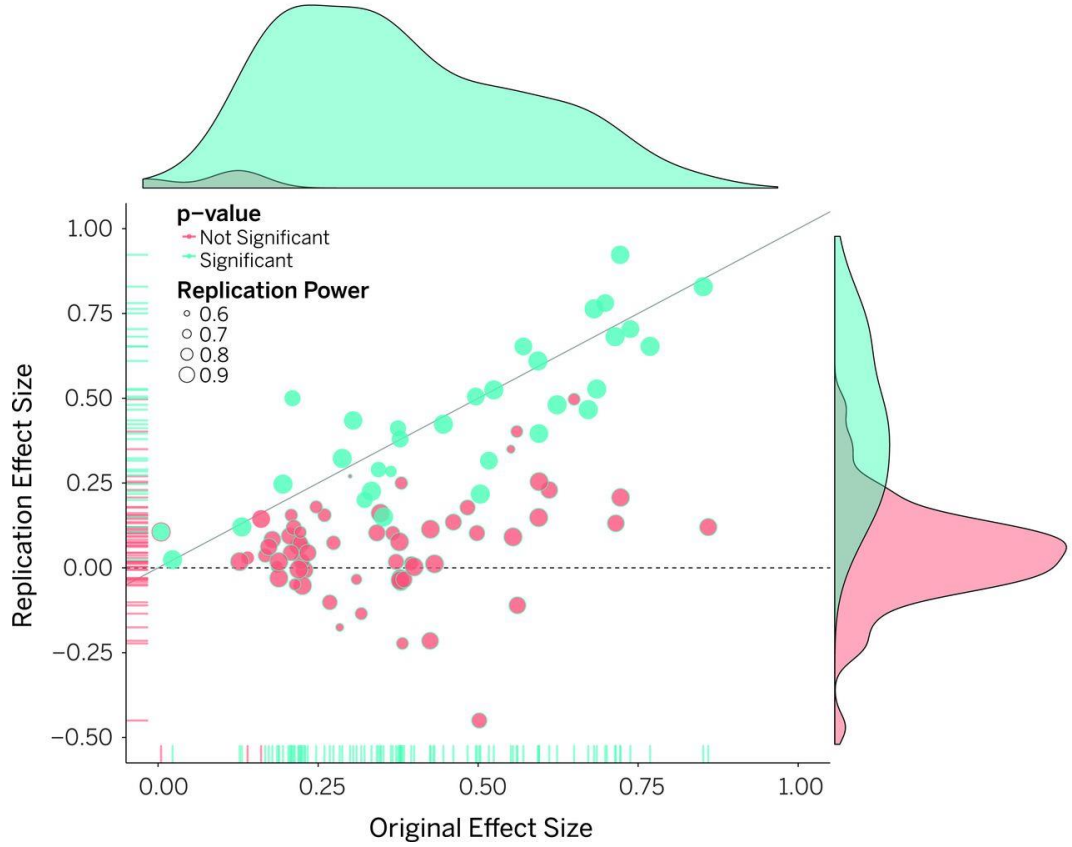
Původní velikost efektů:

- Průměrná velikost účinku
  M_r = 0,403; *SD* = 0,188
- Statistická signifikance: 97 % studií $p < 0,05$

Replikovaná velikost efektů:

- Průměrná velikost účinku
  M_r = 0,197; *SD* = 0,257
- Statistická signifikance: 36 % studií $p < 0,05$

Hodnota velikostí účinku z původních studií se nacházela v 95% intervalu spolehlivosti při replikaci v 47 % případů.



Pochybné praktiky ve výzkumu

„In a poll of more than 2000 psychologists, prevalences of 'Deciding whether to collect more data after looking to see whether the results were significant' and 'Stopping data collection earlier than planned because one found the result that one had been looking for' were subjectively estimated at 61% and 39%, respectively."

- John, Loewenstein, & Prelec, cit. dle Asendorpf et al., 2013

Questionable research practices.

Podvodné vs. pochybné jednání?

- *„Fraud is typically limited to cases in which researchers create false data."*
- *„In contrast, QRPs typically involve the exclusion of data that are inconsistent with a theoretical hypothesis. QRPs are treated differently than fraud because QRPs can sometimes be used for legitimate purposes."* (John, Loewenstein, & Prelec, 2012)

Kde je zakopaný pes?

Questionable Research Practices	OK
1. Not reporting "failed" studies.	83%
2. Not reporting DVs if not significant	92%
3. Not reporting Conditions that "did not work"	89%
4. Excluding data based on effect on p-value.	81%
5. Stopping data collection when significant.	89%
6. Reporting unexpected results "as predicted"	75%

(John, Loewenstein, & Prelec, 2012)

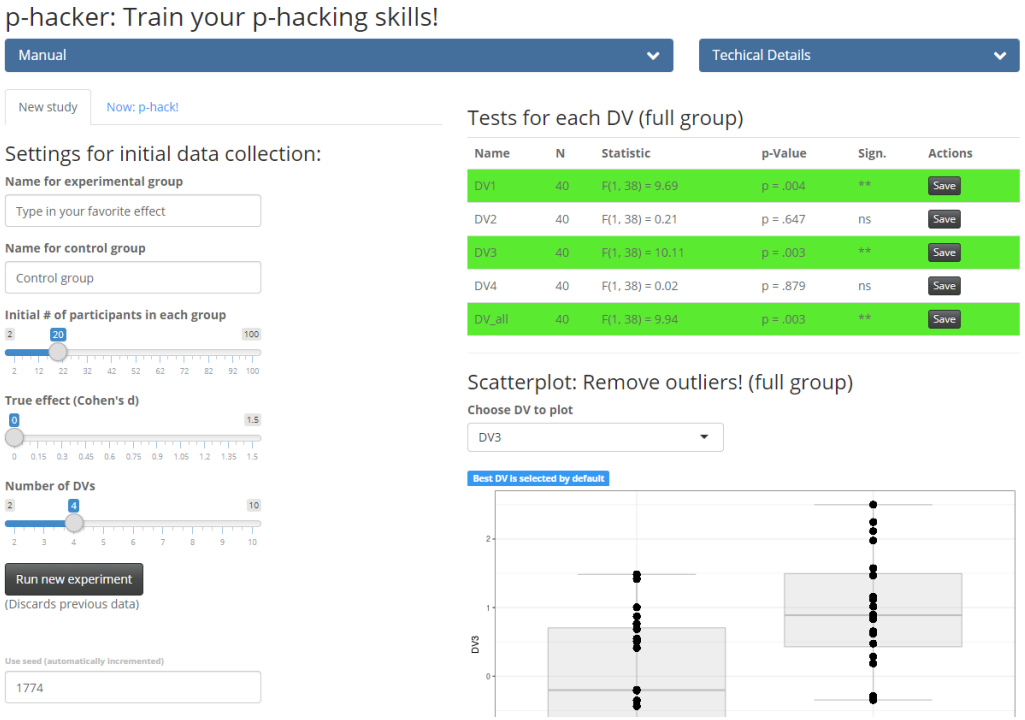
Table 1. Likelihood of Obtaining a False-Positive Result

	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%

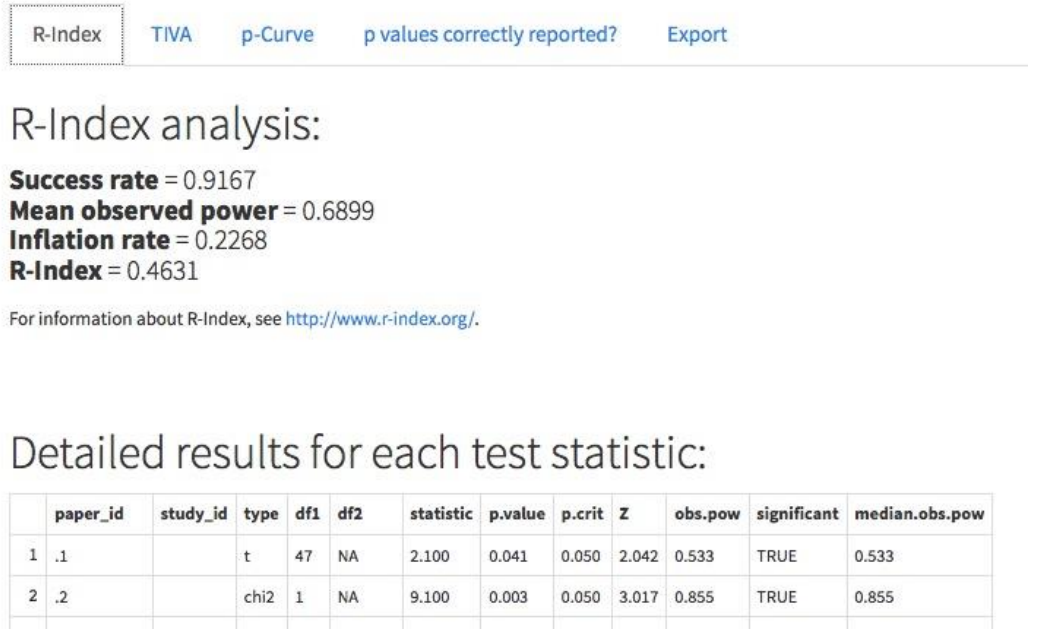
(Simmons, Nelson, & Simonsohn, 2011)

Kontrola předchozích zjištění

P-HACKER



P-CHECKER



	paper_id	study_id	type	df1	df2	statistic	p.value	p.crit	Z	obs.pow	significant	median.obs.pow
1	.1		t	47	NA	2.100	0.041	0.050	2.042	0.533	TRUE	0.533
2	.2		chi2	1	NA	9.100	0.003	0.050	3.017	0.855	TRUE	0.855

Příklady nereplikovatelny

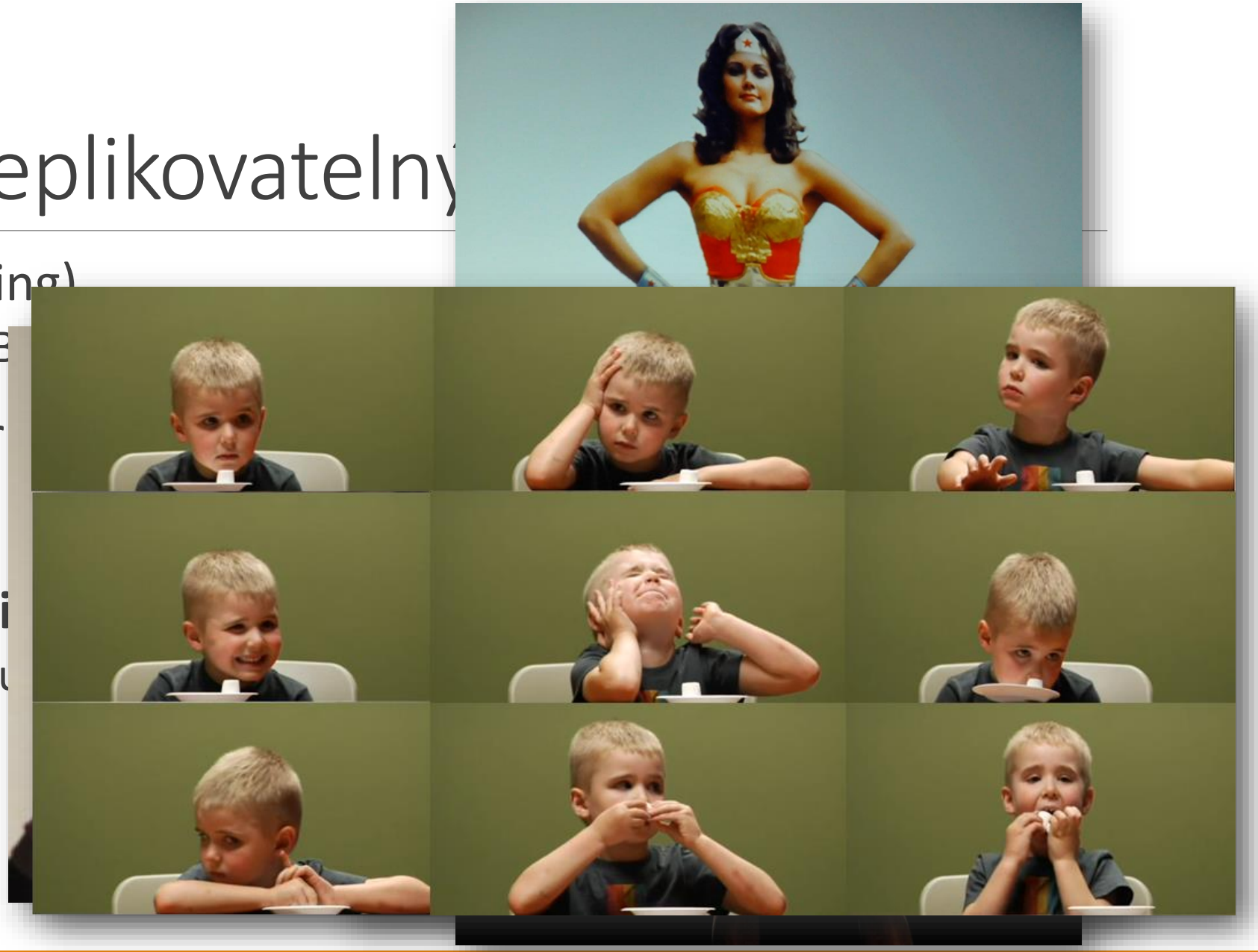
Priming (social priming)
- elderly priming, MacB

Ego deplation (vyčer

Power posing

Vybrané aspekty **faci**
- „smiling will make you

Marshmallow test



Příklady nereplikovatelných efektů

Priming (social priming).

- elderly priming, MacBeth effect, cleanliness priming, money priming...

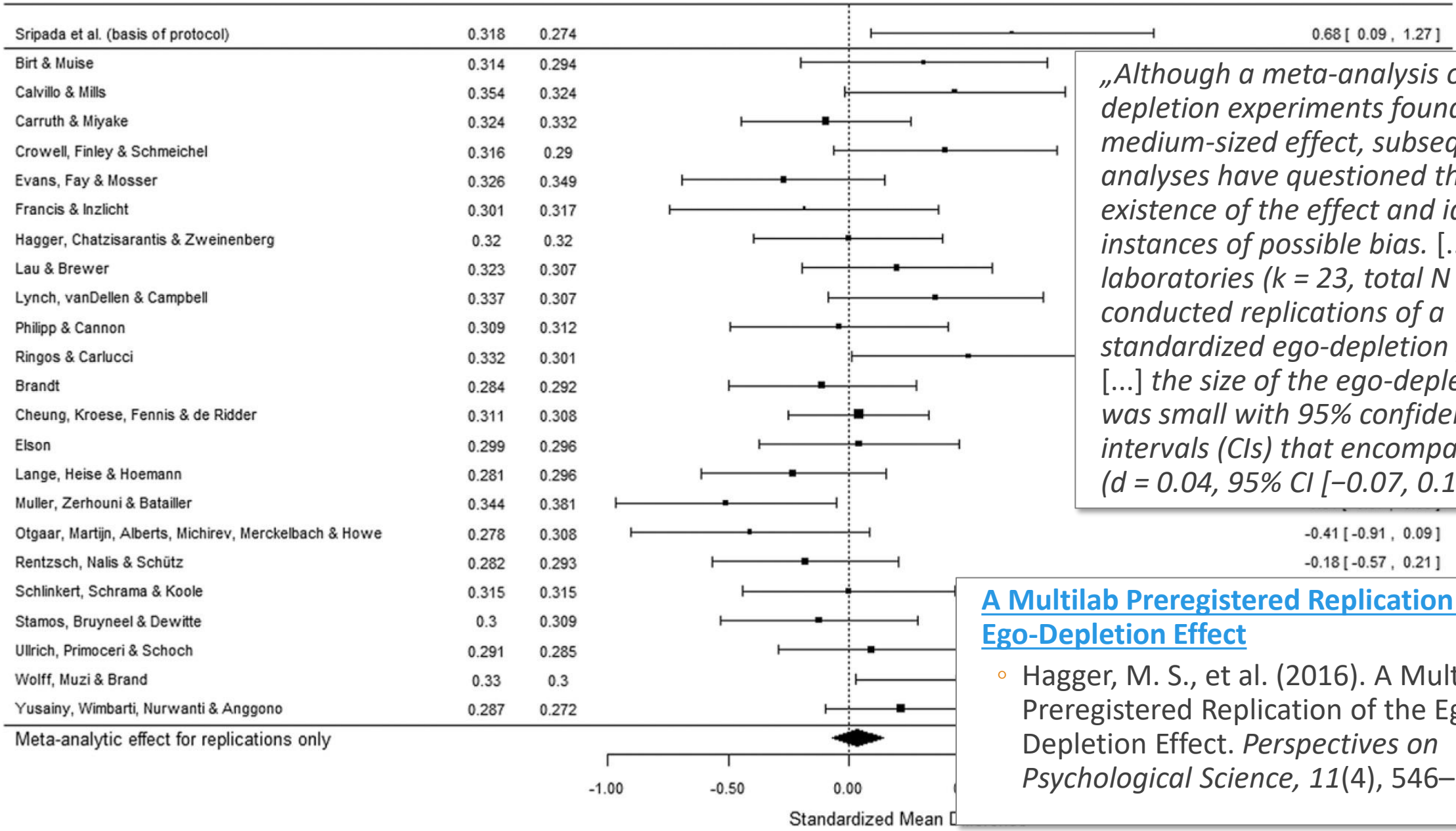
Ego deplation (vyčerpání ega).

Power posing

Vybrané aspekty **facial-feedback hypothesis**

- „smiling will make you feel happier“

Marshmallow test



„Although a meta-analysis of ego-depletion experiments found a medium-sized effect, subsequent meta-analyses have questioned the size and existence of the effect and identified instances of possible bias. [...] *Multiple laboratories (k = 23, total N = 2,141) conducted replications of a standardized ego-depletion protocol* [...] *the size of the ego-depletion effect was small with 95% confidence intervals (CIs) that encompassed zero (d = 0.04, 95% CI [−0.07, 0.15].“*

A Multilab Preregistered Replication of the Ego-Depletion Effect

- Hagger, M. S., et al. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science, 11*(4), 546–573.

„This research tested variation in the replicability of 13 classic and contemporary effects across 36 independent samples totaling 6,344 participants [...] We compared whether the conditions such as lab versus online or US versus international sample predicted effect magnitudes. By and large they did not."

Many Labs 1

- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142-152.

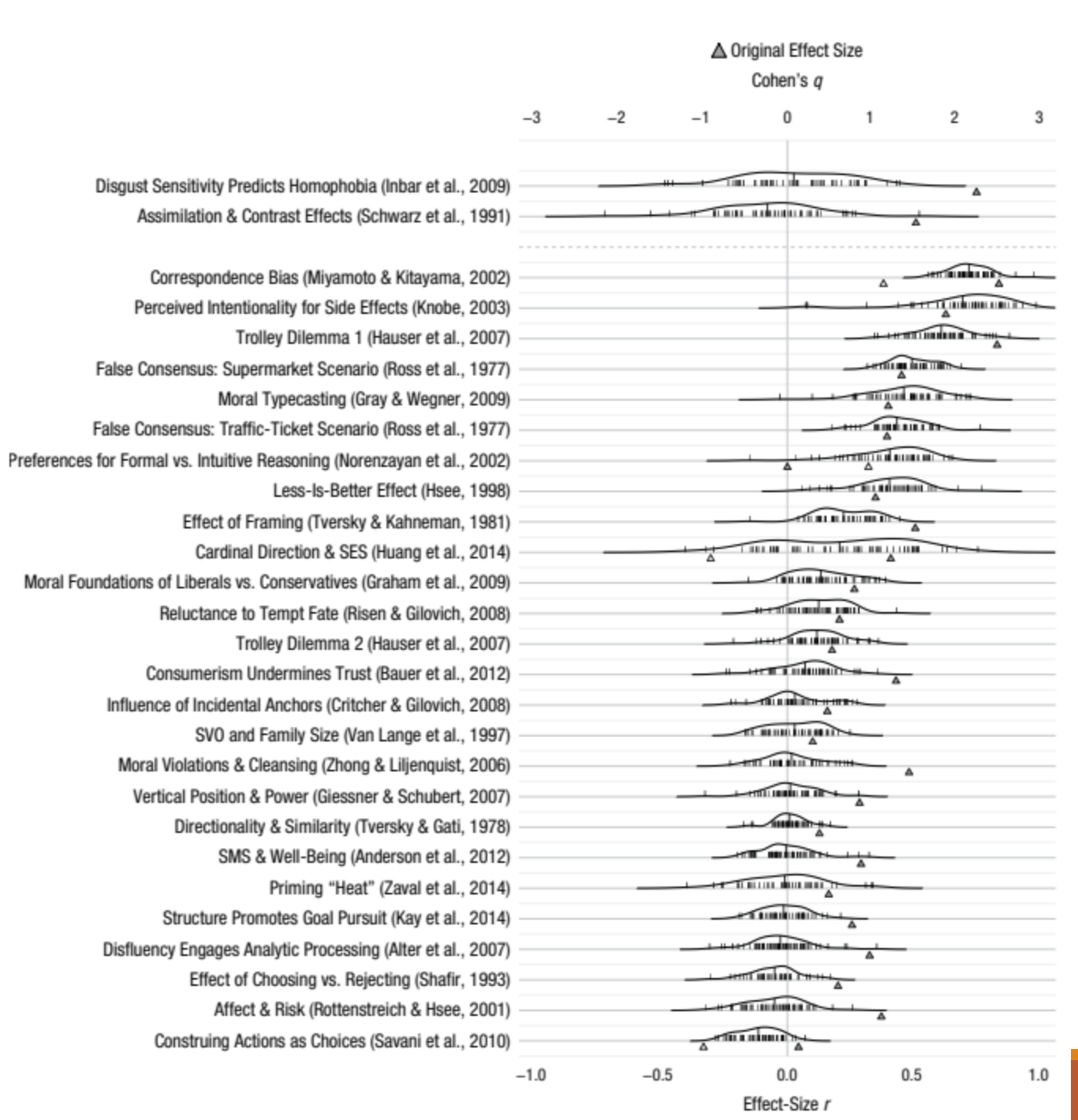
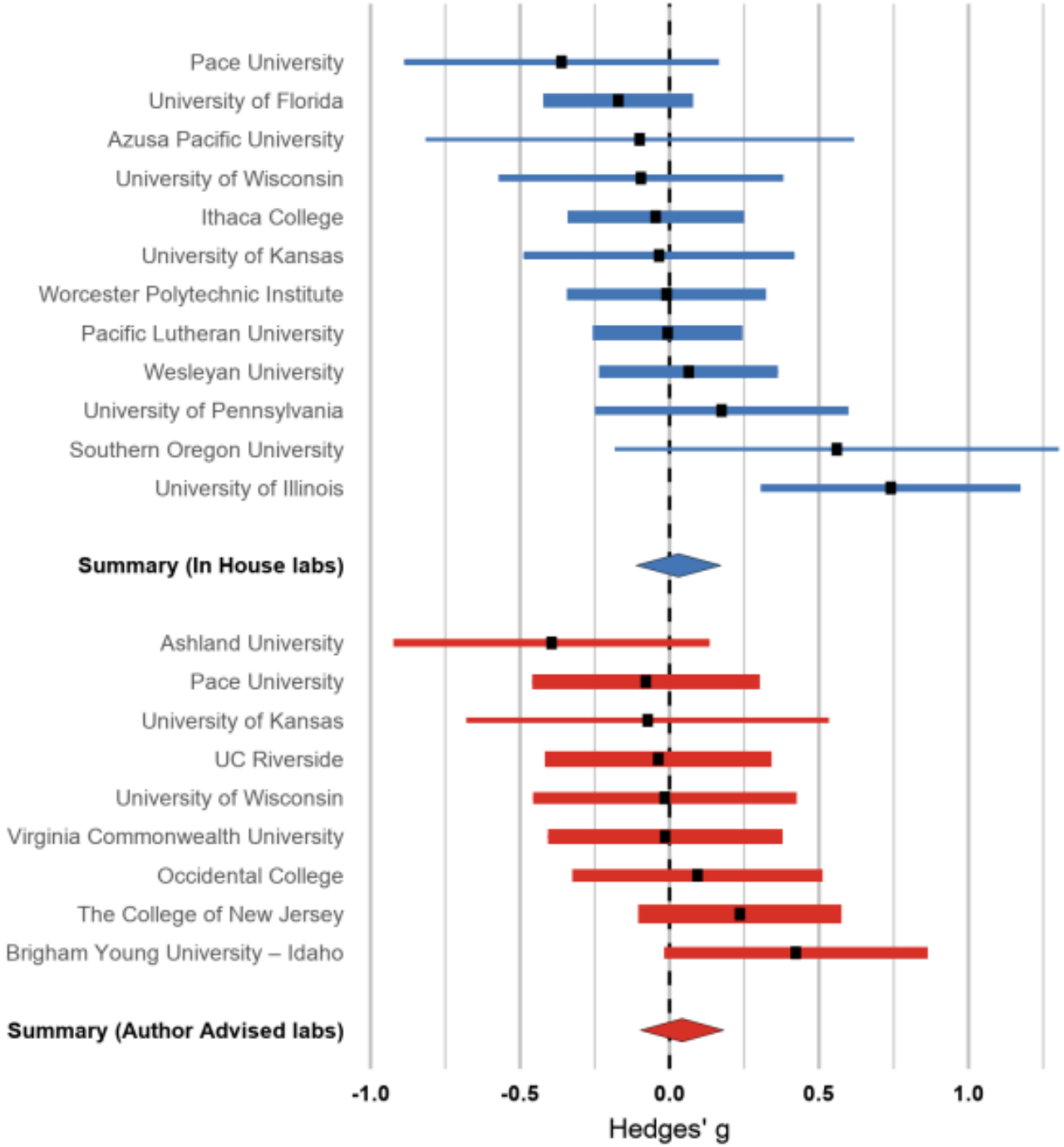


Fig. 2. Effect-size distributions for the 28 effects. The effect size for each replication sample is plotted as a short vertical line; the aggregate estimates are plotted as longer, thick vertical lines. Results for samples with fewer than 15 participants because of exclusions are not plotted, and some samples were excluded because of errors in administration. A detailed accounting of all exclusions is available at https://manylabsopenscience.github.io/ML2_data_cleaning. Positive effect sizes indicate effects consistent with the direction of the original findings in

„Across settings, the Q statistic indicated significant heterogeneity in 11 (39%) of the replication effects, and most of those were among the findings with the largest overall effect sizes; only 1 effect that was near zero in the aggregate showed significant heterogeneity according to this measure. [...] Moderation tests indicated that very little heterogeneity was attributable to the order in which the tasks were performed or whether the tasks were administered in lab versus online. [...] Cumulatively, variability in the observed effect sizes was attributable more to the effect being studied than to the sample or setting in which it was studied."

Many Labs 2

- Klein, R. A., et al. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490.



„We (N = 21 Labs and N = 2,220 participants) experimentally tested whether original author involvement improved replicability of a classic finding from Terror Management Theory (Greenberg et al., 1994). Our results were non-diagnostic of whether original author involvement improves replicability because we were unable to replicate the finding under any conditions. This suggests that the original finding was either a false positive or the conditions necessary to obtain it are not yet understood or no longer exist.“

Many Labs 4

- Klein, R. A., et al. (2019, December 11). Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement.
- preprint

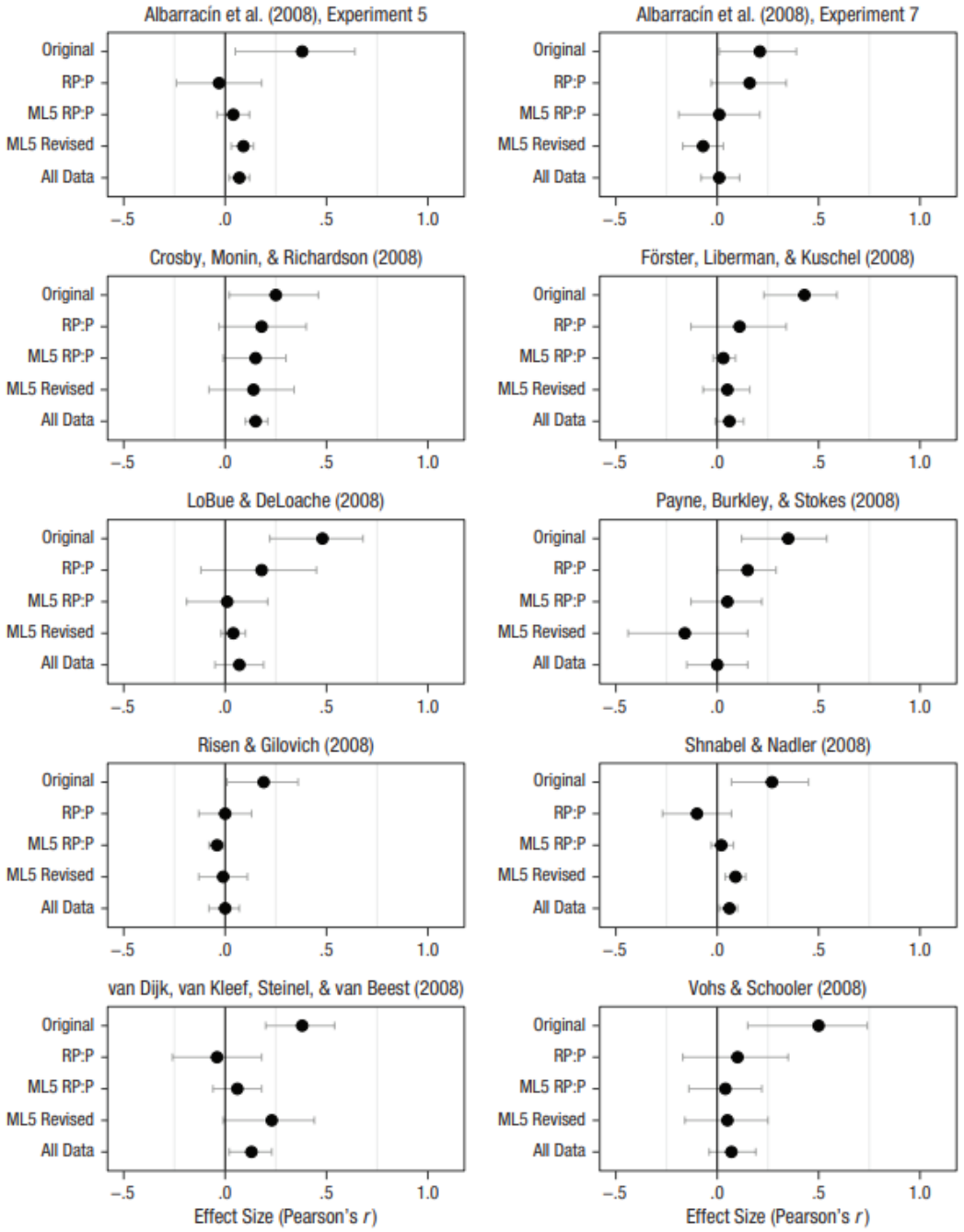
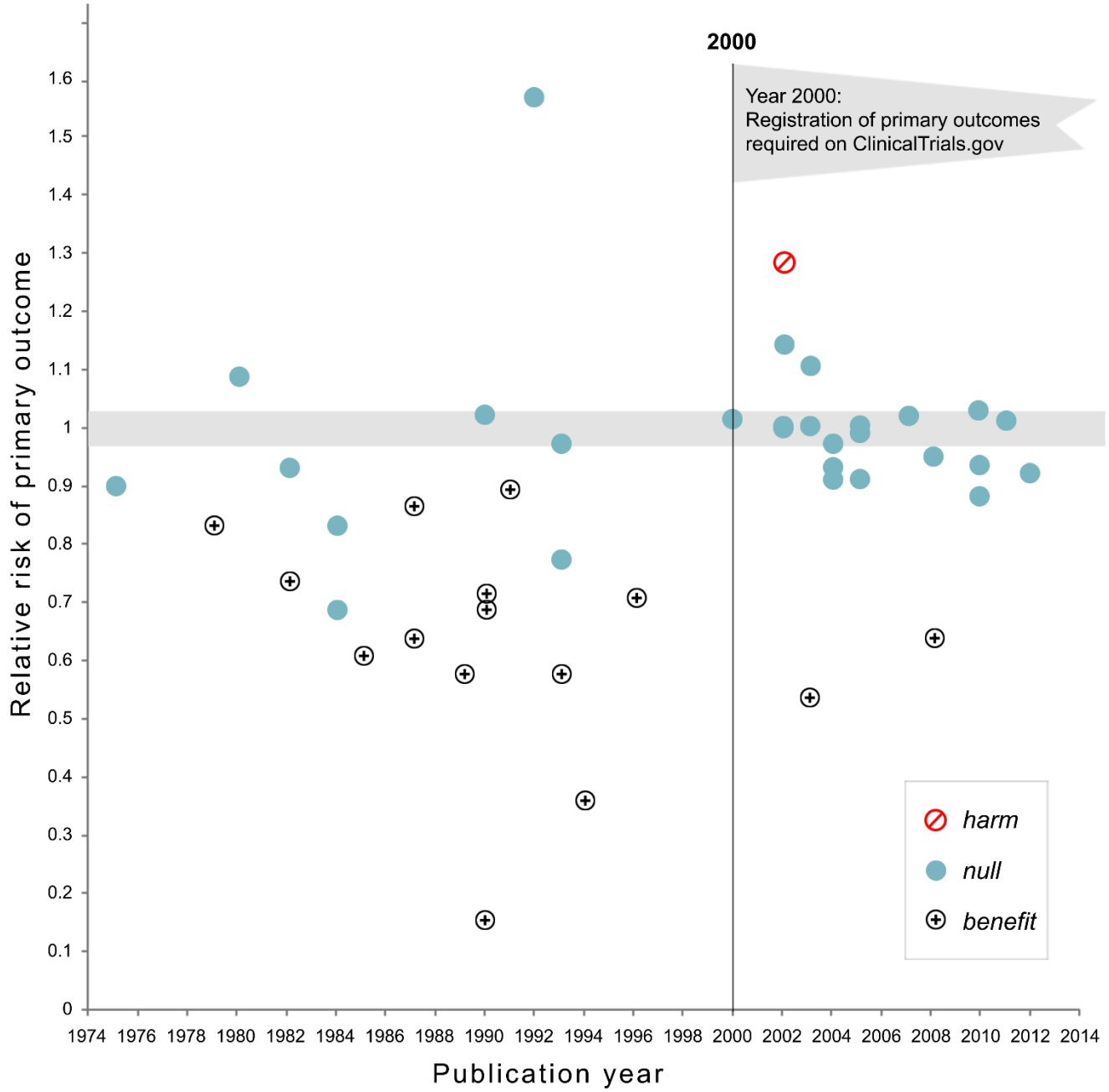


Fig. 2. Effect sizes from the 10 original studies and their replications in the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015) and the Many Labs 5 (ML5) protocols. The "All Data" results are estimates from random-effects meta-analyses including the original studies and their replications. Error bars represent 95% confidence intervals.

„If these [replication] *studies use methods that are unfaithful to the original study or ineffective in eliciting the phenomenon of interest, then a failure to replicate may be a failure of the protocol rather than a challenge to the original finding. Formal pre-data-collection peer review by experts may address shortcomings and increase replicability rates.* [...] *Overall, following the preregistered analysis plan, we found that the revised protocols produced effect sizes similar to those of the RP:P protocols (Δr = .002 or .014, depending on analytic approach).“*

Many Labs 5

- Ebersole, C.R., et al. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. Advances in Methods and Practices in Psychological Science, 3(3), 309–331.



„We identified all large NHLBI supported RCTs between 1970 and 2012 evaluating drugs or dietary supplements for the treatment or prevention of cardiovascular disease. Trials were included if direct costs >$500,000/year, participants were adult humans, and the primary outcome was cardiovascular risk, disease or death. [...] *The number NHLBI trials reporting positive results declined after the year 2000. Prospective declaration of outcomes in RCTs, and the adoption of transparent reporting standards, as required by clinicaltrials.gov, may have contributed to the trend toward null findings.“*

Replikační krize nejen v psychologii.

- Kaplan, R.M., Irvin, V.L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. PLoS ONE 10(8): e0132382.

Aktuální kauza: Dan Ariely

Simonsohn, U., Nelson, L., & Simmons, J. (Srpen 17, 2021). Evidence of Fraud in an Influential Field Experiment About Dishonesty. *Data Colada.* https://datacolada.org/98

This is Table 1 in Kristal et al. (2020), reporting their re-analysis of Shu et al. (2012)

	Sign-at-the-bottom, means (SD)	Sign-at-the-top, means (SD)	Two-sided *t* test, values
Baseline odometer reading (*t*0)	75,034.50 (50,265.35)	59,692.71 (49,953.51)	$t_{(13,474)} = 17.78$, $P < 0.0001$
New odometer reading (*t*1)	98,705.14 (51,934.76)	85,791.10 (51,701.31)	$t_{(13,475)} = 14.47$, $P < 0.0001$
Difference in odometer readings; i.e., miles driven (*t*1–*t*0)*	23,670.64 (12,621.38)	26,098.40 (12,253.37)	$t_{(13,448)} = -11.331$, $P < 0.0001$

*This row was the outcome reported in the original paper.

Figure 1. Histogram of Miles Driven - Car #1 (N=13,488)

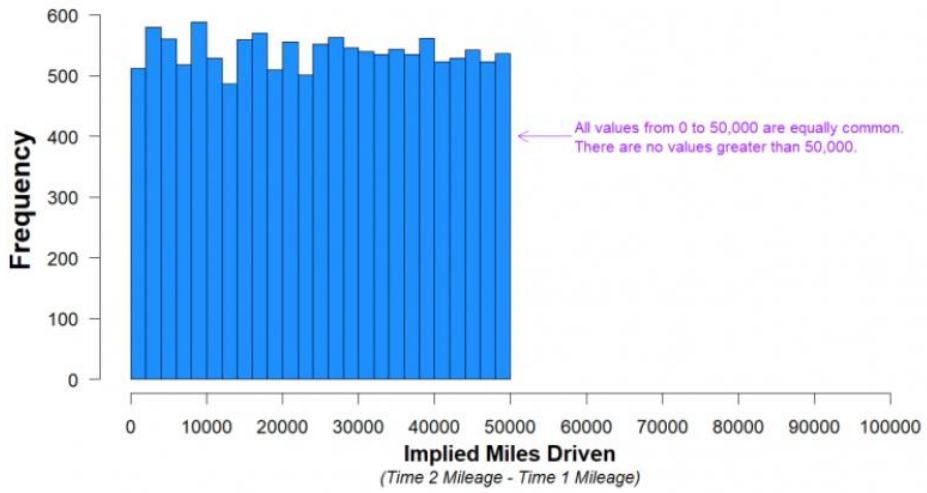
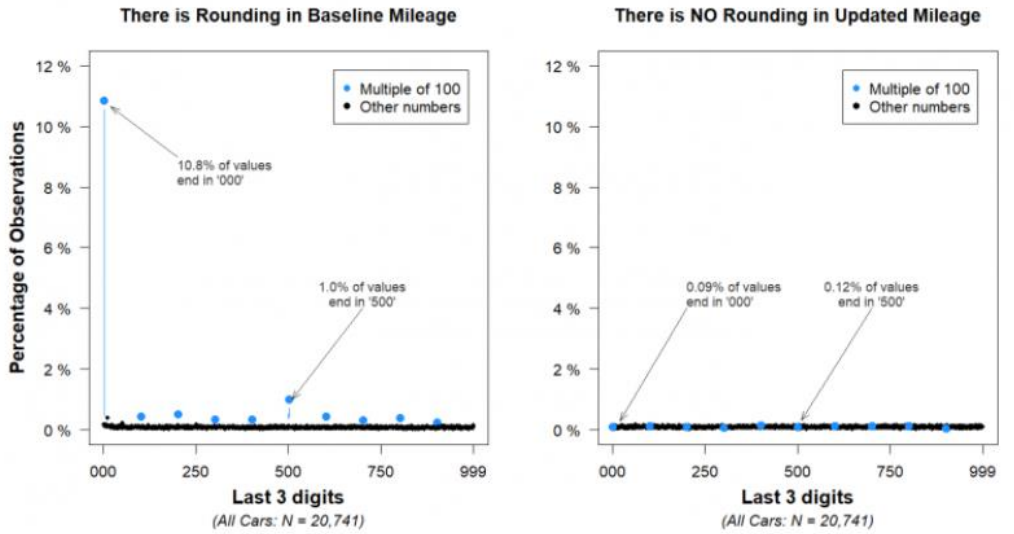


Figure 3. Last Three Digits at Baseline (Time 1) vs Updated (Time 2)



Disclaimer

Susan Fiske: „Metodological terrorism“, „self-appointed data police“.

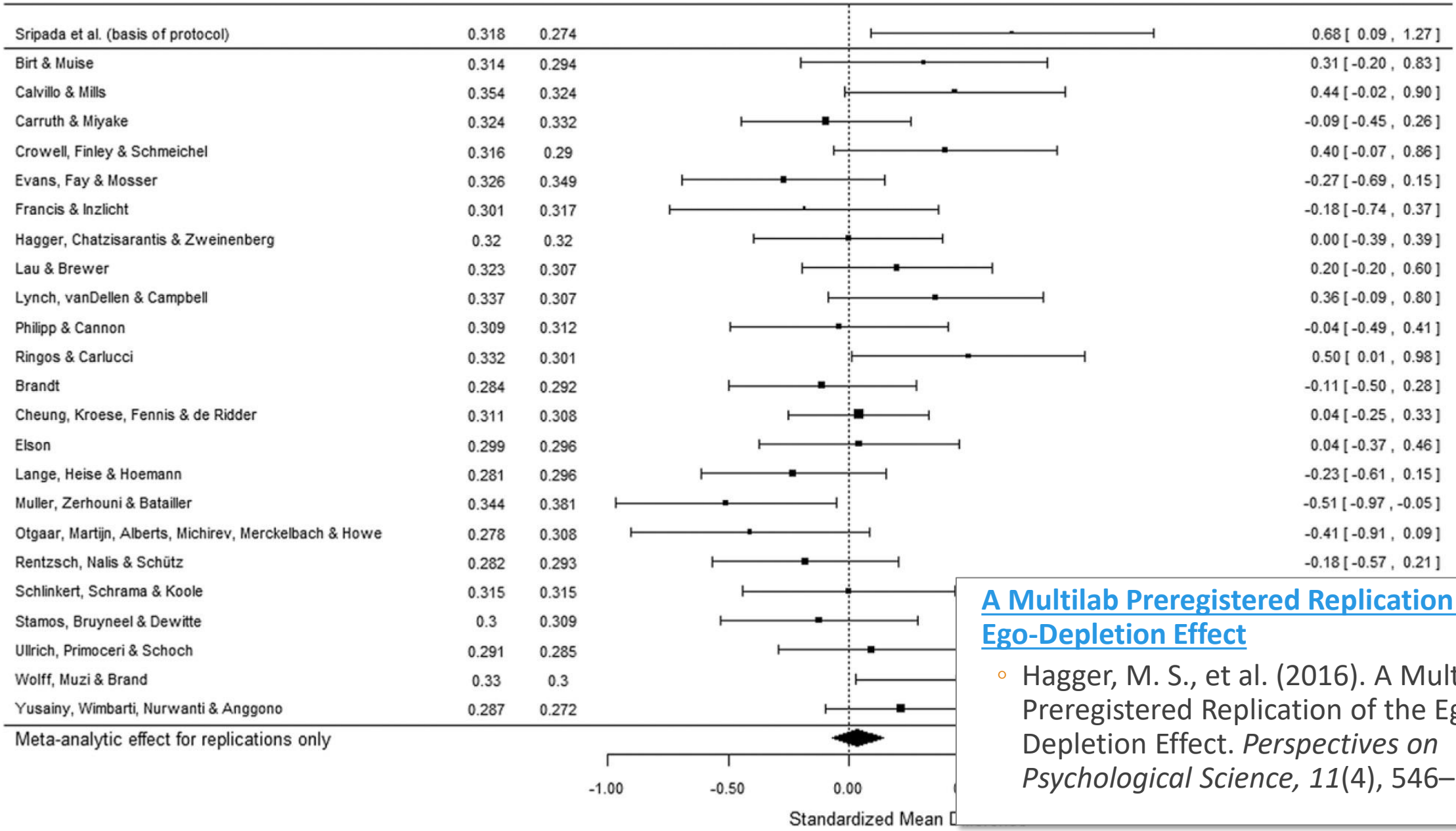
Kontroverze.

Ztráta důvěry ve vědu.

Osobní zodpovědnost výzkumníků?

„Tak se to dělalo...“

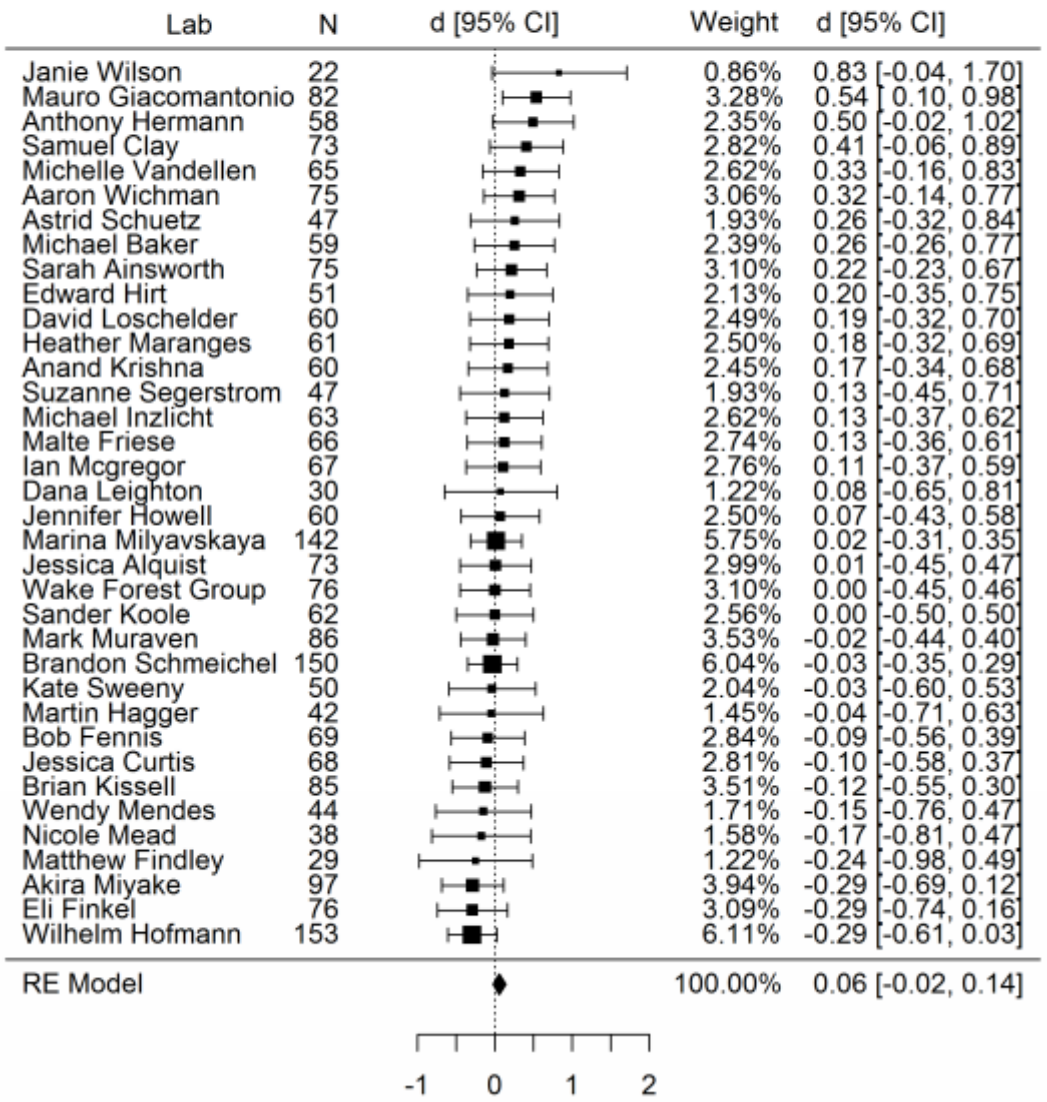
Běžná praxe.



A Multilab Preregistered Replication of the Ego-Depletion Effect

- Hagger, M. S., et al. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science, 11*(4), 546–573.

Figure 1. *Forest Plot of Performance Outcome by Laboratory*. The box plots and numerical values illustrate the same effect size estimates. For the plots, the size of the box represents its weighted contribution to the overall effect and its whiskers display 95% CIs. The dotted line represents a zero effect size. Numerical values show standardized mean differences between depletion and non-depletion conditions expressed in Cohen's *d* (with 95% CIs). The diamond is the overall meta-analytic effect derived from a random-effects model.



„We conducted a preregistered multi-laboratory project (k = 36; N = 3531) to assess the size and robustness of ego depletion effects using a novel replication method, termed the paradigmatic replication approach. [...] *non-significant result, d = 0.06. Confirmatory Bayesian meta-analyses using an informed prior hypothesis (δ = 0.30; SD = 0.15) found the data were four times more likely under the null than the alternative hypothesis. Hence, preregistered analyses did not find evidence for a depletion effect.“*

Vohs, K., et al. (2021). A multi-site preregistered paradigmatic test of the ego depletion effect. *Psychological Science.* https://doi.org/10.1177/0956797621989733

Nástroje k odhalení QRP

Egerův test (z-test) a funnel plot.

P-curve: Rozložení (resp. zešikmení) p-hodnot $p < 0,05$.

- Dobré rozložení: zprava zešikmené. QRP: zleva zešikmené (většina p-hodnot blízko cut-offu).

Z-curve: Srovnání pozorovaného „success-rate“ a mediánu statistické síly.

- R-index: Odhad podílu studií, které by bylo možné replikovat.

„Test of insufficient variance“ (TIVA):

- P-hodnoty převedené na z-skóry by měly být normálně rozdělené (SD=1).

GRIM test: Detekce nemožných průměrů.

- Některé hodnoty desetinných míst nejsou přípustné v případě malých vzorků.
- http://www.prepubmed.org/grim_test/

P-checker: https://shinyapps.org/apps/p-checker/

Reproducibility, replicability, generalizability

Reproducibility (Reprodukovatelnost)

- *„Researcher B must have the following: (a) the **raw data**; (b) the **code book** (variable names and labels, value labels, and codes formissing data); and (c) knowledge of **the analyses** that were performed by Researcher A (e.g. the syntax of a statistics program).“*

Replicability (Replikovatelnost)

- *„The **finding can be obtained with other random samples** drawn from a multidimensional space that captures the most important facets of the research design. In psychology, the facets typically include the following: (a) **individuals** (or dyads or groups); (b) **situations** (natural or experimental); (c) **operationalizations** (experimental manipulations, methods, and measures); and (d) **time points**.“*

Generalizability (Zobecnitelnost)

- *„It does not depend on an originally unmeasured variable that has a systematic effect. In psychology, generalizability is often demonstrated by showing that a **potential moderator variable has no effect** on a group difference or correlation.“*

Asendorpf et al. (2013)

„Slavíme“ 10 let replikační krize v psychologii (2011–2021)

Hlavní změna paradigmatu:

Replikační krize → **krize zobecnitelnosti** nebo též **krize důvěryhodnosti**.

- (*generalizability* or *credibility* crisis)

Měření v psychologii a replikovatelnost

Lilienfeld, S. O., & Strother, A. N. (2020). **Psychological measurement and the replication crisis: Four sacred cows.** *Canadian Psychology, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

„Questionable Measurement Practices“ (QMP)

Namísto „measurement“ často spíše *„schmeasurement“* (Flake & Field, 2020).

Lilienfeld & Strother (2020): Nedostatečná kvalita měření...

- ... snižuje věrohodnost výzkumných zjištění a ohrožuje interní validitu výzkumu;
- ... snižuje a zkresluje velikosti pozorovaných efektů;
- ... a snižuje reprodukovatelnost a hlavně zobecnitelnost výzkumných zjištění.

QMP mohou být jednou z dílčích příčin krize zobecnitelnosti.

V důsledku pak nedostatky v *měření* snižují kvalitu *vědy*, protože měření v širším slova smyslu je základním nástrojem vědy.

„Posvátné krávy“ měření v psychologii

1. Obsahová validita a spoléhání se na „název“ škál.
   - Škály se stejným názvem nemusí měřit to stejné.
   - Pro připomenutí: klasická testová teorie a operacionalismus.

2. Ignorování chyby měření a reliability v laboratorních experimentech.
   - Přesvědčení, že pro výzkum postačuje nižší reliabilita (rovněž i Helmstadter).
   - Behaviorální pozorování (vysoce reliabilní) není totožné s měřeným rysem (vztah může být vágní).
   - A jaká je reliabilita experimentální manipulace?

4. Důraz na konvergentní, nikoli divergentní validitu.
   - Konstruktově irelevantní rozptyl, nedostatek diferenciální validity.
   - Potíže zejména při výzkumu silně korelovaných jevů.

(3. Náročnost sběru dat opravňuje malé velikosti vzorku.)

Lilienfeld, S. O., & Strother, A. N. (2020). **Psychological measurement and the replication crisis: Four sacred cows.** *Canadian Psychology, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

Krize replikovatelnosti: jeden z příznaků **krize zobecnitelnosti**

Yarkoni, T. (2020). **The generalizability crisis.** *Behavioral and Brain Sciences* [preprint], 1–37. https://doi.org/10.1017/S0140525X20001685

Psychologický výzkum je příliš orientovaný na pozorované proměnné namísto na konstrukty.

- 1. Nedostatek konstruktové validity ve smyslu Cronbacha a Meehla.
- 2. Zanedbání hypotetických zdrojů variability výsledků.

Statistické modely jsou jen alternativním „jazykem“ k popisu skutečnosti.

- Při „překladu“ našich otázek do jazyka statistiky a výsledků zpět dochází k chybám.

Doporučuji Yarkoniho číst **až po** přednáškách o epistemologii a teorii zobecnitelnosti.

Klíčové příznaky krize zobecnitelnosti

#1: Psychologové zanedbávají, že různé stimuly, položky dotazníku, operacionalizace konstruktů apod. jsou pouze „vzorky“ z univerza/domény „přípustných“ vzorků.

- Při „překladu“ VO do statistického modelu nejsou operacionalizovány informace o tomto „náhodném“ výběru vzorku pozorování.
- Při překladu výsledků zpět nejsou brány v potaz limity vyplývající z operacionalizace.

#2: Ignorace náhodného výběru zkresluje odhady parametrů. Druhy efektů[1]:

- **Pevné (fixed) efekty:** zpravidla zkoumaný efekt. Není vybrán z domény, je specifický pro danou situaci. Výsledky *nechceme generalizovat* na jiné pevné efekty.
- **Náhodné (random) efekty:** kontrolují náhodu spjatou s výběrem prvků z domény. *Chceme zobecňovat* efekt i na jiné prvky/výběry z dané domény.

„**Fixed-effect fallacy**“: V psychologii bývá zpravidla kontrolovaná náhoda spjatá pouze s between-subject variabilitou (lidmi/subjekty).

- Méně často se situací, laboratoří, stimuly a podobně („stimulus-as-fixed effect fallacy“).

[1] Ve shodě s Yarkonim (2020) používám terminologii generalizovaného lineárního smíšeného modelu (GLMM).

Příklad 1: Stroopův efekt

Příklad: Stroopův efekt.

- Simulace: 20 simulovaných datasetů o 20 osobách.
- Osa X: pozorovaný efekt ve studii.
- Osay Y: číslo experimentu.

Vlevo: between-subject variabilita je ignorovaná.

- Heterogenní výsledky studií.
- Neumožňuje zobecňovat na lidi obecně, ale jen „uvnitř“ vzorku.

Vpravo: Rozdíl lidí byl do modelu vložen jako náhodný efekt.

- Homogenní výsledky studií.
- Lze zobecňovat na lidi obecně v dané populaci.

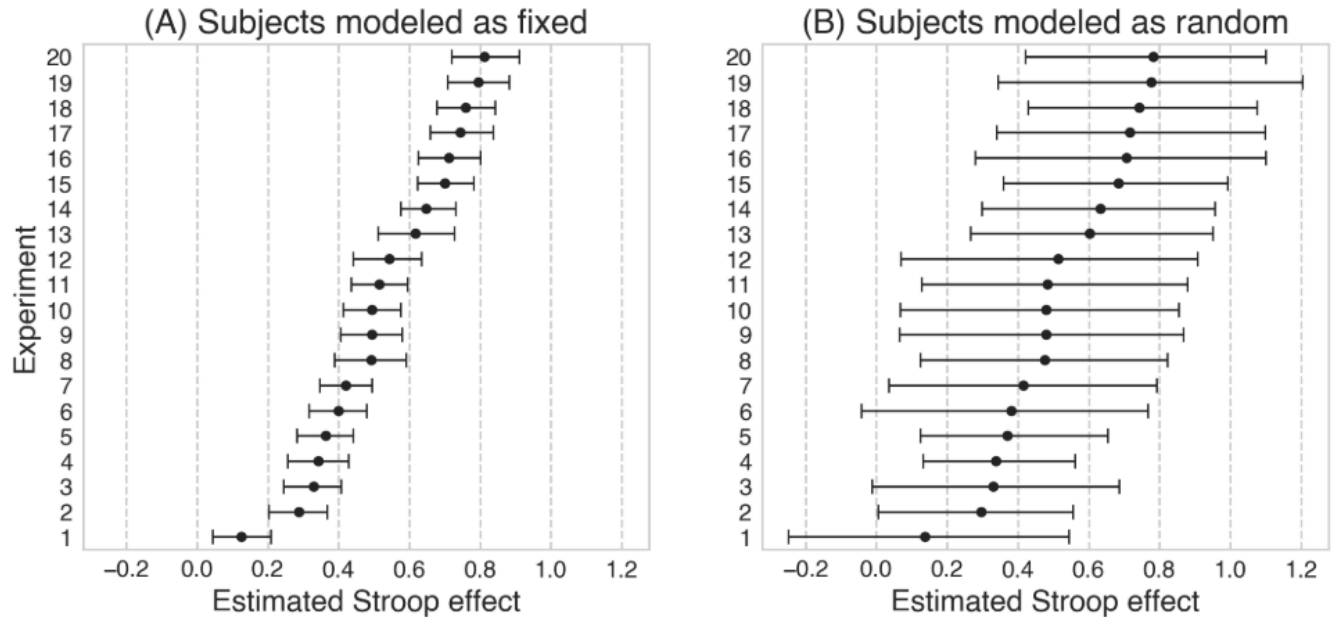


Figure 1: Co
a simulated St
ground truth f
posterior dens
ordered by the magnitude of the point estimate for visual cla
does not account for random subject sampling, and consequer
effect of interest. (B) The random-effects specification in Eq.
appropriately calibrated uncertainty estimates.

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + e_{ij}$$
$$e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$$

del specific
subjects r
l experime
condition

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0i} + u_{1i} X_{ij} + e_{ij}$$
$$u_{0i} \sim \mathcal{N}(0, \sigma_{u_0}^2)$$
$$u_{1i} \sim \mathcal{N}(0, \sigma_{u_1}^2)$$
$$e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$$

Příklad 1: Stroopův efekt

Yarkoni (2020, pp. 6):

- *„... it is the mismatch between our generalization intention and the model specification that introduces **an inflated risk of inferential error**, and not the model specification alone.“*
- *„Empirical studies in domains ranging from social psychology to functional MRI have demonstrated that test **statistic inflation of up to 300% is not uncommon**, and that, under realistic assumptions, **false positive rates in many studies could easily exceed 60%** (Judd et al., 2012; Westfall, Nichols, & Yarkoni, 2016; Wolsiefer, Westfall, & Judd, 2017).“*

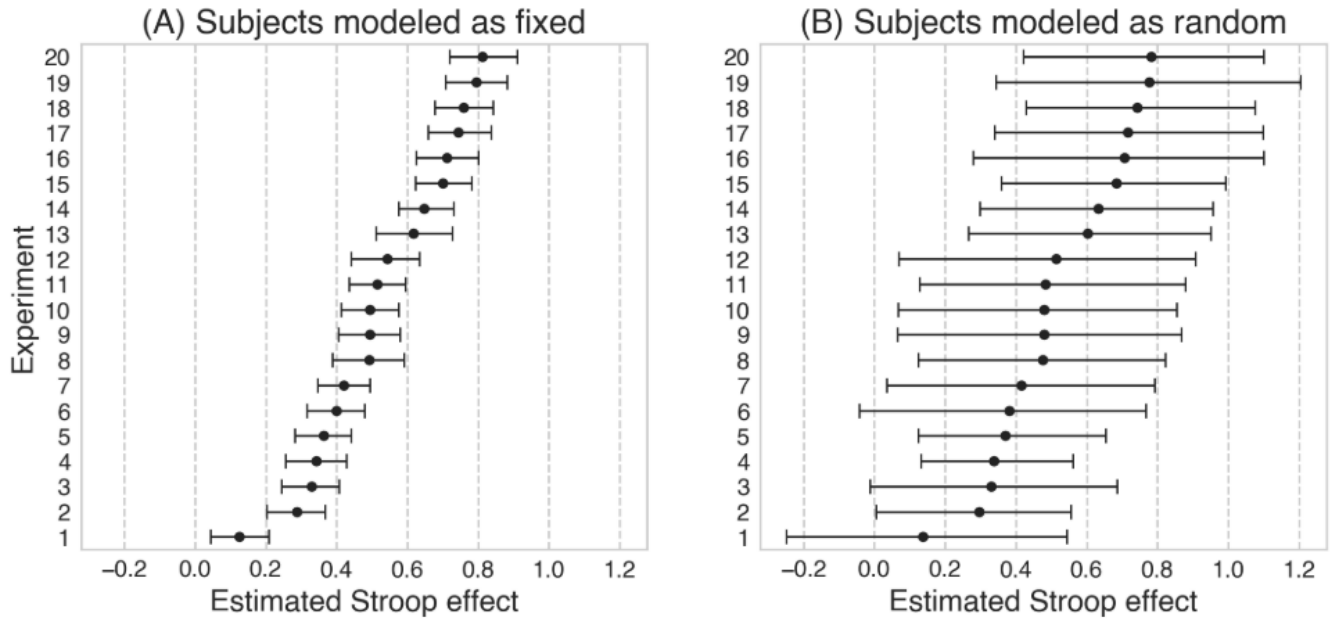


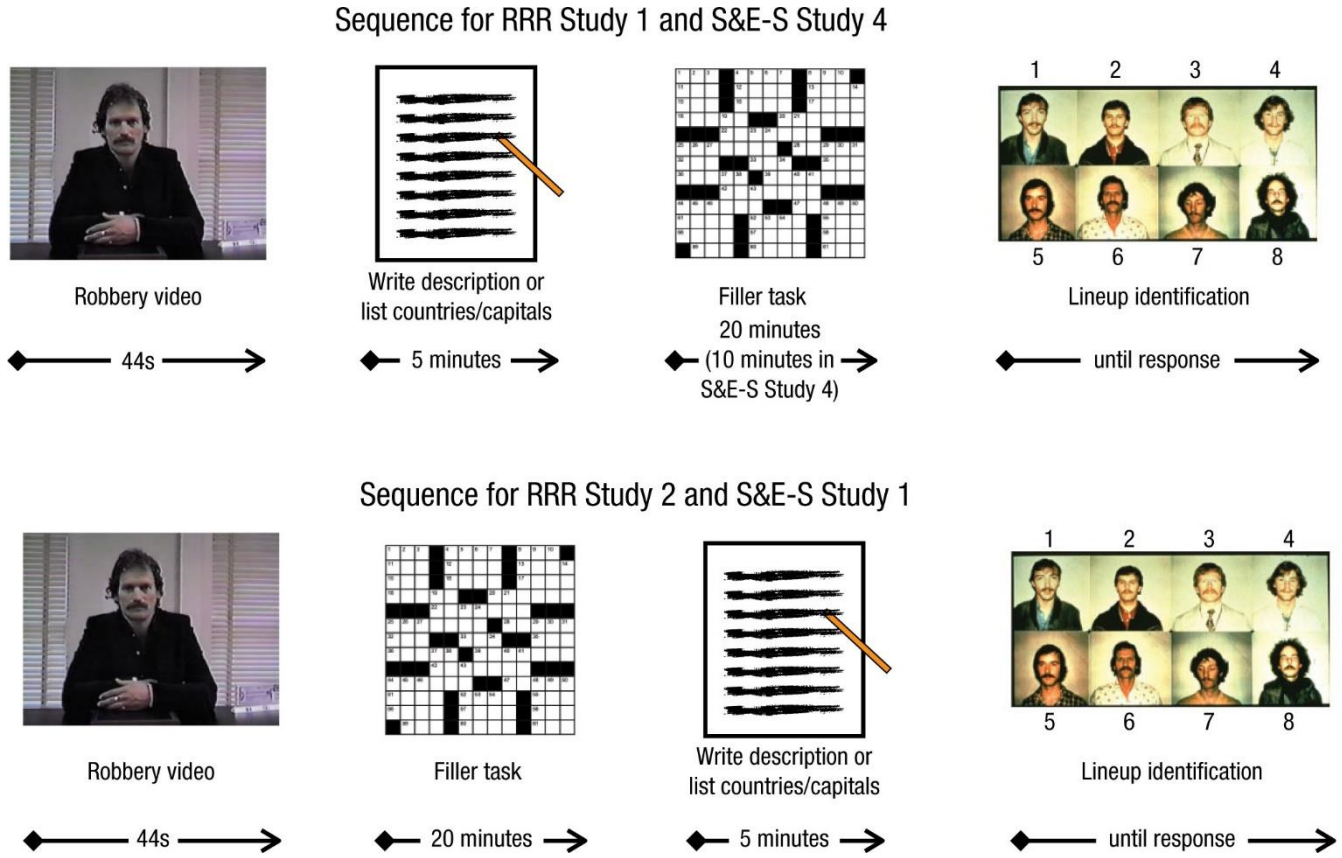
Figure 1: Consequences of mismatch between model specification and generalization intention. Each row represents a simulated Stroop experiment with $n = 20$ new subjects randomly drawn from the same global population (the ground truth for all parameters is constant over all experiments). Bars display the estimated Bayesian 95% highest posterior density (HPD) intervals for the (fixed) condition effect of interest in each experiment. Experiments are ordered by the magnitude of the point estimate for visual clarity. (A) The fixed-effects model specification in Eq. (1) does not account for random subject sampling, and consequently underestimates the uncertainty associated with the effect of interest. (B) The random-effects specification in Eq. (2) takes subject sampling into account, and produces appropriately calibrated uncertainty estimates.

Příklad 2: Verbal overshadowing

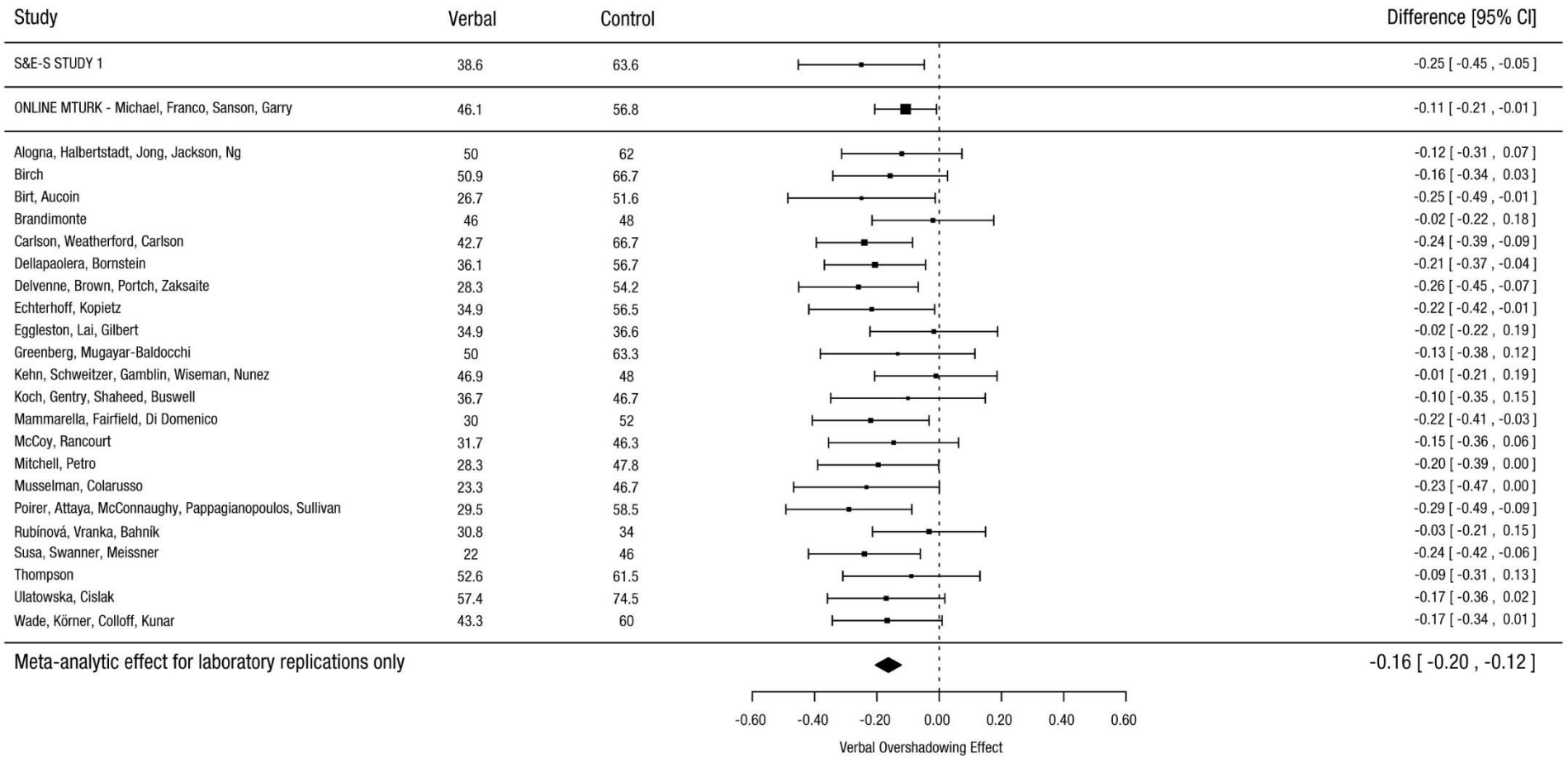
Velká replikační studie „verbálního zastínění“.

- Replikace: Alogna a kol. (2014).
- Originální studie:
  Schooler a Engstler-Schooler (1990)
- 31 laboratoří, N_{tot} > 2000.

„Original authors showed that participants who were asked to verbally describe the appearance of a perpetrator caught com- mitting a crime on video showed poorer recognition of the perpetrator following a delay than did partic- ipants assigned to a control task (naming as many countries and capitals as they could).“



Příklad 2: Verbal overshadowing



Příklad 2: Verbal overshadowing

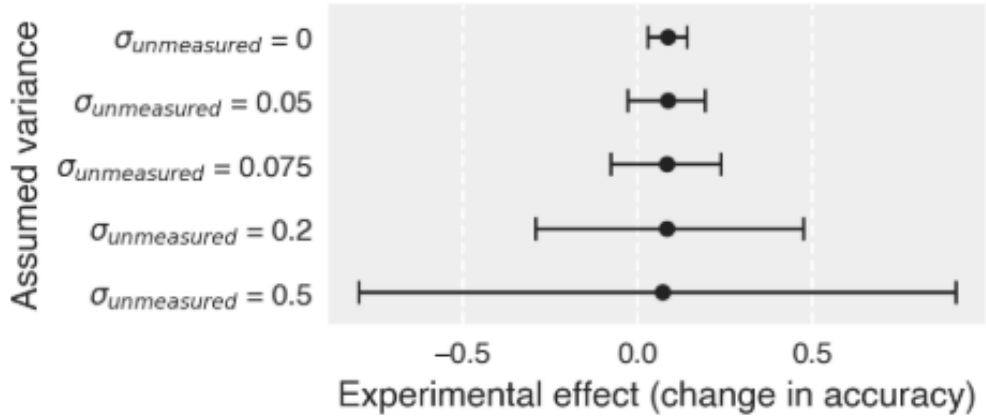
Silný důkaz pro existenci efektu. Sice nižší než originální, ale rostoucí v čase.

Nulová heterogenita výsledků napříč laboratořemi a to včetně MTurk, $I^2 = 0$.

Ale: **Ve shodě s originálními autory pouze jediná nahrávka a jediný line-up.**

- *„The strict conclusion* [...] *is that there is at least one particular video containing one particular face that, when followed by one particular lineup of faces, is more difficult for participants to identify if they previously verbally described the appearance of the target face than if they were asked to name countries and capitals. This narrow conclusion does not preclude the possibility that the observed effect is specific to this one particular stimulus, and that many other potential stimuli the authors could have used would have eliminated or even reversed the observed effect.“* (Yarkoni, 2020, pp. 8).

Pokud by nekontrolované rozdíly ve stimulech (tvářích) měly velmi malý vliv na pozorování SD=0,05 (ve srovnání se zvýšením přesnosti o cca 0,1), souhrnný efekt přestane být signifikantní.



Doporučení pro zvýšení replikovatelnosti psychologického výzkumu

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). **Recommendations for Increasing Replicability in Psychology.** *European Journal of Personality, 27*(2), 108–119. https://doi.org/10.1002/per.1919

Doporučení: Design a analýza

Zmenšit chybu měření

- ... zvýšením velikosti vzorku;
- ... zvýšením statistické síly;
- ... zvýšením reliability měřícího nástroje;
- ... korektním užíváním korekcí pro vícenásobná srovnání,
  - Užívání postupů typu Bonferroniho korekce snižuje statistickou sílu

Od "p < 0,05" k...

- ... reportování skutečné velikosti p-hodnoty;
- ... důrazu na ukazatele velikosti účinku;
- ... důrazu na intervaly spolehlivosti apod.

Doporučení: Publikační proces

Autoři studií, výzkumníci: **transparence**.

- Literature review ve vztahu k dosavadnímu stavu replikace.
  - Existují dřívější replikační studie? Podařilo se původní výsledek replikovat? Apod.
- Zdůvodnění volby velikosti vzorku
- Zveřejnění dat, postupů analýz, work-in-progress, pre-registrací
- Provádění replikací, účast na diskuzích odborné veřejnosti atd.

Žurnály, recenzenti, editoři: **Podpora dobrých výzkumných praktik.**

- Publikování replikací a podpora autorů v této činnosti
- Ústup od konfirmačního zkreslení v publikačním procesu

Doporučení: Vyučující metodologie

Aneb: **Co mají studenti chtít po svých učitelích?**

Rigorózní výuka metodologie, statistické analýzy dat apod.

- Statistická síla, velikost účinku, zobecnitelnost atd.
- Informace o replikovatelnosti efektů při výuce jiných kurzů.

Podpora **transparentnosti**.

- Publikování dat, skriptů apod., analýza takovýchto souborů.

Podpora **studentských replikací**.

- Přínos pro studenty i pro obor.

Podpora **kritického myšlení**.

- Obsahuje studie veškeré podstatné informace? Zvolili výzkumníci vhodnou proceduru pro ověření stanovené hypotézy? Jsou závěry korektně interpretovány?
- Na úrovni jednotlivých studií i v rámci meta-analýz

Doporučení: Instituce

Změna Publish or Perish politiky:

- Počet publikací a impact faktor jako rozhodující proměnná při přidělování grantů, přijetí do zaměstnání či kariérním postupu

Alternativy:

- Oceňování a podpora replikační činnosti
- Vynaložení části prostředků v rámci výzkumu na replikaci

Doporučení: Obor

Přesun od efektů k teoriím.

Přesun od dílčích studií k agregaci výzkumného poznání.

Větší důraz na způsob, kvalitu a podstatu měření.

- Vzhledem k měřenému atributu.

Větší míra standardizace výzkumných nástrojů.

Adekvátní statistické postupy.

Příklady dobré praxe

Registered Replication Report

Registered Replication Reports

Multi-lab, high-quality replications of important experiments in psychological science along with comments by the authors of the original studies.

Quick Links

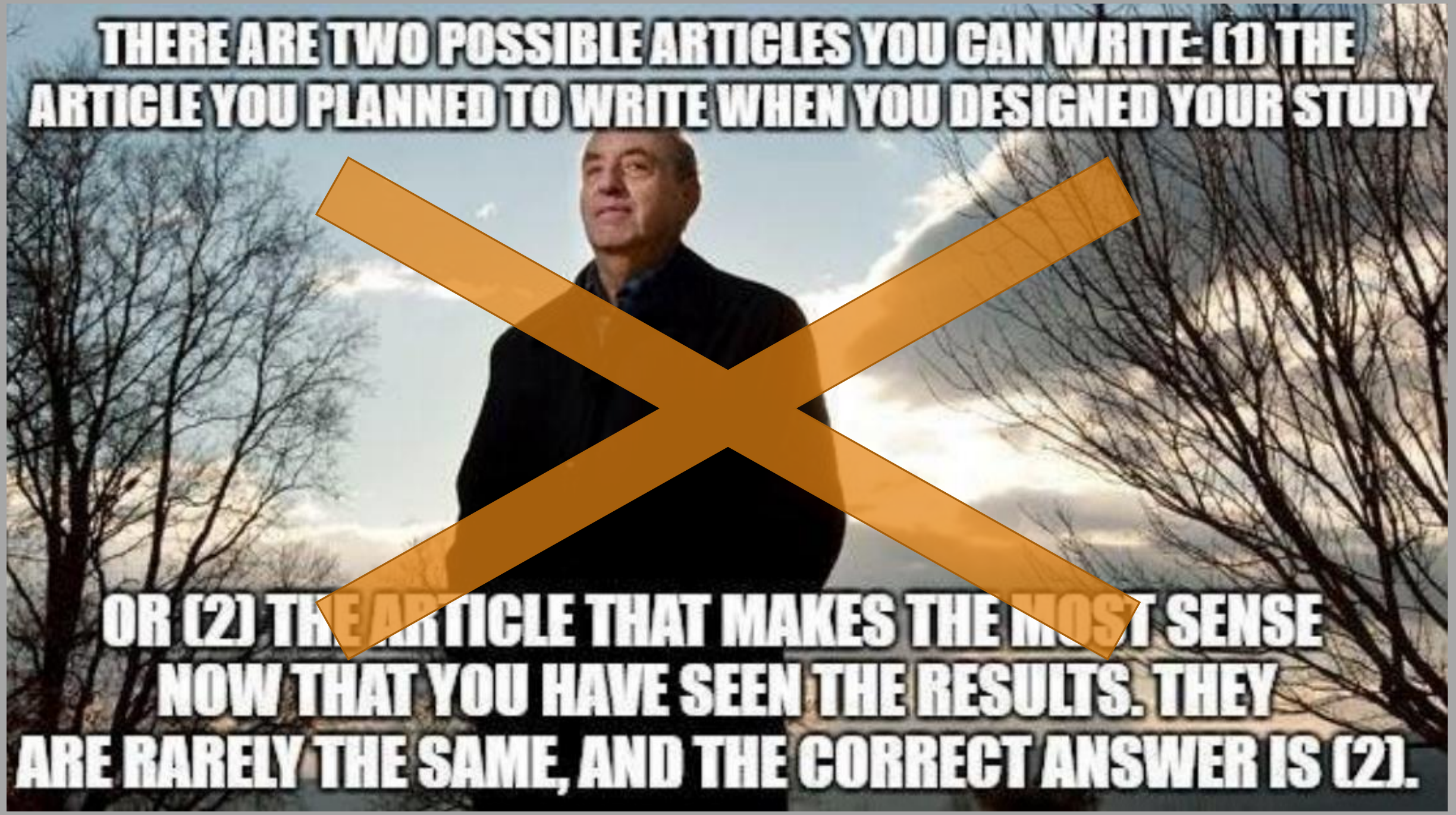
Mission Statement

Replicability is a cornerstone of science. Yet replication studies rarely appear in psychology journals. The new **Registered Replication Reports** article type in *Perspectives on Psychological Science* fortifies the foundation of psychological science by publishing collections of replications based on a shared and vetted protocol. It is motivated by the following principles:

• Psychological science should emphasize findings that are robust, replicable, and generalizable.

• Direct replications are necessary to estimate the true size of an effect.

• Well-designed replication studies should be published regardless of the size of the effect or statistical significance of the result.


THERE ARE TWO POSSIBLE ARTICLES YOU CAN WRITE: (1) THE ARTICLE YOU PLANNED TO WRITE WHEN YOU DESIGNED YOUR STUDY
OR (2) THE ARTICLE THAT MAKES THE MOST SENSE NOW THAT YOU HAVE SEEN THE RESULTS. THEY ARE RARELY THE SAME, AND THE CORRECT ANSWER IS (2).

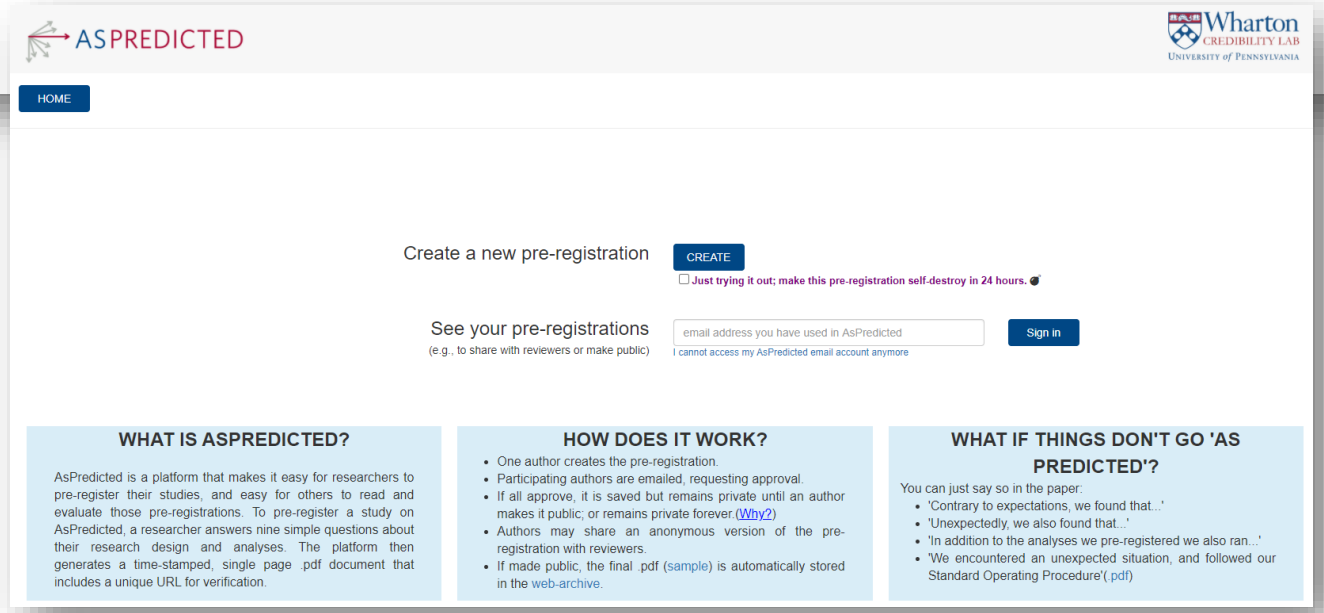
What is Preregistration?

When you preregister your research, you're simply specifying your research plan in advance of your study and submitting it to a registry.

Preregistration separates *hypothesis-generating* (exploratory) from *hypothesis-testing* (confirmatory) research. Both are important. But the same data cannot be used to generate *and* test a hypothesis, which can happen unintentionally and reduce the credibility of your results. Addressing this problem through planning improves the quality and transparency of your research. This helps you clearly report your study and helps others who may wish to build on it.

For additional insight and context, you can read The Preregistration Revolution. (preprint)

https://www.cos.io/initiatives/prereg



https://aspredicted.org/

Curate Science BETA

Search for articles or authors

Curated Replications (Table Vie

Curated List of Large-Scale Replication Efforts

Searchable table of N=1,127 replications of 168 effects from the cognitive and soci

Examples: **"RPP"** for *Reproducibility Project: Psychology*; **"ML1"** or **"ML3"** for *Many Labs 1* or 3; **"RRR"** for *Reg*
Psychology's Special Issue. For topical searches, try "priming", "anchoring", "gambler's fallacy", "love", "moral"

= open data; = open study materials; = preregistered study protocol; = associated replication collec

Search:

target.effect	orig.study.number	o.N	orig.effect.size	r.N	rep.effect.size	rep.study.num
(un)accomplished goal action effect	Koo & Fishbach (2008) Study 4	246	**η² = .041**	768...	**OR = .159**	Kidwell & Dodsc (2015)
achievement priming	Bargh et al. (2001) Study 1	78	**d = .70 ± .46**	106	**d = -.24 ± .40**	Harris, Rohrer, & Pashler (2013) S 1
achievement priming (5-minute delay...	Bargh et al. (2001) Study 3	72	**d = NR**	66	**d = -.03 ± .49**	Harris, Rohrer, & Pashler (2013) S 2
action priming boosts # of thoughts...	Albarracin et al. (2008) Study 7	98	**r = .21 ± .19**	109	**r = .16 ± .19**	Voracek & Sonnleitner (20

Recently Added

FILTER SORT BY

Filters: Replication CLEAR ALL

Replication 1 — Added June 25, 2020

Unlearning implicit social biases during sleep: A failure to replicate
GB Humiston & EJ Wamsley **(2019)**
PLOS ONE
pdf html

Replication 2 — Added March 11, 2020

Does honesty require time? Two preregistered direct replications of Experiment 2 of Shalvi, Eldar, and Bereby-Meyer (2012)
I Van der Cruyssen, J D'hondt, E Meijer, & B Verschuere **(2020)**
Psychological Science
pdf

Replication 1 — Added March 10, 2020

A replication attempt of "Does curiosity tempt indulgence"? The problem of hidden confounds (DR4)
J Simmons & L Nelson **(2020)**
preprint

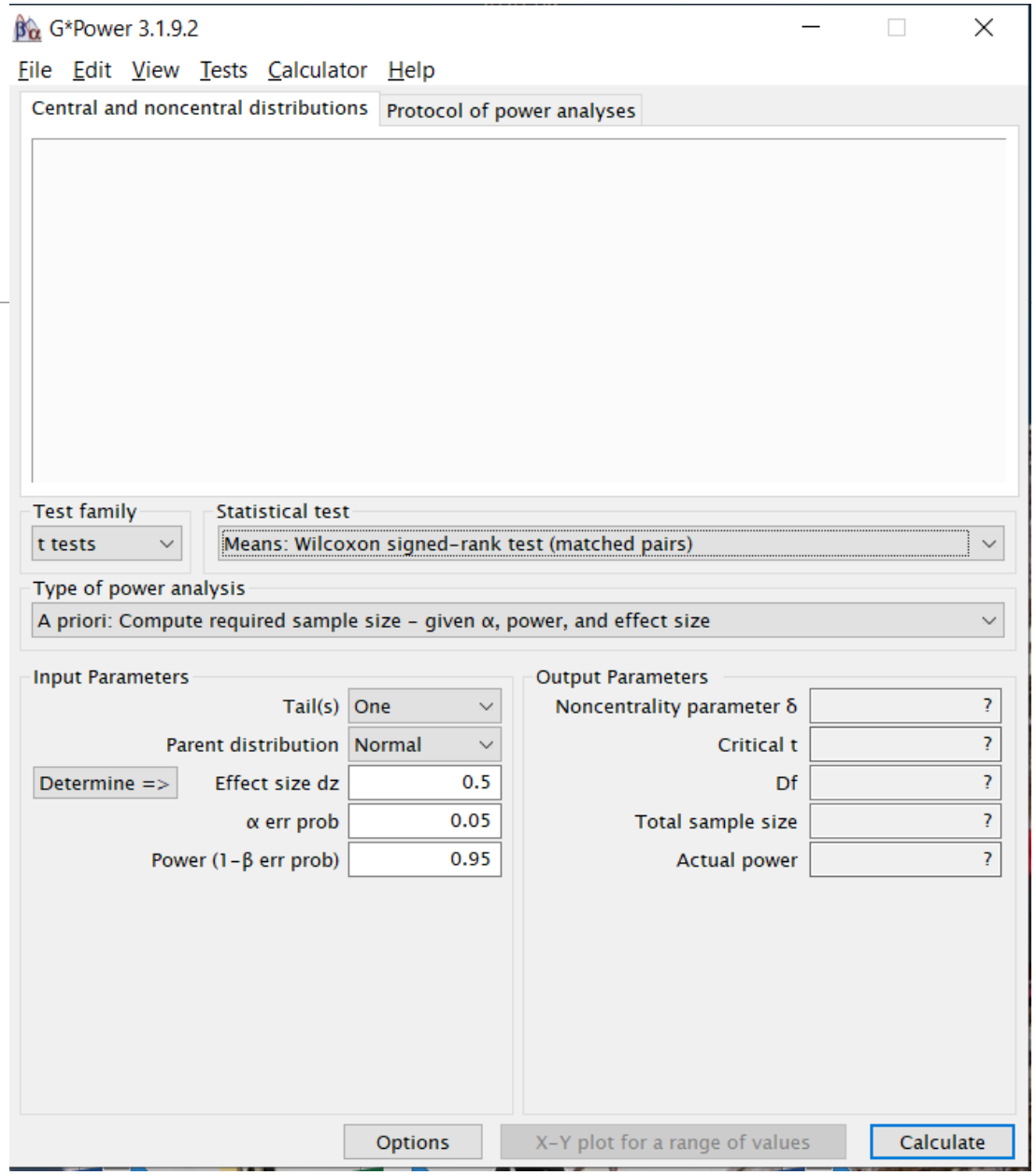
Replication 1 — Added March 3, 2020

Impact of ownership on liking and value: Replications and

Velikost vzorku

Používání větších datových souborů.

Pečlivá power-analýza.



A 21 Word Solution

Simmons, Joseph P. and Nelson, Leif D. and Simonsohn, Uri, A (2012). 21 Word Solution. *SSRN*. http://dx.doi.org/10.2139/ssrn.2160588

Choir: There is no need to wait for everyone to catch-up with your desire for a more transparent science. If *you* did not *p*-hack a finding, *say it*, and your results will be evaluated with the greater confidence they deserve.

If you determined sample size in advance, *say it.*
If you did not drop any variables, *say it.*
If you did not drop any conditions, *say it.*

These 21 words in a Methods section can *say it* succinctly:

"We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."

When needed, supplemental materials can be used to ensure the 21 words are accurate.
When sample size is not determined in advance, one could write:
"We added 50 observations after analyzing the first 100".

www.metascience2021.org

Psychology's crisis of confidence: Measurement edition
Diskutují: Jessica K. Flake, Eiko Fried, Andrea Helena Stoevenbelt
Moderátor: Esther Maassen
https://metascience2021.org/events/psychologys-crisis-of-confidence-measurement-edition/

Připomenutí