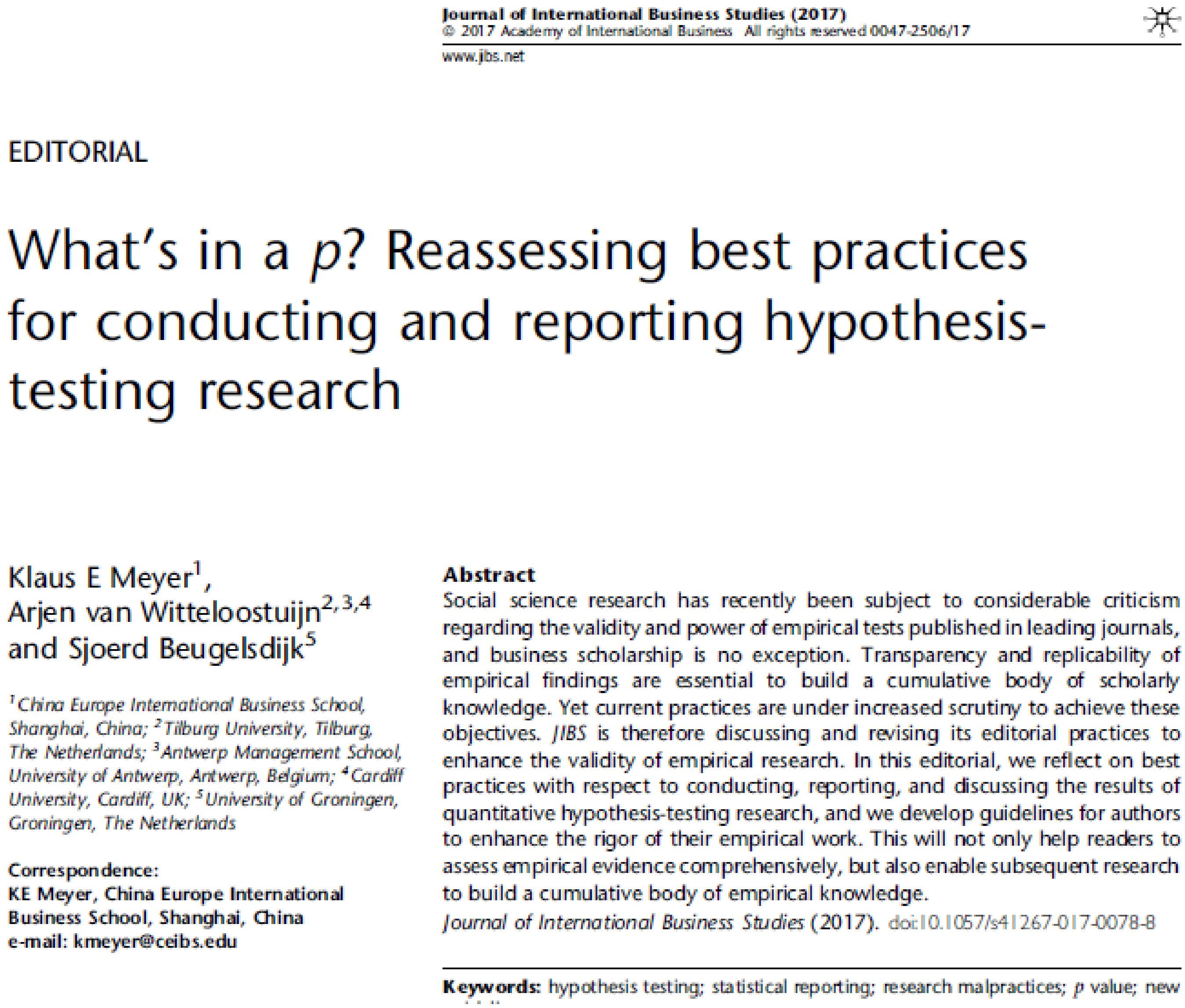
What's in a P?

Let's improve reporting practices for empirical research!

Klaus Meyer
Sjoerd Beugelsdijk

AIB 2017

Meyer, Klaus E., van Witteloostuijn, Arjen & Beugelsdijk, Sjoerd. 2017, *Journal of International Business Studies*, 48(5), 535-551.



Journal of International Business Studies (2017)
© 2017 Academy of International Business All rights reserved 0047-2506/17
www.jibs.net

EDITORIAL

What's in a *p*? Reassessing best practices for conducting and reporting hypothesis-testing research

Klaus E Meyer[1], Arjen van Witteloostuijn[2,3,4] and Sjoerd Beugelsdijk[5]

[1] China Europe International Business School, Shanghai, China; [2] Tilburg University, Tilburg, The Netherlands; [3] Antwerp Management School, University of Antwerp, Antwerp, Belgium; [4] Cardiff University, Cardiff, UK; [5] University of Groningen, Groningen, The Netherlands

Correspondence:
KE Meyer, China Europe International Business School, Shanghai, China
e-mail: kmeyer@ceibs.edu

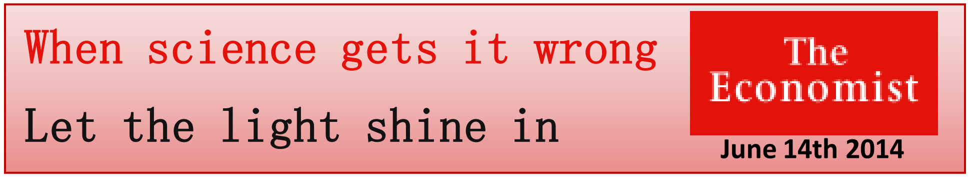
Abstract

Social science research has recently been subject to considerable criticism regarding the validity and power of empirical tests published in leading journals, and business scholarship is no exception. Transparency and replicability of empirical findings are essential to build a cumulative body of scholarly knowledge. Yet current practices are under increased scrutiny to achieve these objectives. *JIBS* is therefore discussing and revising its editorial practices to enhance the validity of empirical research. In this editorial, we reflect on best practices with respect to conducting, reporting, and discussing the results of quantitative hypothesis-testing research, and we develop guidelines for authors to enhance the rigor of their empirical work. This will not only help readers to assess empirical evidence comprehensively, but also enable subsequent research to build a cumulative body of empirical knowledge.

Journal of International Business Studies (2017). doi:10.1057/s41267-017-0078-8

Keywords: hypothesis testing; statistical reporting; research malpractices; *p* value; new

The Problem: Scientific journals are reporting far too many false positives!



When science gets it wrong

Let the light shine in

The Economist

June 14th 2014

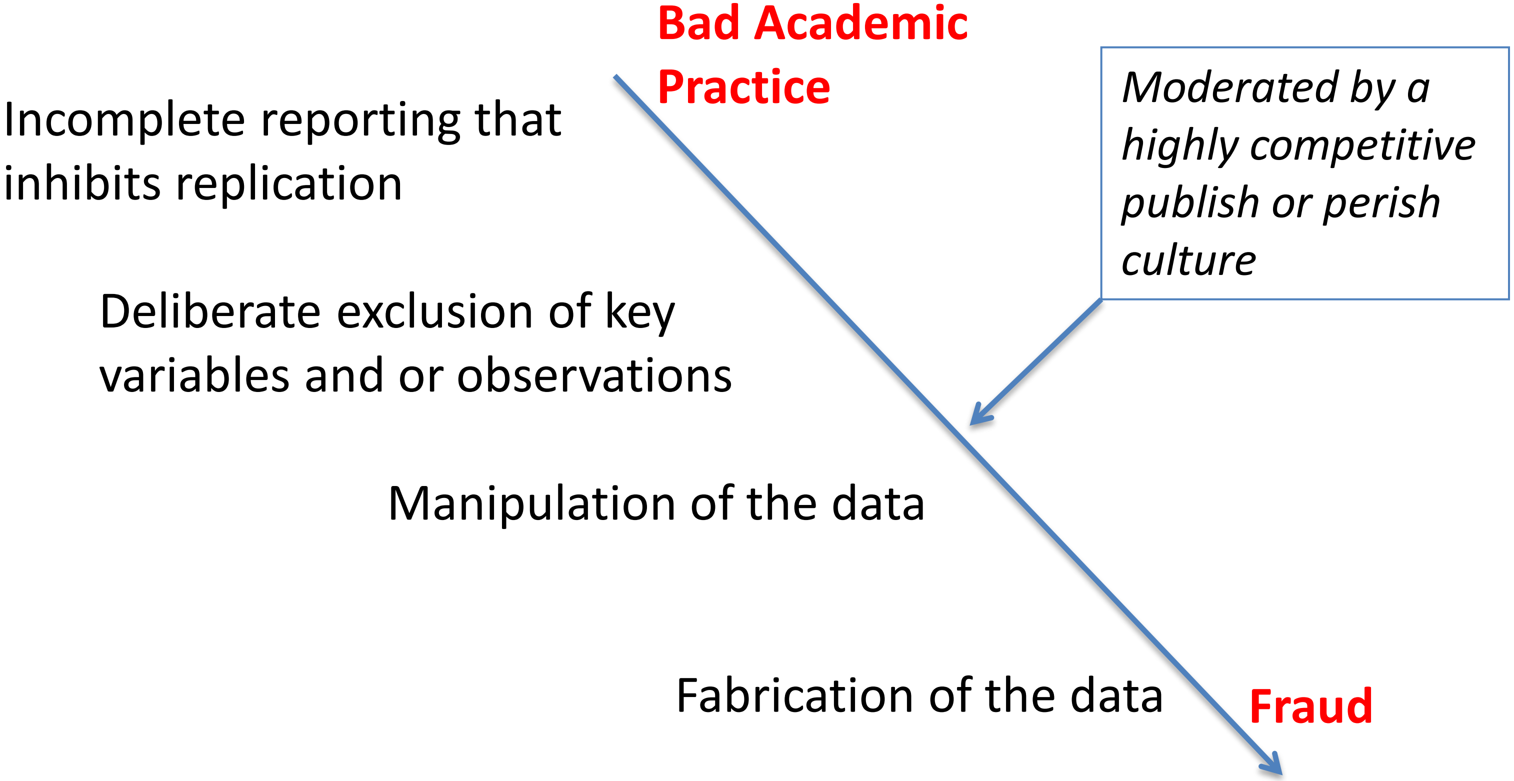
Publication-selection bias

Papers where hypotheses are “confirmed” have a higher probability of being accepted.

HARKING
Hypothesizing
After the Results
are Known

P-hacking
Manipulating a regression
until p-value crosses a
desired threshold

There is a slippery slope, but we are not talking about fraud cases



Bad Academic Practice

Incomplete reporting that inhibits replication

Moderated by a highly competitive publish or perish culture

Deliberate exclusion of key variables and or observations

Manipulation of the data

Fabrication of the data

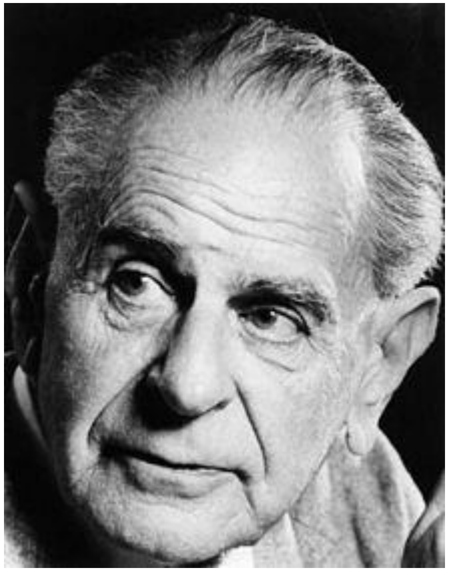
Fraud

Why do we test hypotheses in the first place?
To FALSIFY the statement that there is ‘no effect’

Deductive
Theory → Design Test → Construct Data
→ Conduct Test → Falsification

Inductive
Data → Interpretation → Theory *to be tested*

Unscientific
Data → Theory → Test on the same data



Karl Popper (1902-1994)

The American Statistical Association got seriously worried about the statistics practice!

... issued a formal statement partly reproduced in our appendix

Wasserstein, R.L., & Lazar, N.A. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose, *American Statistician*, 70(2): 129-133.

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

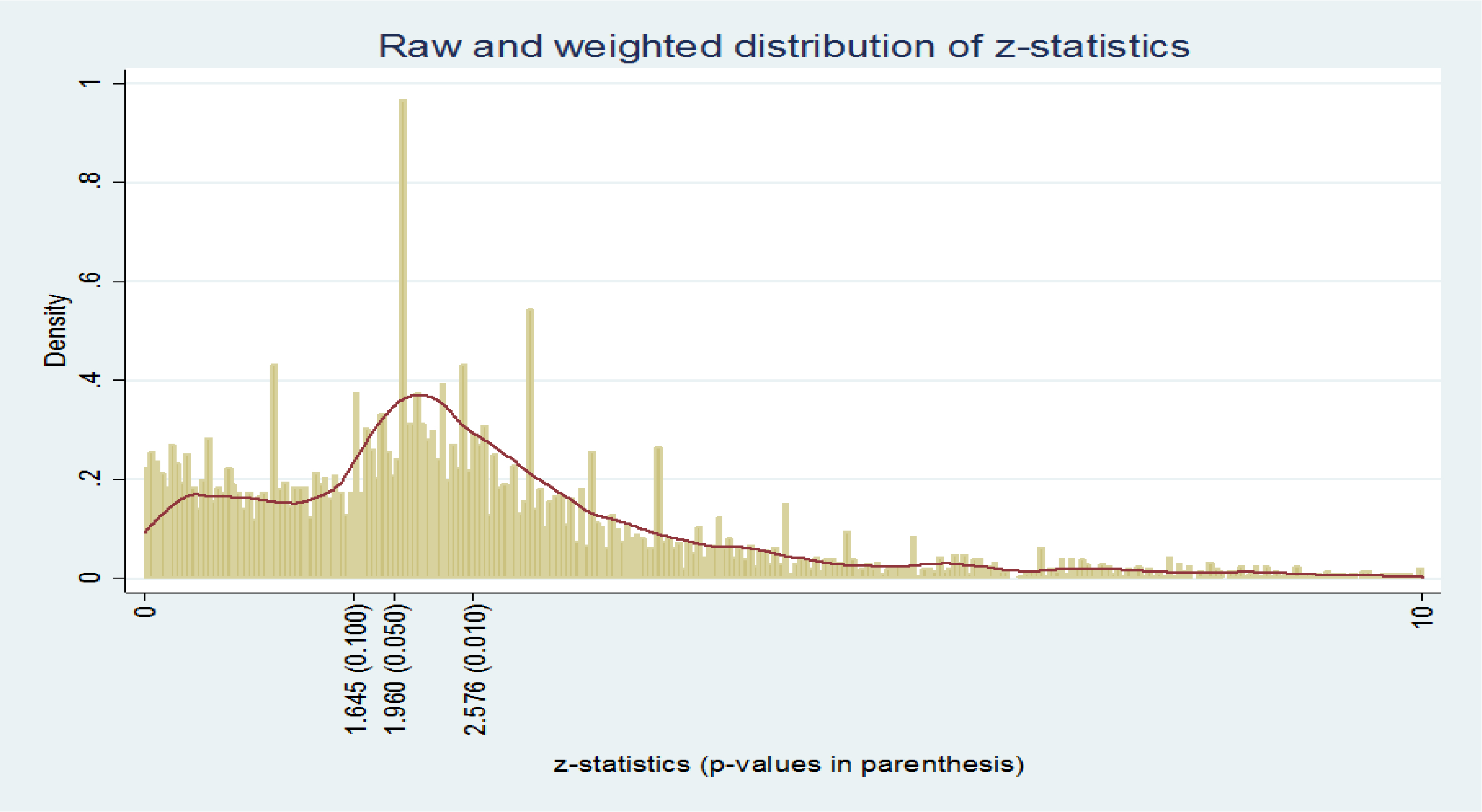
Where does the obsession with p<0.05 come from?

> *"The value for which p = .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not"*
>
> *(Ronald Fischer, 1925)*

A legacy of the pre-computer age:

Before the development of modern econometrics software, calculating p-values was quite complicated. Thus, econometrics textbooks provided appendix tables with threshold values for p that were used decide if a hypothesis was rejected or not.

Evidence: Also in IB we observe anomalous patterns in the levels of statistics reported.



Note: The graph shows the histogram as well as the kernel density plot of the weighted distribution of z-scores in all hypotheses testing articles published in *JIBS*, *OrgSci*, and *SMJ* in 2015 and 2016. (313 articles; 5,579 null hypothesis tests)

Similar patterns have been observed in many fields:

Economics:	Brodeur et al. 2016; *American Economic Journal - Applied Economics*
Psychology:	Ferguson & Heene, 2012; *Perspectives on Psych Science*
Political Science:	Gerber, Green & Nickerson, 2001; *Political Analysis*

Meta-analyses typically control for selection bias, i.e. working papers versus published articles (see e.g. Görg & Strobl, 2001)

Precision and transparency are an essential foundation for all scientific endeavors!

***Guideline 1**: At a basic level, all regression analyses should include, for each coefficient, standard errors (as well as mention the confidence intervals for the variable of interest) and, for each regression model, the number of observations as well as the* R^2 *statistics or equivalent.*

***Guideline 2**: Authors should refer to the actual* p*-value rather than the threshold* p*-value when assessing the evidence for and against their hypothesis.*

***Guideline 3**: Authors should not report asterisks to signal* p*-value thresholds.*

Explain whether the effect you are testing is important, i.e. ‘large’ relative to other influences

Guideline 4*: Reflections on effect sizes are included, reporting and discussing whether the effects (the coefficients and, if appropriate, marginal effects) are substantive in terms of the research question at hand.*

> ***“Ceteris paribus, a one standard deviation increase of cultural distance (which is comparable with a change in distance from, say, US-UK to, e.g., US-Italy) reduces the longevity of joint ventures with two to four years.”***

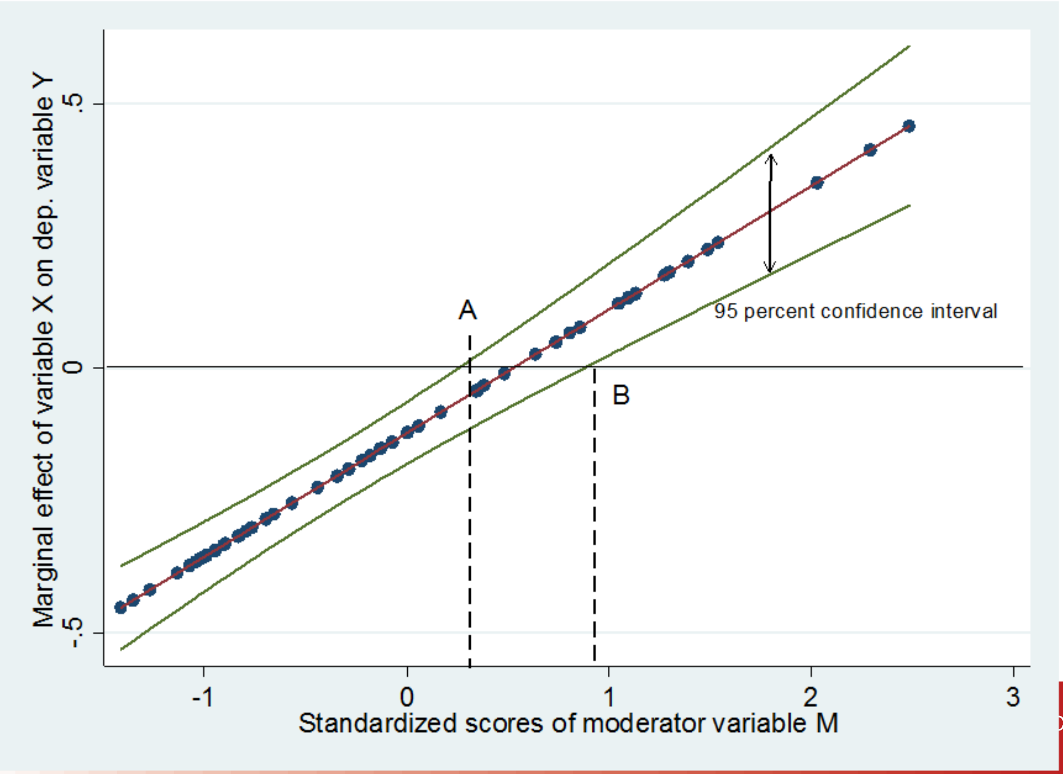
Suggested wording

➔ **Best practice on how this is to be done is evolving!**

Explain whether the effect you are testing is important, i.e. ‘large’ relative to other influences!

***Guideline 4a**: When discussing effect size, authors should take the confidence interval associated with the estimated coefficient into account as well as the minimum and maximum effect (not just one standard deviation above and below the mean), thus providing a range of the strength of a particular relationship. This may be done graphically for more complex models.*

Illustration of the effect size in an interaction model



Explain whether the effect you are testing is important, i.e. ‘large’ relative to other influences!

Guideline 4b*: When discussing effect sizes, where possible and relevant, authors should compare the range of the effect size of the variable of interest with other variables included in the regression model.*

> *“Ceteris paribus, a one standard deviation increase of cultural distance (which is comparable with a change in distance from, say, US-UK to, e.g., US-Italy) reduces the longevity of joint ventures with two to four years.* ***For comparability, the effect of a similar increase of one standard deviation of geographic distance results in a reduction of joint longevity by eight years.”***

Suggested wording

P-values are not the only interesting result in your regressions!

***Guideline 5**: Outlier observations are discussed carefully, especially when they have been eliminated from the sample (e.g., through technical practices such as ‘winzorizing).*

***Guideline 6**: Null and negative findings are equally interesting as are positives, and hence are honestly reported, including a discussion of what this implies for theory.*

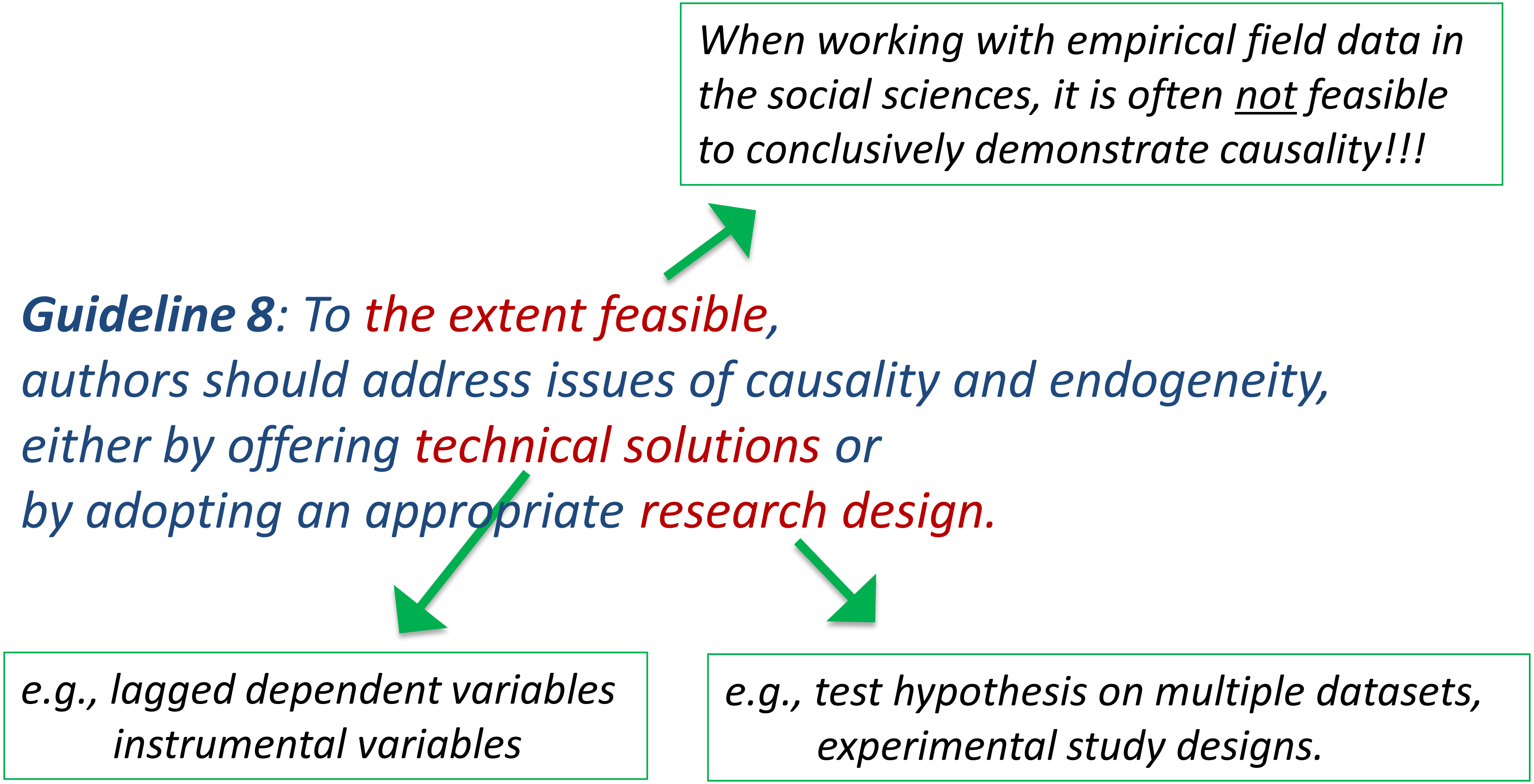
Don’t jump from a statistical association to a statement about causality!

Guideline 7*: In the absence of a clear strategy designed explicitly to identify causes and effects, authors should be careful in using terminology suggesting causal relationships between variables of interest, and accordingly adjust their language in the wording of the hypotheses and in the discussion of the empirical results.*

DO	“association”, “relation”
DON’T	“determinant” and “effect” or “affect”

… for the results section of the paper!

Don’t jump from a statistical association to a statement about causality! (2)



When working with empirical field data in the social sciences, it is often not feasible to conclusively demonstrate causality!!!

***Guideline 8**: To the extent feasible, authors should address issues of causality and endogeneity, either by offering technical solutions or by adopting an appropriate research design.*

e.g., lagged dependent variables instrumental variables

e.g., test hypothesis on multiple datasets, experimental study designs.

But: these “solutions” have their own methodological challenges that need to be properly addressed!

Robustness tests enhance reviewers confidence in your work!

***Guideline 9**: Authors are expected to conduct a variety of robustness tests to show that the significant finding is not due to an idiosyncrasy of the selected empirical measures, model specifications and/or estimation strategy.*

For example:

- **alternative proxies of focal constructs**, especially for not directly measurable constructs
- **alternative sets of control variables**, especially when correlation is present
- **alternative functional forms** of the regression models, especially for non-linear, moderating and mediating effects

We need to develop stronger research practices for inductive theory building based on data!

***Guideline 10**: HARKing is a research malpractice. Theory developed by interpreting empirical phenomena or results should be reported as such (for example, in the discussion section).*

Post-hoc theorizing

Phenomenon-driven research

Hollenbeck, J. R. & Wright, P. M. 2017. Harking, S-harking, and T-harking. *Journal of Management*, 43(1): 5-18.

On Causality and Endogeneity

- Certo, S.T., Busenbark, J.R., Woo, H.S. & Semadeni, M. 2016. Sample selection bias and Heckmann models in strategic management research, *Strategic Management Journal*, 37(13): 2639-2657.
- Reeb, D., Sakakibara, M, & Mahmood, I.P. 2012, Endogeneity in IB research, *Journal of International Business Studies*, 43(3): 211-218.
- Thomas, D.C., Cuervo-Cazurra, A. & Brannen, M.Y. 2011. Explaining theoretical relationships in IB research: Focusing on the arrows not the boxes, *Journal of International Business Studies*, 42(9): 1073-1078.

On the Appropriate Empirical Model

- Andersson, U., Cuervo-Cazurra, A., & Nielsen, B.B. 2014. Explaining interaction effects within and across levels of analysis, *Journal of International Business Studies*, 45(9): 1063-1071.
- Cortina, J.M., Köhler, T., & Nielsen, B.B. 2015, Restriction of variance interaction effects and their importance for international business, *Journal of International Business Studies*, 46(8): 879-885.
- Haans, R.F.P., Pieters, C., & He, Z.L. 2016, Thinking about U: Theorizing and testing U- and inverted U-shaped Relationships in Strategy Research, *Strategic Management Journal*, 37(7): 1177-1196.
- Meyer, K.E. 2009. Motivating, testing, and publishing curvilinear effects in management research, *Asia Pacific Journal of Management*, 26(2): 187-193.
- Petersen, M.F., Arregle, J.L., & Martin, X. 2012. Multilevel models in IB research, *Journal of International Business Studies*, 43(5): 451-457.

Best Practice is Evolving!

We try to push the evolution forward, but our guidelines should not be seen as definitive solution to the methodological challenges faced by the social sciences.

Meyer, K.E., van Witteloostuijn, A. & Beugelsdijk, S. 2017,
Journal of International Business Studies, 48(5), 535-551.