Open Science in Business Economics

Joachim Gassen
Humboldt-Universität zu Berlin
TRR 266 "Accounting for Transparency"

March 17, 2020

SFB/Transregio 266
ACCOUNTING FOR TRANSPARENCY

Agenda

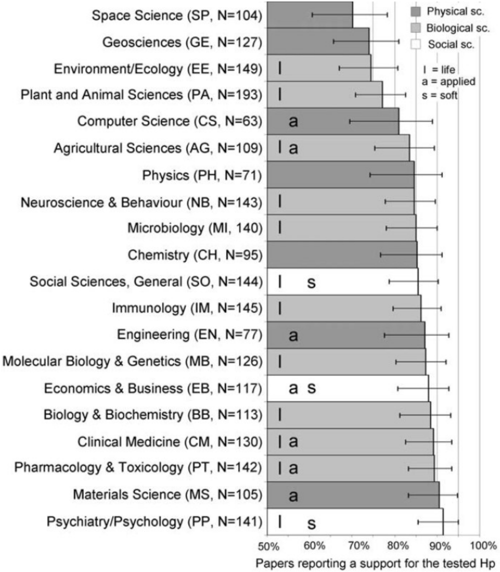
- What is open science and why should we care?
- Introducing our case: The Preston curve
- Sorry but your code does not run: Research containers
- Making data FAIR: Issues in economics
- Working around unFAIR data: The 'ExPanDaR' package
- Increase research transparency: The 'rdfanalysis' package

What is open science and why should we care?

Open Science

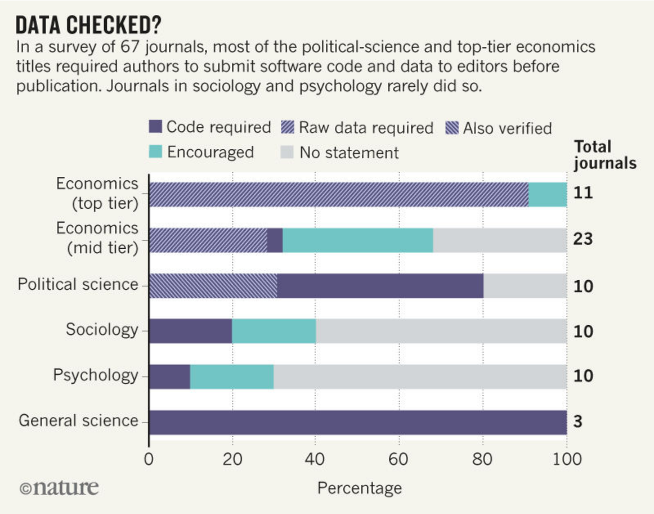
> Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.
>
> — FOSTER, https://www.fosteropenscience.eu/

Too good to be true?



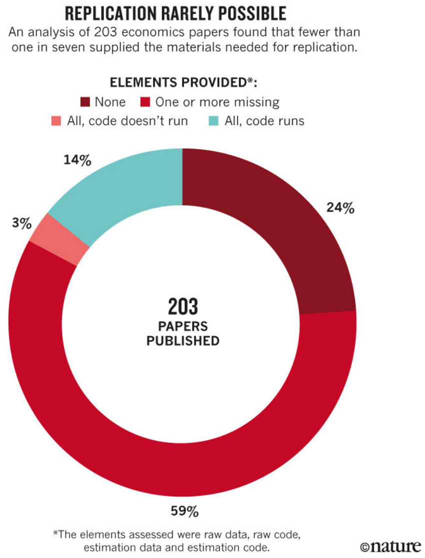
Fanelli: "Positive" Results Increase Down the Hierarchy of the Sciences (PloS, 2010)

Data and code repositories are on the rise . . .



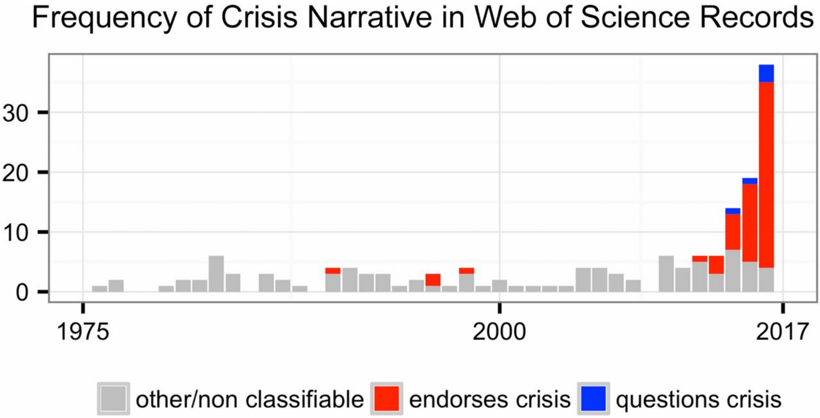
Gertler, Galiani and Romero (Nature, 2018)

. . . but yet fail to guarantee reproducible results



Gertler, Galiani and Romero (Nature, 2018)

But: The narrative of a "research crisis" ...



Fanelli (PNAS, 2018)

. . . might be exaggerated

> The new "science is in crisis" narrative is not only empirically unsupported, but also quite obviously counterproductive. Instead of inspiring younger generations to do more and better science, it might foster in them cynicism and indifference. Instead of inviting greater respect for and investment in research, it risks discrediting the value of evidence and feeding antiscientific agendas.
>
> . . .
>
> Therefore, contemporary science could be more accurately portrayed as facing "new opportunities and challenges" or even a "revolution". Efforts to promote transparency and reproducibility would find complete justification in such a narrative of transformation and empowerment, a narrative that is not only more compelling and inspiring than that of a crisis, but also better supported by evidence.
>
> — Fanelli (PNAS, 2018)

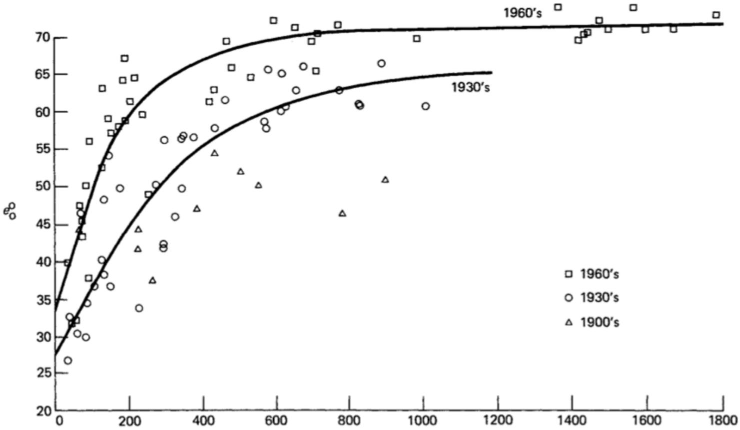
Instead: Open Collaboration is a success story!

- Wikipedia
- GNU/Linux
- Git(hub)
- R/RStudio
- Python
- ...

Introducing our case:
The Preston curve

Life expectancy is associated with GDP per capita

Scatter-diagram of relations between life expectancy at birth (e_0^o) and national income per head for nations in the 1900s, 1930s, and 1960s.



National income per head, 1963 U.S. dollars.

Preston (1975): The Changing Relation between Mortality and level of Economic Development, Population Studies (29): 235.

Descriptive Statistics

	N	Mean	Std. dev.	Min.	25 %	Median	75 %	Max.
Life expectancy	4,144	68.0	10.0	26.2	61.3	70.6	75.4	84.3
GDP per capita	4,144	11.8	17.3	0.2	1.3	3.9	14.0	112.0
Years of schooling	4,144	7.8	3.0	0.3	5.6	8.4	10.3	13.2
Unemplyoment rate	4,144	8.2	6.4	0.2	3.6	6.5	10.8	37.9

Note: The data is obtained from the World Bank and the Wittgenstein Center. The sample covers 179 countries and the period 1991 to 2015. *GDP per captia* are in constant 2010 thousand U.S. dollars.

Regression Results

	Dependent variable: *Life expectancy*			
	(1)	(2)	(3)	(4)
ln(*GDP per capita*)	5.232^{***}	3.723^{***}	3.754^{***}	1.537^{***}
	(0.062)	(0.082)	(0.034)	(0.501)
ln(*Years of schooling*)		6.456^{***}	6.036^{***}	6.897^{***}
		(0.240)	(0.080)	(1.696)
ln(*Unemployment rate*)		-0.976^{***}	-0.896^{***}	0.027
		(0.108)	(0.123)	(0.354)
Constant	24.296^{***}	26.064^{***}		
	(0.526)	(0.497)		
Fixed effects	None	None	Year	Country, Year
Std. errors clustered	No	No	Year	Country, Year
Observations	4,144	4,144	4,144	4,144
R^2	0.632	0.688	0.696	0.968
Adjusted R^2	0.632	0.687	0.694	0.966

Note: The dependent variable is the average life expectancy at birth in years. OLS coefficients are reported together with standard errors in parentheses. $^{*}/^{**}/^{***}$ indicate two-sided significance levels of 10/5/1 %, respectively.

Sorry but your code does not run: Research containers

What is the problem?

- All code is based on a development environment (e.g., your computer)
- Not everybody has access to your development environment (matter-of-factly most likely nobody has)
- Designing projects so that they can be easily ported across environments is hard. Too hard for most of us.
- The easier way: Develop your project and its code in a standardized environment. Ship that environment with your code.

Meet Docker

- Take a look at this file: https://github.com/joachim-gassen/opsci_talk/blob/master/docker/Dockerfile

Making data FAIR: Issues in economics

What makes data FAIR?

- Data is easily findable both by experts in the field and by researchers from other domains
- Data is accessible for interested researchers meaning that they can access, understand and use the data
- Data is inter-operable so that it can be merged with other data sources
- Data is reusable so that it can be used for other research projects besides the research project for which they were initially collected

Why is it hard to establish FAIR data in economics?

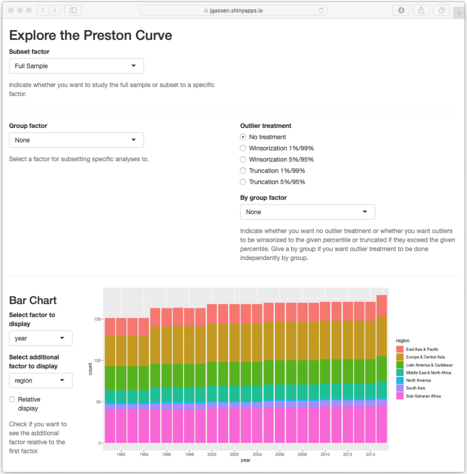
- Many data are commercialized, especially in the area of capital markets
- Survey data often contains personal information
- Experimental data is collected based on informed consent from participants that did not include later sharing of the data
- Data is stored on author's web pages or similar, with very little meta data, leading to data that only experts in the field are able to find and understand
- Lack of clear identifiers for main units of analyses make it inherently hard to merge data from different sources
- Lack of information on data generating processes reduces re-usability

Working around unFAIR data: The ‘ExPanDaR’ package

The idea

- When data cannot be shared, you can still provide an interface to the data so that others can explore it
- That way you enable users to assess the robustness of empirical evidence without providing them with access to the underlying data
- Besides that, the 'ExPanDaR' package also provides a toolbox for researchers to explore data on the fly, allowing them to download R notebook code that reflects their analysis

Meet the ExPanD app



https://jgassen.shinyapps.io/preston

Increase research transparency:
The ‘rdfanalysis’ package

Empirical research requires making choices

> What troubles me about using opinions is their whimsical nature. Some mornings when I arise, I have the opinion that Raisin Bran is better than eggs. By the time I get to the kitchen, I may well decide on eggs, or oatmeal. I usually do recall that the sixteenth president distinguished himself. Sometimes I think he was Jackson; often I think he was Lincoln.
>
> A data analysis is similar. Sometimes I take the error terms to be correlated, sometimes uncorrelated; sometimes normal and sometimes nonnormal; sometimes I include observations from the decade of the fifties, sometimes I exclude them; sometimes the equation is linear and sometimes nonlinear; sometimes I control for variable *z*, sometimes I don't. Does it depend on what I had for breakfast?

Leamer (AER 1983: 37f.)

Slightly more systematic . . .

TABLE 1 | Checklist for different types of researcher degrees of freedom in the planning, executing, analyzing, and reporting of psychological studies.

Code	Related	Type of degrees of freedom
Hypothesizing		
T1	R6	Conducting explorative research without any hypothesis
T2		Studying a vague hypothesis that fails to specify the direction of the effect
Design		
D1	A8	Creating multiple manipulated independent variables and conditions
D2	A10	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators
D3	A5	Measuring the same dependent variable in several alternative ways
D4	A7	Measuring additional constructs that could potentially act as primary outcomes
D5	A12	Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)
D6		Failing to conduct a well-founded power analysis
D7	C4	Failing to specify the sampling plan and allowing for running (multiple) small studies
Collection		
C1		Failing to randomly assign participants to conditions
C2		Insufficient blinding of participants and/or experimenters
C3		Correcting, coding, or discarding data during data collection in a non-blinded manner
C4	D7	Determining the data collection stopping rule on the basis of desired results or intermediate significance testing
Analyses		
A1		Choosing between different options of dealing with incomplete or missing data on *ad hoc* grounds
A2		Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an *ad hoc* manner
A3		Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner
A4		Deciding on how to deal with outliers in an *ad hoc* manner
A5	D3	Selecting the dependent variable out of several alternative measures of the same construct
A6		Trying out different ways to score the chosen primary dependent variable
A7	D4	Selecting another construct as the primary outcome
A8	D1	Selecting independent variables out of a set of manipulated independent variables
A9	D1	Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)
A10	D2	Choosing to include different measured variables as covariates, independent variables, mediators, or moderators
A11		Operationalizing non-manipulated independent variables in different ways
A12	D5	Using alternative inclusion and exclusion criteria got selecting participants in analyses
A13		Choosing between different statistical models
A14		Choosing the estimation method, software package, and computation of SEs
A15		Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)
Reporting		
R1		Failing to assure reproducibility (verifying the data collection and data analysis)
R2		Failing to enable replication (re-running of the study)
R3		Failing to mention, misrepresenting, or misidentifying the study preregistration
R4		Failing to report so-called "failed studies" that were originally deemed relevant to the research question
R5		Misreporting results and *p*-values
R6	T1	Presenting exploratory analyses as confirmatory (HARKing)

Wicherts et al. (Frontiers in Psychology 2016: 3)

Objective of the `rdfanalysis` package

Development of a statistical computing framework for a "multiverse analysis" (Steeger et al., Persp on Psych Sci 2016) that

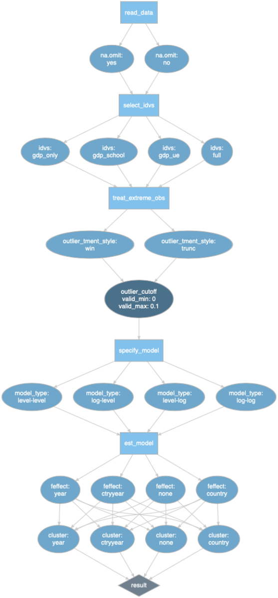
- supports research design development separate from data,
- facilitates a priori power analysis,
- promotes unit testing,
- generates well documented and easily portable code,
- makes researcher degrees of freedom (Simmons et al., Psyc Science 2011) explicit in code, and
- allows for rigorous robustness checking by exhausting all these degrees of freedom algorithmically.

Note: Code based walk-through on what follows: https://joachim-gassen.github.io/rdfanalysis/articles/analyzing_rdf.html

Workflow

1. Develop your `design` by specifying the `result` that you want to estimate and the required `steps` to do that.
2. The package generates a code skeleton based on your `design`.
3. Implement each `step` starting from this skeleton. Assign positive weights to all design choices that you consider to be supported best by theory ex ante.
4. Use package functions to document and test your code for internal consistency.
5. Optional: Run code on simulated data to test its computational correctness and assess the power of your `design`.
6. Run `design` with real data and generate a weighted estimate of the `result`.
7. Exhaust `design` across all `protocols` and use the visualization tools of the package to assess the robustness of your `result` across all researcher degrees of freedom.

The case steps and their discrete/continuous choices



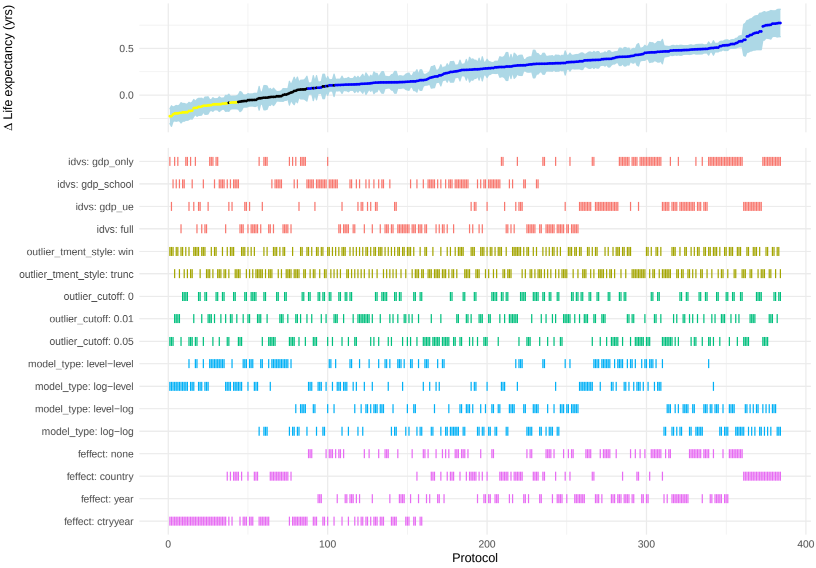
Estimate the effect size for the weighted protocols

na.omit	idvs	outlier_tment_style	outlier_cutoff	model_type	feffect	cluster	weight	est	lb	ub
yes	full	win	0.00	level-log	ctryyear	ctryyear	0.05	0.146	0.048	0.245
yes	full	trunc	0.00	level-log	ctryyear	ctryyear	0.05	0.146	0.048	0.245
yes	full	win	0.01	level-log	ctryyear	ctryyear	0.10	0.127	0.042	0.211
yes	full	trunc	0.01	level-log	ctryyear	ctryyear	0.10	0.137	0.052	0.221
yes	full	win	0.05	level-log	ctryyear	ctryyear	0.10	0.103	0.023	0.183
yes	full	trunc	0.05	level-log	ctryyear	ctryyear	0.10	0.101	0.017	0.184
yes	full	win	0.00	log-log	ctryyear	ctryyear	0.05	0.168	0.052	0.284
yes	full	trunc	0.00	log-log	ctryyear	ctryyear	0.05	0.168	0.052	0.284
yes	full	win	0.01	log-log	ctryyear	ctryyear	0.10	0.138	0.047	0.229
yes	full	trunc	0.01	log-log	ctryyear	ctryyear	0.10	0.148	0.065	0.232
yes	full	win	0.05	log-log	ctryyear	ctryyear	0.10	0.096	0.012	0.181
yes	full	trunc	0.05	log-log	ctryyear	ctryyear	0.10	0.108	0.032	0.183

And the weighted average estimate is:

est	lb	ub	n
0.127	0.039	0.215	12

Some but not all of the researcher degrees of freedom



And an interactive variant of it

https://jgassen.shinyapps.io/shiny_rdf_spec_curve/

For further info and code

- Code for this talk:
  https://github.com/joachim-gassen/opsci_talk
- 'ExPanDaR' package:
  https://joachim-gassen.github.io/ExPanDaR
- 'rdfanalysis' package:
  https://joachim-gassen.github.io/rdfanalysis
- Online course for open science methods:
  https://github.com/joachim-gassen/sposm
- Blog on R and Open Science https://joachim-gassen.github.io
- Twitter: https://twitter.com/JoachimGassen