Should we retire null hypothesis significance testing in (some) social policy research?

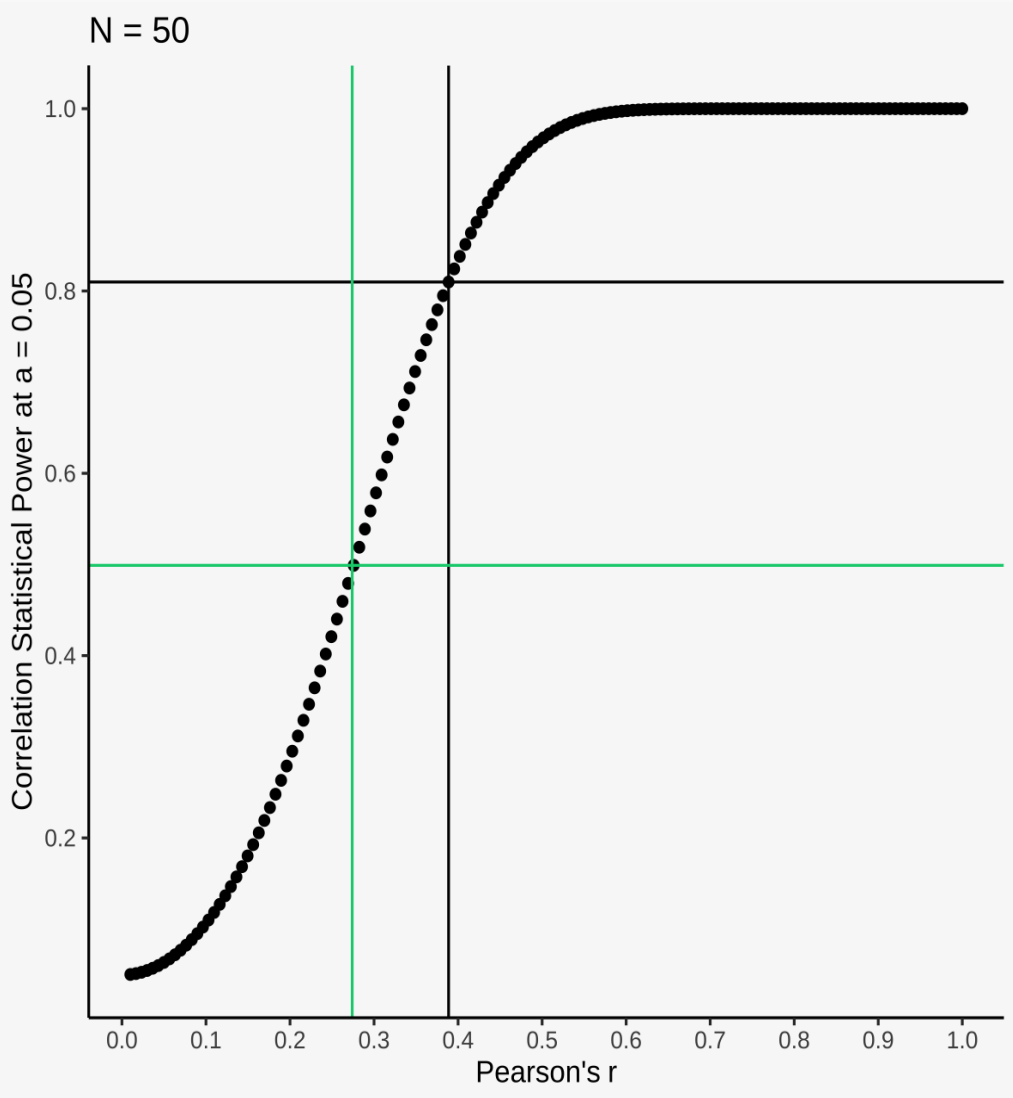
Dr. Calum Webb
Sheffield Methods Institute, the University of Sheffield
c.j.webb@sheffield.ac.uk

Working Paper Available on SPA2023 Conference Platform | Code published on Github

What is the effect of family support services spending on rates of children in the care system?

	B	S.E.	t	p	
(Intercept)	0.00	0.13	0.00	1.00	
Child poverty rate z	0.40	0.13	3.15	0.00	*
Inequality (Gini) z	0.30	0.13	2.36	0.02	*
Spending z	-0.15	0.13	-1.18	0.24	
Staffing z	0.05	0.13	0.39	0.70	

N = 50



- Only effects with a correlation around **±0.4 or greater** will be **detectable** ("p<0.05") **at least 80% of the time** with a sample size of N = 50.
- Only effects from a sample with a correlation **above around ±0.275** will be **detectable** ("p<0.05") **at all** with a sample size of N = 50 (Lakens, 2017).
- If we could randomly re-sample the data 1,000 times, the spending effect would **only be statistically significant in 23% of the samples.**

Our study is underpowered. What can we do?

The textbook approach:

- Do a **power analysis** beforehand and repeat the study with a larger sample and adequate power to detect the small/modest effect.
- Only one problem...

We can't collect a bigger 'random' sample: we have data from all 50 administrative units (states, local authorities, countries, etc.). This is an apparent population (Berk, Western & Weiss, 1995a)

Some suggestions...

Bin statistical inference (e.g. our data are the population, no inference is needed [Berk, Western & Weiss, 1995a])

Change the ecological unit (e.g. individual level survey, neighbourhoods rather than local authorities or states, etc.)

Pool data over several years (e.g. data for states/LAs between 2015-2021) [Bollen, 1995]

Use meta-analysis (Several non-significant results can illuminate a significant one [Bollen, 1995])

Bin null hypothesis significance testing

Bin statistical inference
A theoretical/philosophical argument

A brief recap of frequentist NHST:

- There is a true population parameter.
- Our estimates of this parameter vary because our samples of data vary.
- The aim of statistical inference is to generalise from our sample to a population (e.g. a random sample of 500 people to the entire population of the UK)
- We test the probability that the estimate might be observed from a random sample of size N if the true population parameter is equal to 0.

Is this approach to statistical inference meaningful when our data are the complete population?

> "If the data (the apparent population) are a census of a population and interest lies in a population value, then descriptive statistics are all that is needed. No inference is needed since nothing is unknown." (Berk, Western & Weiss, 1995b: 483)

But...

- The 'population' could refer to a 'superpopulation' — the underlying data generating process.
- Probably undesirable to have no measure of statistical uncertainty, especially if estimates are 'noisy'

Change the ecological unit
At risk of fallacy

- Sometimes appropriate, e.g. unemployment rates might be substituted with individual probability of being unemployed.
- Often not: e.g. how to meaningfully disaggregate state or LA spending to the individual or neighbourhood level.
- Uncomfortable implications for social policy/public administration research: macro- national or meso- policy and socioeconomic conditions don't matter because they're not convenient for our statistical practice?

Change the ecological unit
At risk of fallacy

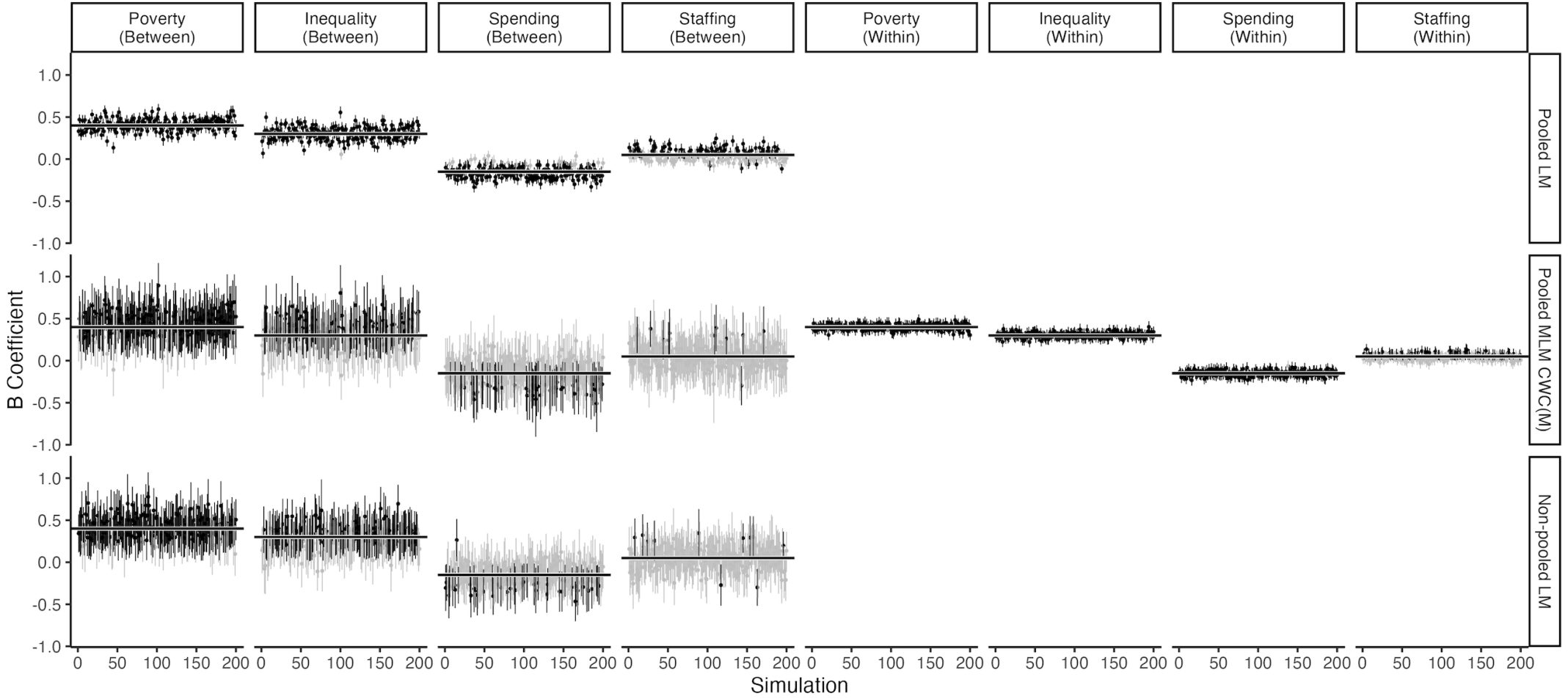
- Sometimes appropriate, e.g. unemployment rates might be substituted with individual probability of being unemployed.
- Often not: e.g. how to meaningfully disaggregate state or LA spending to the individual or neighbourhood level.
- Uncomfortable implications for social policy/public administration research: macro- national or meso- policy and socioeconomic conditions don't matter because they're not convenient for our statistical practice?

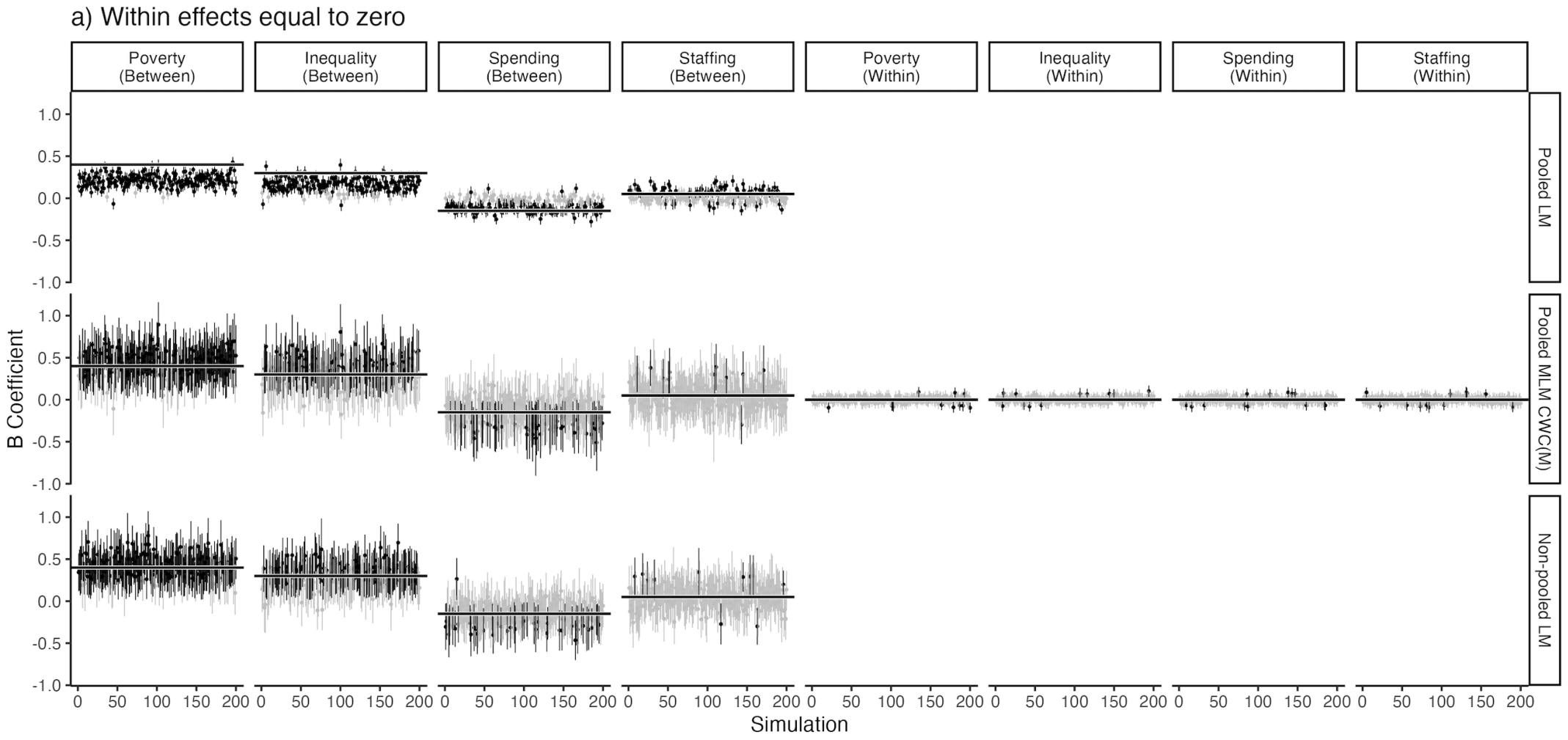
Pool data over several years

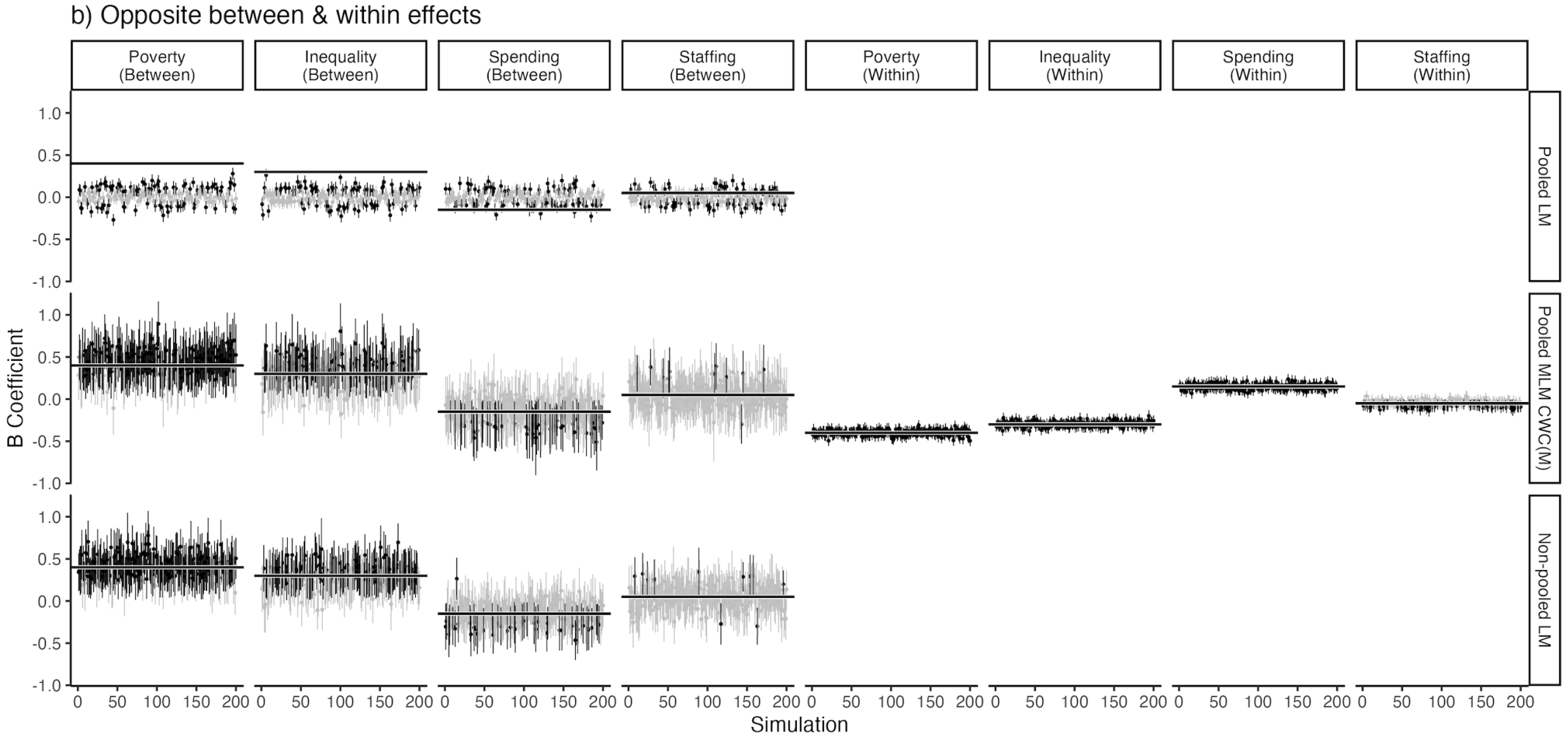
The problem with pooling

- Pool panel data for the same countries/LAs/states, etc. over multiple years to increase the sample size.
- Ignores the multilevel structure of the data...

c) Within and between effects are exactly the same



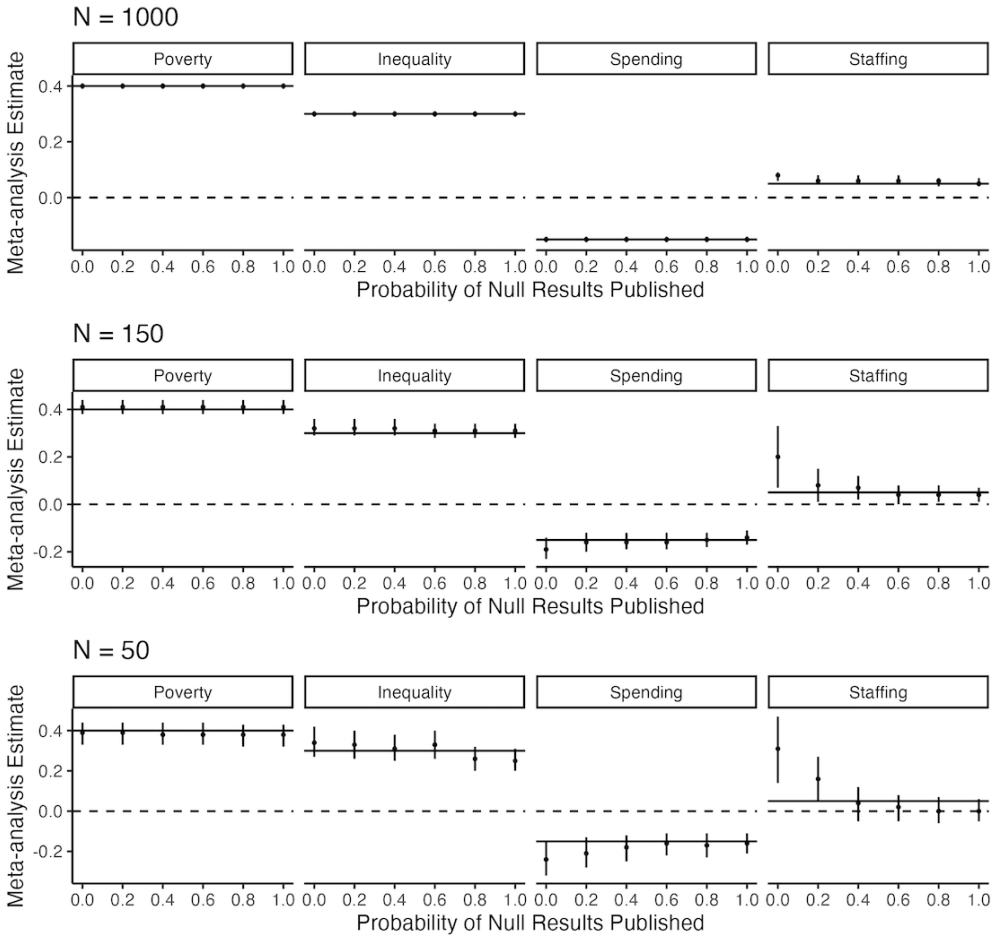

a) Within effects equal to zero
Poverty (Between)
Inequality (Between)
Spending (Between)
Staffing (Between)
Poverty (Within)
Inequality (Within)
Spending (Within)
Staffing (Within)
Pooled LM
Pooled MLM CWC(M)
Non-pooled LM
B Coefficient
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
200
Simulation


b) Opposite between & within effects
Poverty (Between)
Inequality (Between)
Spending (Between)
Staffing (Between)
Poverty (Within)
Inequality (Within)
Spending (Within)
Staffing (Within)
Pooled LM
Pooled MLM CWC(M)
Non-pooled LM
B Coefficient
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
200
Simulation

Meta-analysis

In theory and practice...

- It is true that we could repeat a study every year for 20 years and even if all of the coefficients from each individual study were non-significant a meta-analysis of all of these studies would be likely to be significant.
- It is also true that doing this would be incredibly boring, unlikely to attract funding, and unlikely to get written up and published...



- This would lead to considerable publication bias which will

At this point, you may be wondering...

Is this one of those depressing methods papers where someone points out everything that is wrong with what we are doing and offers no practical alternatives?

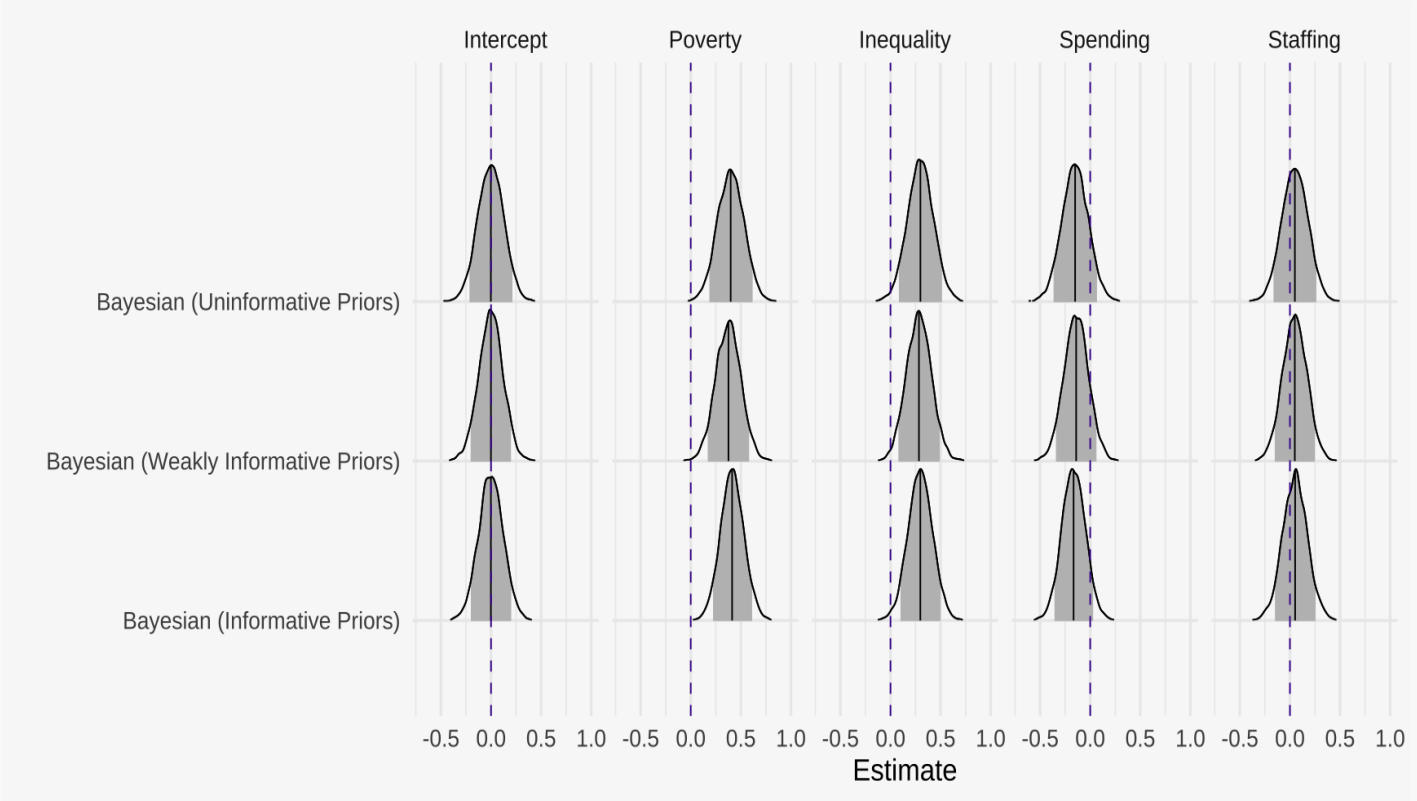
Fair point.

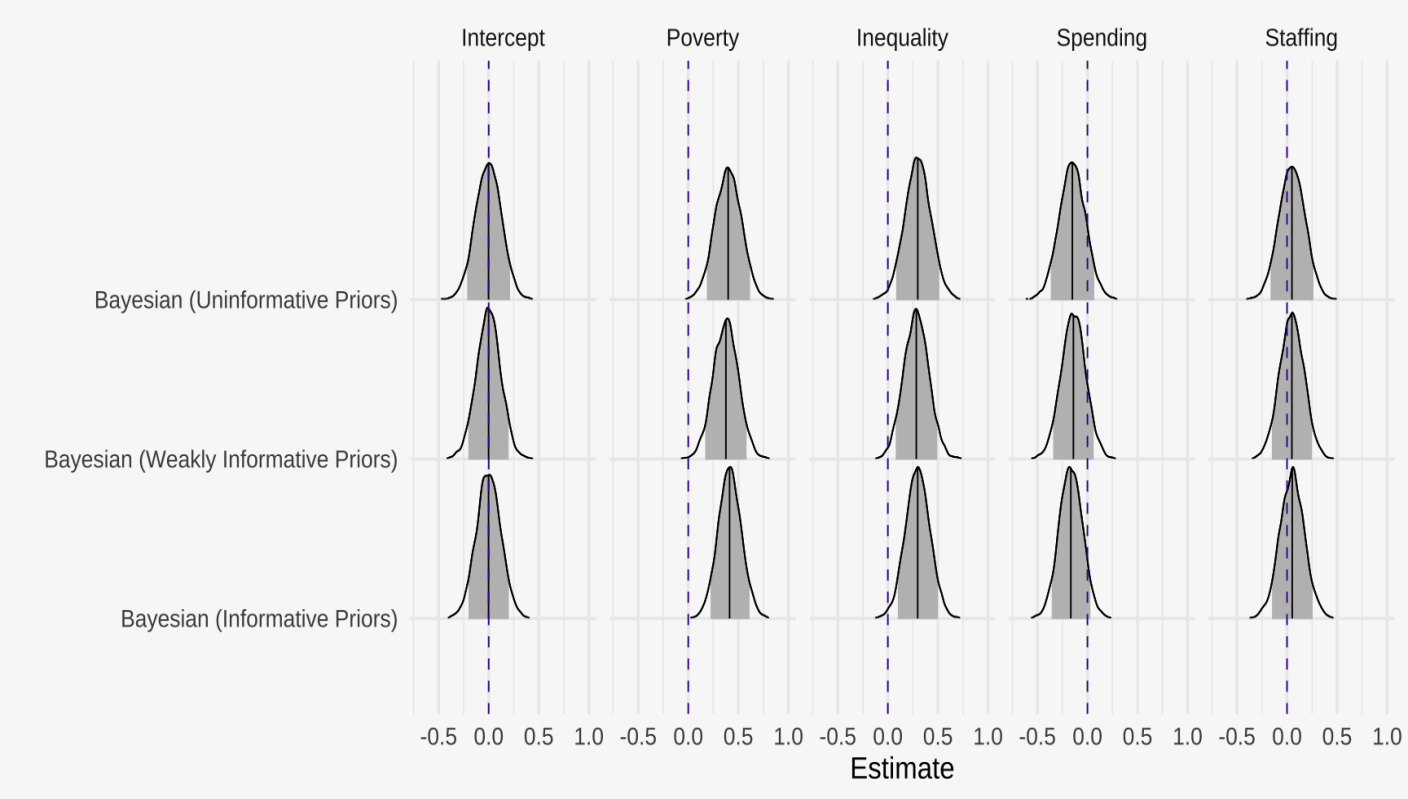
But no, it is not!

An alternative: Describing, rather than deciding, uncertainty

From a pragmatic standpoint, the root of this problem is the use of a decision threshold (i.e. the Null Hypothesis Significance Test).

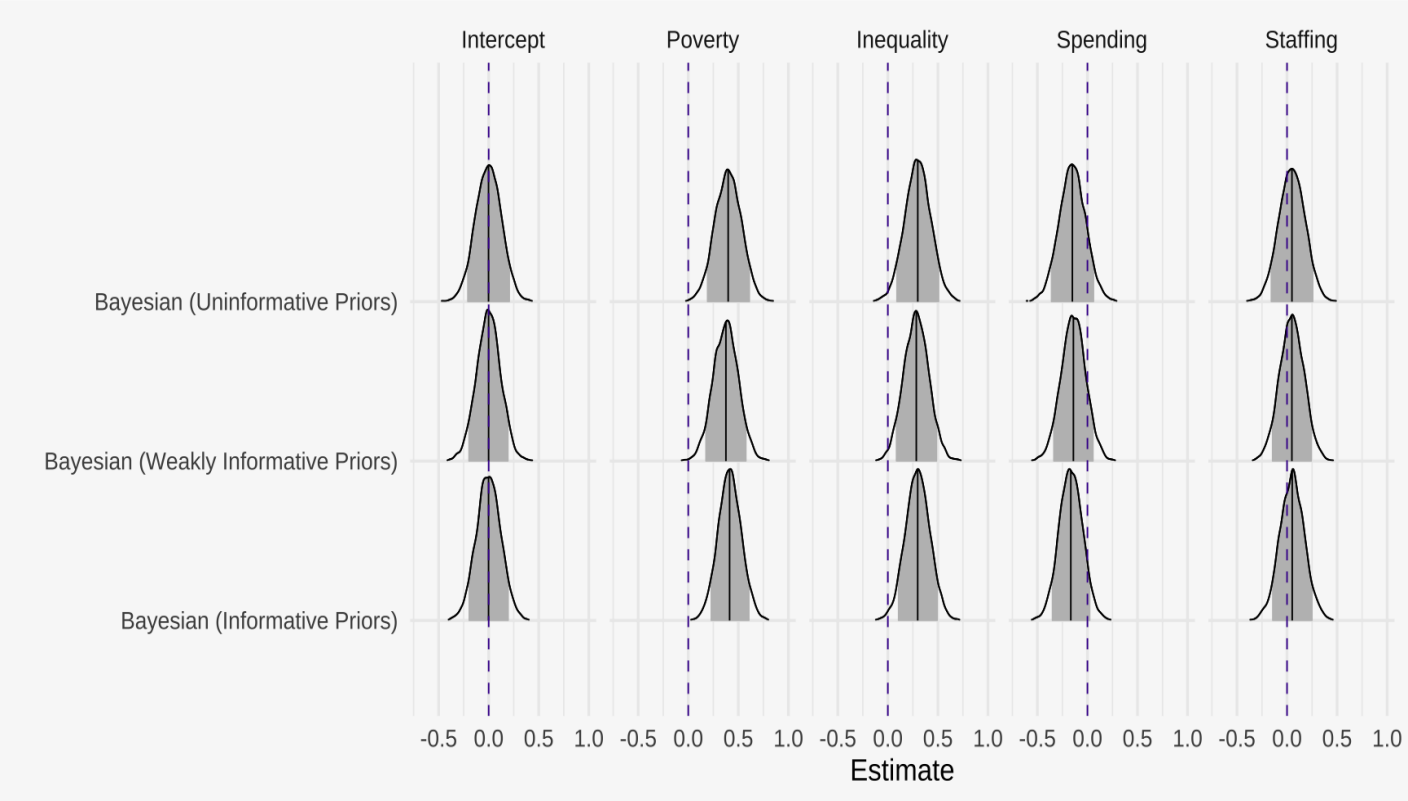
But there are alternatives!





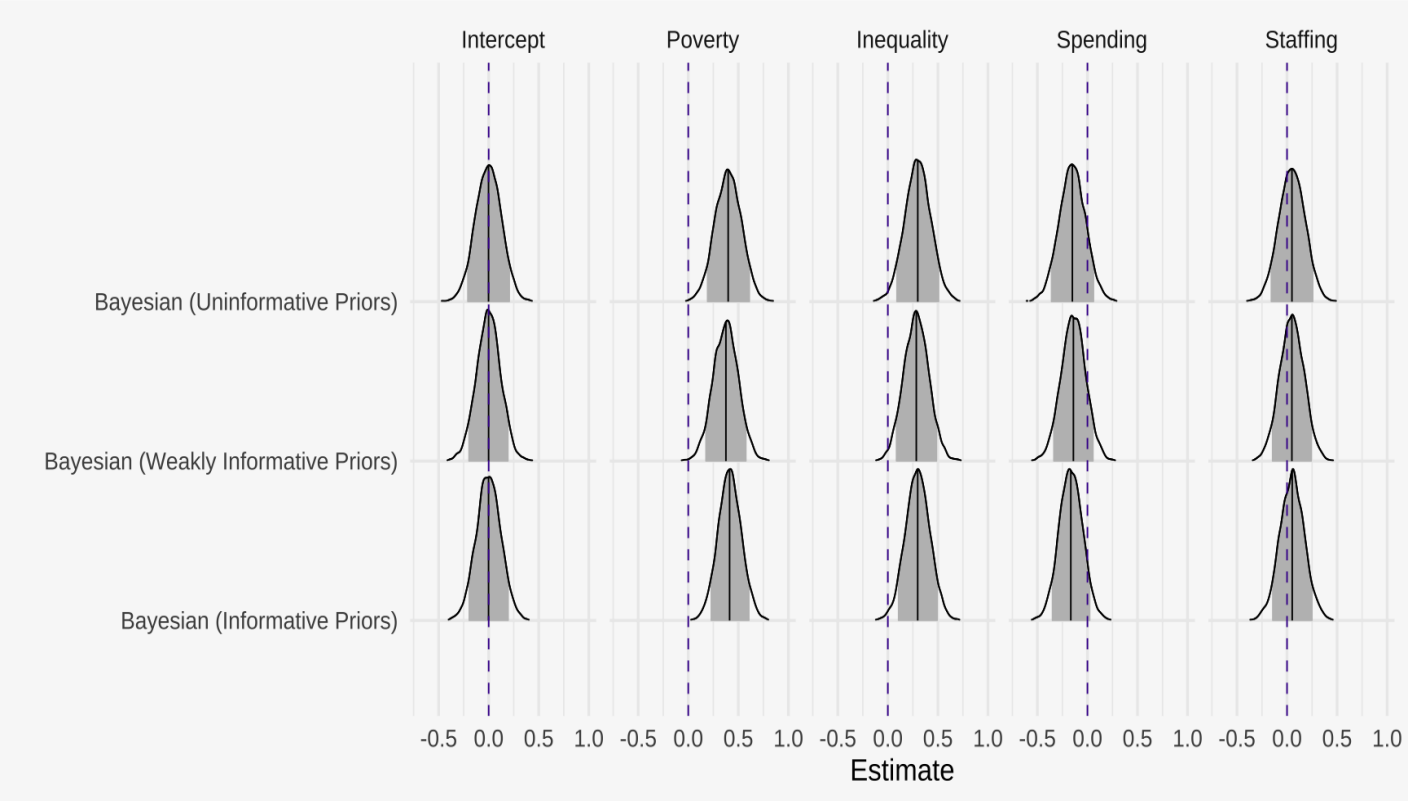
Frequentist NHST Interpretation

- There was no statistically significant association between spending on family support services and rates of children in care (p = 0.244)



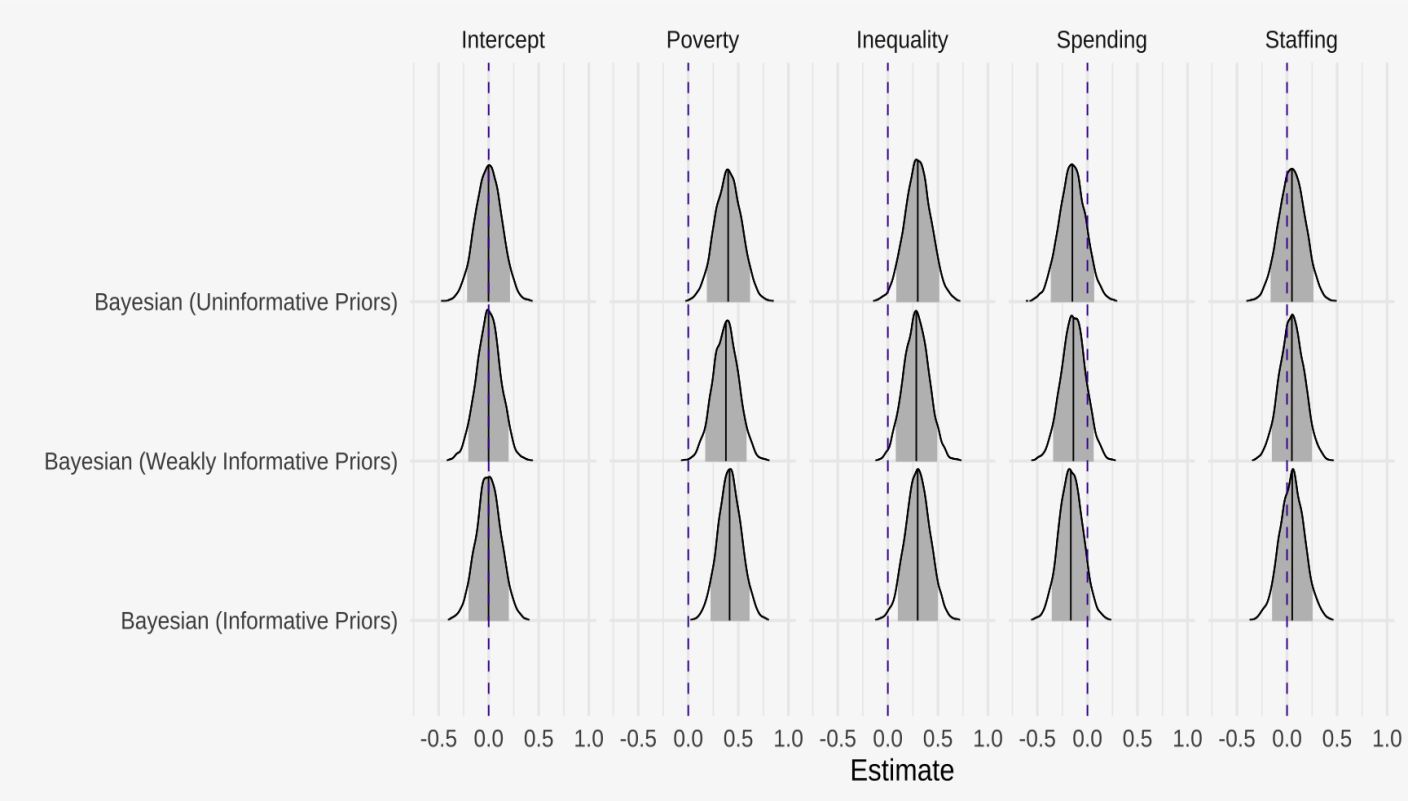
Bayesian descriptive interpretation

- 87% of the posterior distribution was consistent with spending having a negative effect on rates of children in care; greater spending on preventative services is around 6.7 times as likely to reduce rates of care entry than to increase them (*Probability of direction*).



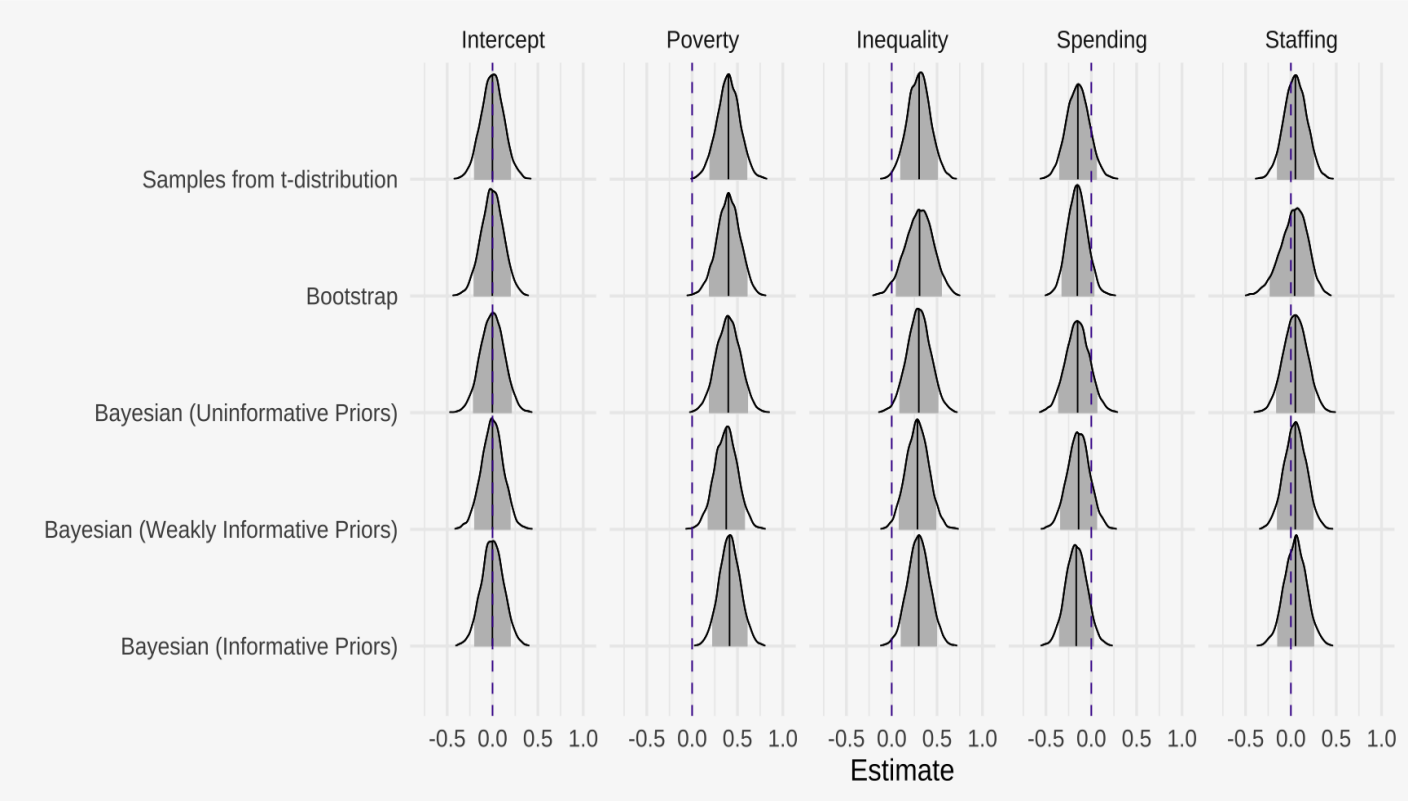
Bayesian descriptive interpretation

- 65% of the posterior distribution was associated with a negative effect of spending beyond what would be considered practically equivalent to zero. There is only a 3% probability that increasing spending would increase rates of care (*Region of Practical Equivalence*).



Bayesian descriptive interpretation

- There is a 76.5% probability that increasing spending would decrease rates of children in care to such an extent to be cost-neutral (< -0.05sd) (*Policy Relevant Value*).



The straightforwardness of these statements is possible because of the Bayesian interpretation of probability: the parameter varies but the data are fixed, our knowledge of the parameter's value is uncertain, rather than the variability of the sample.

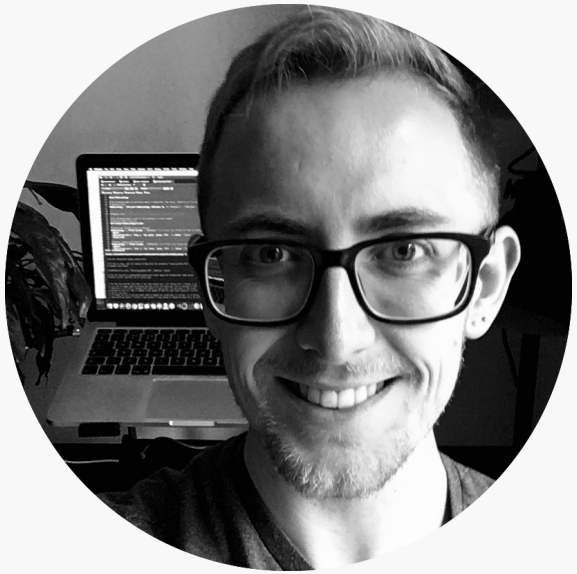
However, that doesn't mean that frequentist descriptions of uncertainty are impossible to imagine. The challenge is the dominance of NHST and difficulty of interpretation.

Conclusions & Reflections

These kinds of apparent populations are common in social policy research. Their units are also meaningful for policy development.

A reliance on NHST, inappropriate at worst and underpowered at best, indirectly generates more evidence for individualised/atomised policies and framing of social problems (where there is more statistical power). How many effects stood no chance of being detected?

There are no shortcuts for making up for the shortfall of statistical power in apparent populations, but alternative descriptive accounts of uncertainty are possible and desirable.



Dr. Calum Webb

Sheffield Methods Institute
The University of Sheffield
The Wave, 2 Whitham Road
Sheffield
S10 2AH

c.j.webb@sheffield.ac.uk

If this presentation has piqued your interest even slightly, I would recommend learning Bayesian data analysis via Richard McElreath's Statistical Rethinking textbook and lecture series followed by checking out the `brms` `R` package.

Remember, the full working paper is available from the SPA2023 conference website and the code for all simulations is available on github

References

Berk, R. A., Western, B., & Weiss, R. E. (1995a). Statistical inference for apparent populations. *Sociological Methodology*, **25**, 421-458.

Berk, R. A., Western, B., & Weiss, R. E. (1995b). Reply to Bollen, Firebaugh, and Rubin. *Sociological Methodology*, **25**, 481-485.

Bollen, K. A. (1995). Apparent and nonapparent significance tests. *Sociological Methodology*, **25**, 459-468.