FRISCHSENTERET
Ragnar Frisch Centre for Economic Research

Oslo, 10.3.2023

Trustworthy and reproducible research at the Frisch Centre

Aim

Empirical research at the Frisch Centre aims to describe and understand the world through statistical analyses informed by relevant institutional, social, and economic contextual facts. This work should strive to be trustworthy, which in turn requires that the work is transparent, reproducible and methodologically sound. This will enable the scientific community to evaluate the work and the extent to which it will generalize outside the analysis data. While not always achievable in practice, this would ideally include:

- **Description of the research workflow**, i.e., clarifying whether the study was pre-registered (and if so, at what level of detail and where the pre-plan is available), whether the model specification was refined while working with the analysis sample, whether the study used holdout samples or other validation techniques, etc.
- **Verifiable information sources** for claims regarding the subject matter of a study (e.g., dates and content of reforms) or its methodology. This is typically in the form of citations or other references to other research, legal documents, newspaper articles or named experts.
- **Clearly described and well-supported statistical techniques** that are justified explicitly through a combination of relevant references and logic
- **Raw data files** with documentation
- **All computer code** used to combine and process the raw data, perform statistical analyses, and generate all published results (e.g., tables, summary statistics and plots)

Making such resources available is required for research to be *verifiable* and for the overall scientific knowledge to be *self-correcting.* This reflects a longstanding principle of science, already recognizable in the 17th century motto of the British Royal Society: "Nullius in verba." ("Take nobody's word for it.").

In the following note, we briefly describe the background for this strategic effort, note some issues of particular importance for the research conducted at the Frisch Centre, and list a set of concrete initiatives and efforts taken to improve the transparency and reproducibility of our work.

Background

The "credibility revolution" in econometrics (Angrist and Pischke 2009; 2010) raised the bar for empirical research – highlighting the need for a more sophisticated understanding of causal inference and statistical analysis amongst applied empirical researchers. This required researchers to shift to new research designs while staying up-to-date on

methodological developments and refinements, as illustrated by past and continuing developments of difference-in-differences techniques (Bertrand, Duflo, and Mullainathan 2004; Roth et al. 2022).

In an analogous way, the "replication crisis" is raising the bar for empirical research by highlighting the need for a more sophisticated understanding of how research workflows and transparency influence the expected validity of results. This, too, will require changes to how we work as applied empirical researchers, and the Frisch Centre aims to play an active part in recognizing the challenges as well as in innovating better work and reporting practices.

The overarching concern of the replication crisis is that the published research base is biased towards statistically significant and "exciting" results, with researchers unwittingly adopting and spreading work practices that raise the chances of publication in high profile journals even as these reduce the credibility and external validity of the research. Comparing published p-values across empirical economics papers in high-ranking journals finds marked clumping of published p-values below conventional significance thresholds (Brodeur et al. 2016; Brodeur, Cook, and Heyes 2020), with weakly powered studies systematically reporting larger effect magnitudes (Ioannidis, Stanley, and Doucouliagos 2017). Across various sub-fields of economics, the average published estimate exceeds the most precise estimates by a factor of 2-4 (Ioannidis, Stanley, and Doucouliagos 2017).

Part of this reflects the importance of researcher degrees-of-freedom – the many judgment based decisions required to specify a statistical model and process raw data into analysis files. "Many-lab" experiments have found large variation in the estimates produced by multiple research groups analyzing the same question in the same data (Silberzahn et al. 2018) or even implementing the same research design on the same data (Huntington-Klein et al. 2021).

This sensitivity of results to a myriad of detailed choices at a level rarely considered in the review process means that results will vary more across "repeated independent samples" than captured by standard errors and p-values. This can lead to systematic bias if researchers refine their statistical specification iteratively while working with their actual data, gravitating towards choices that make the results look plausible, exciting and statistically significant (Gelman and Loken 2014). Importantly, this does not imply conscious p-hacking or unethical practices: Any work process that in practice raises the odds of high profile publications will lead to prestige and spread as researchers seek out successful role models and mentors – even if these practices serve to weaken the validity of the end product (McElreath and Smaldino 2015).

While awareness of these and related issues is growing, there is a lack of consensus on how they are best addressed, with different approaches likely required to fit different types of research. Pre-registered analysis plans are increasingly required for laboratory and field experiments, for instance, but may be harder to implement for analyses of historical reforms and public programs in data from administrative registers with population coverage. Researchers may already be deeply familiar with such data from prior work (Weston et al. 2019), and when new sources are found they may need to be extensively explored and analyzed before researchers can make informed decisions on feasible study designs and their statistical power.

In light of this, we believe it would be premature to mandate specific work practices and standards at the Frisch Centre. Instead, our strategy will be to combine a) transparency measures that clarify the work practice and availability of replication materials of Frisch Centre research, b) participation in external efforts to improve the trustworthiness and reproducibility of empirical research, and c) internal efforts to develop and support better practices.

Highlighted topics

While there is substantial variation in the types of empirical research conducted at the Frisch Centre, the core of this work uses administrative register data with population coverage to study the causal effects of government programs and historical reforms. In this chapter, we highlight some of the issues that have surfaced as particularly important during internal debates at the Centre.

Pre-registration of research projects

During the research process, the researcher is free to adjust or change the research question, to adjust or change model specifications, sample restrictions and variable definitions. This room for discretion may lead researchers, intentionally or not, to selectively present results and specifications that fit a desired narrative or give a misleading impression of robustness.

To narrow this room for discretion we believe the research process should distinguish more clearly between exploratory and confirmatory analyses. One way this can be achieved is by separating the formulation of hypotheses from their testing by specifying the hypotheses and detailing proposed tests in a pre-registration plan. This reduces opportunities for HARKing (Hypothesizing After the Results are Known), since it ensures that the researcher's guess is made before they peek at the data.

In line with this, pre-registration is most credible in cases where the lack of data-peeking can be verified objectively, e.g. before the outcome data exists (e.g., planned experiment or a recent reform) or before the researcher gains access to the data. In such settings, pre-registration *of some form* should be the default, with the analysis plan deposited in a third-party registry such as the American Economic Association's registry for Randomized Controlled Trials or the Open Science Framework (OSF) website.

Pre-registration is a new practice to most empirical economists, and questions remain regarding its scope and value for empirical economics more generally. Some are concerned that a pre-plan can turn into a straitjacket, requiring researchers to implement specifications and data choices that seemed ex ante sensible but turned out to be suboptimal ex post when work began on the data. Unscrupulous researchers could in theory submit multiple pre-analysis plans for the same project and later refer only to the one that "panned out." Or, if they have access to the data, they could analyze it in secret and post a p-hacked specification as a pre-registered analysis plan. And so on.

Such concerns are understandable, but partly reflect a misunderstanding of the purpose of pre-registered plans. Pre-registration is not a silver bullet that ensures verifiably objective and credible research, and deliberate fraud is possible here as in other parts of science. Pre-plans may nonetheless help to move research in the right direction – with a clearer separation between exploratory data analysis on the one hand and confirmatory tests of pre-registered hypotheses on the other.

In line with this, the Frisch Centre will work to establish an internal resource with examples of different types of pre-registration plans, ranging from "light touch" high-level descriptions of the inference strategy and research design to be pursued in a specified context where the data sources are unfamiliar, to detailed specifications of data processing pipelines and statistical tests. We will also explore solutions that involve *partial access* to data with scrambled or missing outcome variables. As researchers gain experience with different approaches in different contexts, we will periodically reflect on these experiences and discuss further improvements and possibilities.

Importantly – the Frisch Centre does not *mandate* or require pre-registration plans. Transparency on how the research process was conducted, including whether or not a pre-registration were made as well as why it were not, is however always preferable. For this reason, all research output from the Frisch centre will be accompanied with a transparency form (see below).

Statistical significance

Statistical significance and p-values tell us how consistent a sample of observed data is with some specific "null" model, but these concepts have long been subject to misinterpretation (Wasserstein and Lazar 2016). As summarized in an editorial in the American Statistician for a special issue focusing on the role of p-values (Wasserstein, Schirm, and Lazar 2019):

> *If you're just arriving to the debate, here's a sampling of what not to do:*
>
> - *Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the p-value passed some arbitrary threshold such as p<0.05).*
> - *Don't believe that an association or effect exists just because it was statistically significant.*
> - *Don't believe that an association or effect is absent just because it was not statistically significant.*
> - *Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.*
> - *Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).*

Of particular concern is the widespread use of significance thresholds as an indicator of whether some result is "worthy", "true" or "important," a practice that strengthens incentives for selective reporting and for using workflows that raise the probability of significant results (Wasserstein, Schirm, and Lazar 2019). As noted in a statement on p-values from the American Statistical Association, "No single index should substitute for scientific reasoning" (Wasserstein and Lazar 2016).

In line with this, we encourage our researchers to avoid using stars (e.g., "***") to signal low p-values in tables and to avoid an emphasis on purely statistical significance in their discussion of results. The value of empirical research will rarely boil down to the statistical properties of single parameters, but will instead consist of how the research adds to a broader evidence base with multiple lines of evidence including previous research, expert knowledge and patterns of different estimated relationships. In many cases, the biggest sources of uncertainty will likely be related to e.g., data processing, identification strategy, model specification, and context, factors not reflected in p-values or standard errors.

Note that none of this is to deny that p-values and significance levels are valid tools that have their appropriate uses. For instance, a pre-registered experiment designed to estimate a causal effect with sufficient precision to distinguish

meaningful and *practically relevant* effects from zero may use significance thresholds as a way to control the error rate. More commonly, however, researchers have many degrees of freedom to adjust analyses and data processing *after seeing the data.* This may be important and even necessary for exploratory work, but the "price of allowing this flexibility is that the validity of any resulting statistical inferences is undermined" (Tong 2019).

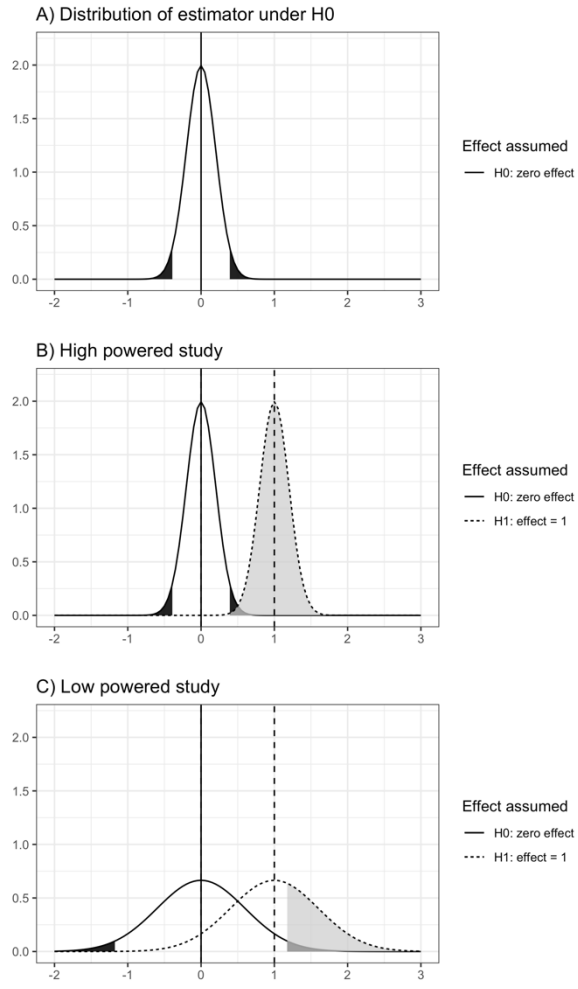
Power calculations

Statistical power can be seen as an indicator of how *informative* a study is expected to be: Will the study have sufficient precision to statistically distinguish effects of plausible size from zero or to provide confidence intervals that are sufficiently small to provide new and useful information?

When planning experiments, power calculations help us to assess whether the experiment will be sufficiently informative to justify its costs, assess how changes to the experimental design or sample size would affect precision, and compare different alternative designs and sample sizes.

In other types of empirical research, the statistical power may be more outside of the researcher's control: The sample size cannot easily be increased if we are already using administrative registers with population coverage, and the type and amount of identifying variation in the data will be whatever it happened to be when the data were recorded.

In these contexts, analyses of statistical power can help researchers make more informed choices when deciding which reforms and causal effects to study. A low-powered study may mean that we are unlikely to see a statistically significant estimate even if the expected effect is truly there. In fact, with low power we may have a design where only *strongly misrepresentative* point estimates reach statistical significance (see fig 1).



An unbiased estimator is distributed around the true effect – which is unknown.

Assuming a zero effect under the null, the shaded tails show values that are unlikely under the null.

A statistically significant result is one that is unlikely under the null hypothesis: equal to or exceeding the values in the shaded tails.

If we believe the true effect is actually equal to 1, the estimator will actually be distributed around the value 1. Significant results (shaded) are those exceeding the significance thresholds.

In a high-powered study, the probability that this estimator will be statistically significant is large (e.g., higher than 80%).

In a low-powered study of the same effect, the estimator will vary more across repeated independent samples.

This means that large estimates are more likely under the null, which means that only larger estimates will be significant.

This means that we are less likely to see significant results, and that only *exaggerated* estimates larger than the true effect will be significant

Figure 1 - Statistical power

These problems with low-powered studies are further compounded when researchers avoid pre-registration and use traditional workflows where the analysis is refined while working with the data. Researchers may (rightly) believe that there is *some* effect operating, and may try different ways of slicing the data and specifying the model to see if they can bring out this effect in the analysis. As an example, consider a study of some treatment on workers' sickness absence: If there is no effect on *total sickness absence days*, maybe there is an effect on the *probability of long* spells (> 16 days), or on the *duration* of long spells, or on the *share* of spells with some specific set of *diagnoses*, or maybe the effect is primarily seen for workers in certain *industries* or *age groups*. While each of these possibilities may be triggered by and be in line with prior research and theory, the risk is that our workflow implicitly selects for statistically significant low-powered estimates that systematically exaggerate the true effect. In addition, we would *de facto* be testing multiple hypothesis – exploring a branching set of possible analyses and retaining a selected subsample without correcting for multiple hypothesis testing. The end result may be research that looks convincing - but where the seemingly strong statistical results may actually reflect selective reporting and work practices that produce inflated estimates with misleading p-values.

To reduce such risks, the Frisch Centre aims to increase the use of power analyses at an early stage of register-based studies. At present, however, there is a lack of simple tools and approaches that make this feasible for register-based studies that plan to exploit identifying variation in historical data, often using subtle and complex models. An internal task-group is planned to explore this question in more detail, with the aim of making a brief "methods note" that describes various approaches.

Reproducibility

> Openness in research is a precondition for scientific development, accountability, and critique. Public sharing of data, research material, and results is a precondition for developing knowledge, comparing research results, and assessing the analyses, interpretations, and conclusions of academic peers. Data material as well as results should therefore be shared with other researchers as openly as possible
>
> (The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH) 2021).

Empirical research typically involves a large number of decisions that are necessarily left out of the research paper due to space limitations. This can make it exceedingly difficult to reproduce published results based on the information in the published text, tables and figures. In many cases, then, a piece of research is only reproducible if other researchers have access to the code used to process the raw data, run the analyses, and construct published tables and figures.

To facilitate reproducibility the Frisch Centre urges our researchers to always include a document listing the data sources used (ideally, for recently acquired data, with information on the specific variables ordered from Statistics Norway), as well as including the programs used. This level of transparency is increasingly being demanded by scientific journals, with AEA journals and Economic Journal as high-profile examples in Economics. To further strengthen this development in the field, the Frisch Centre has also joined an initiative from NHHs compliance officer, Erik Sørensen, for a replication network that will make it easier to establish *proven reproducibility* for research involving Norwegian register data.

Description of institutional details

Many research papers study the consequences of variation in institutional arrangements and/or changes thereof. Such contextual information will often be the evidence that supports the identifying assumptions of causal research designs, and the credibility of the research will thus hinge on whether the institutional details are correct, precise and comprehensive. In contrast to most other issues in a research paper, (international) peer review will in many cases be unable to assess and verify institutional details embedded in Norwegian policy documents and expertise. The description of institutional details should be supported by referenced, relevant sources (preferably legal documents), and ideally verified by (named) subject matter experts.

Comment articles

Most research will have its blind or weak spots, and even the best researchers will make mistakes. Correcting the scientific record when such mistakes are found is an essential part of the scientific process. As stated by the *Guidelines for Research Ethics in the Social Sciences and the Humanities*, §2, "Researchers should not withhold substantial critique (…)" (The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH) 2021).

While most researchers will agree to this *in principle,* writing comment articles to communicate critique can be both time consuming and unpleasant, potentially triggering retaliation or animosity from those criticized. Such comments may also be only tangentially related to ongoing research projects with external financing, requiring additional work on top of the regular workload. To encourage our researchers to write such comments, the Centre will thus give weight to such work when distributing internal research funds.

Specific initiatives

Transparency form

A transparency form will be required for all empirical research conducted at the Frisch Centre. This form does not impose or require any specific solutions but clarifies to what extent the work was pre-registered, the availability of computer code, data files and other replication materials, and whether the research has been evaluated by researchers outside the author team. The first version of the form was implemented beginning in January 2023, and the form is expected to be revised over time in light of experience and ongoing changes in identified best practices.

Norwegian Replication Network

High-ranking scientific journals increasingly demand evidence that all published results can be reproduced by applying the author's submitted computer code to the underlying raw data files. This is challenging when the data are on loan from Statistics Norway and come from administrative data registers with population coverage. The Frisch Centre is a participant in a newly formed network of research groups that will jointly test the reproducibility of each other's research in order to document reproducibility in such cases.

Seminars and workshops

The Frisch Centre has a weekly seminar series, which will continue to periodically include external researchers working on research transparency and credibility – with an ambition of at least two talks on related topics annually.

Internal task forces

Implementing and developing these practices will be supported as needed by temporary internal "task forces" asked to review new developments, synthesize experiences and best practices, and gather/develop new tools and resources. These will address specific topics, such as "pre-registration plans and observational data", or "power-calculations for register data studies."

References

Angrist, J. D, and J. S Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *The Journal of Economic Perspectives* 24 (2): 3–30.

Angrist, Joshua David, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Vol. 1. Princeton university press Princeton. https://pup.princeton.edu/titles/8769.html.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249–75.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science: Data-Dependent Analysis--"garden of Forking Paths"--Explains Why Many Statistically Significant Comparisons Don't Hold Up." *American Scientist* 102 (6): 460–66.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–60. https://doi.org/10.1111/ecin.12992.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127 (605): F236–65. https://doi.org/10.1111/ecoj.12461.

McElreath, Richard, and Paul E. Smaldino. 2015. "Replication, Communication, and the Population Dynamics of Scientific Discovery." *PLOS ONE* 10 (8): e0136088. https://doi.org/10.1371/journal.pone.0136088.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *ArXiv:2201.01194 [Econ, Stat]*, January. http://arxiv.org/abs/2201.01194.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56. https://doi.org/10.1177/2515245917747646.

The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH). 2021. "Guidelines for Research Ethics in the Social Sciences and the Humanities." The Norwegian National Research Ethics Committees. https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/guidelines-for-research-ethics-in-the-social-sciences-humanities-law-and-theology/.

Tong, Christopher. 2019. "Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science." *The American Statistician* 73 (sup1): 246–61. https://doi.org/10.1080/00031305.2018.1518264.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. “The ASA Statement on P-Values: Context, Process, and Purpose.” *The American Statistician* 70 (2): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. “Moving to a World Beyond ‘p < 0.05.’” *The American Statistician* 73 (sup1): 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Weston, Sara J., Stuart J. Ritchie, Julia M. Rohrer, and Andrew K. Przybylski. 2019. “Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets.” *Advances in Methods and Practices in Psychological Science* 2 (3): 214–27. https://doi.org/10.1177/2515245919848684.