COLUMBIA | SIPA
Center for Environmental Economics and Policy

CEEP Working Paper Series
Working Paper Number 20

March 2022

Unconfounded but Inflated Causal Estimates

Vincent Bagilet and Léo Zabrocki

https://ceep.columbia.edu/sites/default/files/content/papers/n20.pdf

Unconfounded but Inflated Causal Estimates[†]

Vincent Bagilet[1]

Léo Zabrocki[2]

March 1st, 2022

[Preliminary Work]

Abstract

Convincing research designs make empirical economics credible. To avoid confounding, quasi-experimental studies focus on specific sources of variation. This could lead to a reduction in statistical power. Yet, published estimates can overestimate true effects sizes when power is low. Using fake data simulations, we show that for all causal inference methods, there could be a trade-off between confounding and exaggerating true effect sizes due to a loss in power. We then discuss how reporting power calculations could help address this issue.

[†]**Comments and suggestions are highly welcome.** We are very grateful to Hélène Ollivier and Jeffrey Shrader for their guidance on this project. Many thanks to Jesse McDevitt-Irwin, José Luis Montiel Olea, Claire Palandri, as well as Jeffrey Shrader's lab members for helpful comments and seminars participants at Columbia and Thomas Piketty's lunch seminar at PSE for their feedback.

[1]Columbia University, New York, USA. Email: vincent.bagilet@columbia.edu

[2]Paris School of Economics and École des Hautes Etudes en Sciences Sociales, Paris, France. Email: leo.zabrocki@psemail.eu

One of the main challenges in empirical economics is to reduce confounding to identify causal effects. Identifications strategies based on Regression Discontinuity (RDD), Instrumental Variable (IV) and Difference-in-Differences (DID) can help achieve this goal. To do so, these strategies only use part of the variation in the data. They exploit the exogenous part of the variation in the treatment or decrease the sample size by only considering observations for which the "as if random" assumption is credible. As we point out in this paper, limiting variation can decrease statistical power, *i.e.*, the probability of detecting an effect when there is actually one. This could create a tension between statistical power and reducing confounding.

In settings with low power, statistically significant estimates will exaggerate the true effect size (Ioannidis 2008, Gelman and Carlin 2014). Only estimates at least two standard errors away from zero will be statistically significant at the 5% level. In under-powered studies, these estimates make up a selected sub-sample of all estimates, located in the tails of the distribution of all possible estimates. The average of these statistically significant estimates will differ from the true effect, located at the center of the distribution if the estimator is unbiased. When power is low, obtaining a statistically significant estimate from an unbiased estimator does not guarantee that it will be close to the true effect.

Furthermore, a large literature has underlined the existence of a publication bias towards estimates that are statistically significant (Rosenthal 1979, Andrews and Kasy 2019, Abadie 2020, Brodeur et al. 2020, for instance). This statistical significance filter can lead published estimates from under-powered studies to greatly exaggerate true effect sizes. Gelman and Carlin (2014) calls this inflation of significant estimates type M (Magnitude) error. The exaggeration of significant estimates can be computed as the expected value of statistically significant estimates over the true effect. It increases when power decreases (Lu et al. 2019, Zwet and Cator 2021).

This issue partly explains the current replication crisis affecting various fields such as economics, epidemiology, medicine or psychology (Button et al. 2013,

Open Science Collaboration 2015, Camerer et al. 2016, Chang and Li 2022). Even in experimental economics, with a high level of control and an arguable absence of confounders, estimates published in top economic journals have failed to replicate. Camerer et al. (2016) replicated 18 laboratory experiments and found that the effect size of original studies was on average 1.5 times larger than the replicated one. If we assume that the point estimates obtained in the replication studies represent the true effect, 44% of the original studies would not reach the conventional 80% power threshold. The treatment allocation in these under-powered studies may have by chance or through researchers' degrees of freedom produced estimates in the tails of the distribution, making them more likely to be published.

Quasi-experimental studies could be more prone to this issue since statistical power is not central to the analysis in current practices. Despite usually large sample sizes, Ioannidis et al. (2017) concernedly finds that the median statistical power in a wide range of economic studies is no more than 18% and that nearly 80% of estimates may be exaggerated by a factor of two. Understanding the determinants of low power is key to avoid the inflation of published estimates.

In this paper, we show that design choices in quasi-experimental studies can be seen as a trade-off between avoiding confounding and overestimating true effect sizes due to a resulting loss in power. To limit the threat of confounding, causal inference methods discard variation in the treatment. It can lead to a reduction in statistical power. Due to the statistical significance filter, the resulting published estimates could be inflated and thus deceptive.

In the first section of this paper, we illustrate the existence and consequences of this trade-off using fake-data simulations based on examples drawn from education, labor, environmental and political economics. We consider separately the main causal inference methods used in the economics literature: RDD, IV, DiD, event study, as well as selection on observables through matching. For each identification strategy we discuss the key factors affecting the confounding / exaggeration trade-off. In RD designs, while the initial sample size may be large, we discard part of the variation by only considering observations within the bandwidth, de-

creasing the effective sample size. In an IV setting, we only use part of the variation in the treatment, the portion explained by the instrument. In DiD event studies, the variation used to identify an effect sometimes only comes from a limited number of treated observations. When assuming that all confounders are measured, matching prunes treated units that cannot be matched to untreated ones.

In the second section of the article, we discuss solutions to assess whether statistically significant estimates from observational studies could be inflated. We advocate reporting power calculations. They can be computed before and after the analysis is carried out. By approximating the data generating process, prospective power simulations help identify the design parameters affecting power (Gelman 2020, Black et al. 2021). Retrospective power calculations allow to evaluate whether a study would have enough power to confidently estimate a range of smaller but credible effect sizes (Gelman and Carlin 2014, Stommes et al. 2021). Our companion website describes in details how such solutions can be implemented.

Our paper contributes to three strands of the literature. First, the idea that causal identification estimators, while unbiased, may be imprecise is not new; this is the well-known bias/variance trade-off (Imbens and Kalyanaraman 2012, Deaton and Cartwright 2018, Hernán and Robins 2020, Ravallion 2020). In underpowered studies, resulting estimates have large confidence intervals, suggesting that a wide range of effects are consistent with the data. We approach this literature from a different angle: through the prism of statistical power and publication bias. Not only the limited precision resulting from the use of causal identification methods could make it difficult to draw clear conclusions regarding the exact magnitude of the effect but we argue that it might also inherently lead to inflated published effect sizes.

Second, recent studies discussing the inflation of statistically significant estimates due to low power focused on specific causal identification methods separately (Schell et al. 2018, Black et al. 2021, Stommes et al. 2021, Young 2021). We show that using causal identification methods may in itself cause power issues.

This connection could be exacerbated by the fact that, as noted by Brodeur et al. (2020), publication bias is more prevalent for some methods such as the IV.

Third, our study contributes to the literature on reproducibility in economics (Camerer et al. 2016, Ioannidis et al. 2017, Christensen and Miguel 2018, Kasy 2021). The trade-off presented in this paper may be an additional explanation for observing replication failures in empirical economics, despite the widespread use of convincing causal identification methods.

1 Simulations

The inflation of statistically significant estimates can be measured by the ratio of the estimated effect over the true effect. It can therefore only be computed if the true effect is known, which is never the case in real world settings. We therefore turn to fake data Monte-Carlo simulations to compare the two measures of interest, $\frac{\mathbb{E}[\hat{\beta}]}{\beta_0}$ and $\frac{\mathbb{E}[\hat{\beta}|\text{significant}]}{\beta_0}$.

For clarity, we split the analysis by identification strategy. While the general idea that causal inference methods discard variation to identify effects is shared across strategies, the confounding / exaggeration trade-off is mediated through a distinctive channel for each of them. We build simulations that reproduce real world examples from economics of education for RDD, labor economics for matching, political economy for IV and environmental economics for DiD event studies. Real world settings enable to clearly grasp the relationships between the different variables and to set realistic parameter values. Since our simulations have an illustrative purpose only, we intentionally restrict our simulation exercise to settings in which power can be low. All our models are correctly specified and accurately represent the data generating process, except for the omitted variable bias (OVB).

For each identification strategy, we start by laying out the intuition behind the method and how it enables to estimate causal effects. It naturally points to the key parameter through which the confounding/exaggeration trade-off is mediated. We then briefly describe the example setting considered and our simulation assump-

tions. The exact simulation processes are described in more details on the project's website. We finally display the simulation outputs and discuss the implications of the trade-off that are specific to the identification strategy considered.

1.1 Matching

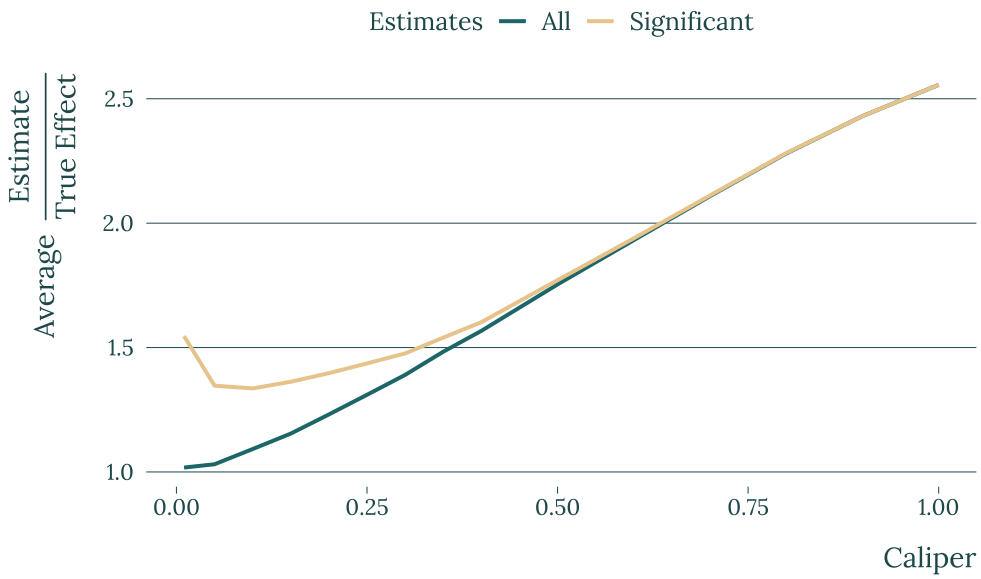
We first focus on the ideal case for which all confounders are assumed to be observed. Under this assumption, one can use matching to estimate a causal effect specific to matched treated units. Contrary to naive regression models, this method makes the common support of the data explicit, avoids model extrapolation and non-parametrically adjusts for observed confounders.

In the case of propensity score matching, the distance metric used to group similar treated and control units is called a caliper. It is expressed in standard deviation of the propensity score distribution. The smaller the caliper, the more comparable units are and therefore the lower the risk of confounding is. Yet, with a stringent caliper, some units may not be matched, decreasing the sample size. This might lead to an important loss in statistical power and inflated estimates. In the case of matching, the confounding / exaggeration trade-off is therefore mediated by the value of the caliper.

We illustrate this issue by simulating a non-randomized labor training program as done by Dehejia and Wahba (1999). Individuals self-select into the program and may therefore have different characteristics from individuals who do not choose to enroll. We first define the treatment status of individuals. We then simulate their probability to be treated, making sure that the propensity score distributions of the two groups only partially overlap. We finally create the two potential revenues for each unit and express the observed revenue according to their treatment status. We generate 1000 datasets and for caliper values ranging from 0 to 1, we regress the observed revenue on the treatment indicator.

Figure 1 indicates that the bias of estimates, regardless of their statistical significance, decreases with the value of the caliper as units become more comparable. As the caliper decreases, statistically significant estimates start being more inflated

Figure 1: Evolution of Bias with the Caliper in Propensity Score Matching, conditional on significativity.



Notes: The green line indicates the average bias for all estimates, regardless of their statistical significance. The beige line represents the inflation of statistically significant estimates at the 5% level. The caliper is expressed in standard deviation of the propensity score distribution. Details on the simulation are available at this link.

than the entire sample of estimates. For large caliper values, units are not comparable enough and confounding bias the effect. For small caliper values, the sample size may become too small to be able to precisely estimate an effect of this magnitude and exaggeration arises. In some settings, statistically significant estimates may thus never get close to the true effect.

1.2 Regression Discontinuity Design

To identify a causal effect, RDD relies on the assumption that for values close to the threshold, treatment assignment is quasi-random. Under this assumption, individuals just below and just above the threshold would be comparable on average and only differ in their treatment status. To avoid confounding, the RDD thus only considers observations within a certain bandwidth around the threshold and discards observations further away. The effective sample size used for identifica-

tion differs from the total sample size. In the case of the RDD the confounding / exaggeration trade-off is mediated by the size of the bandwidth.

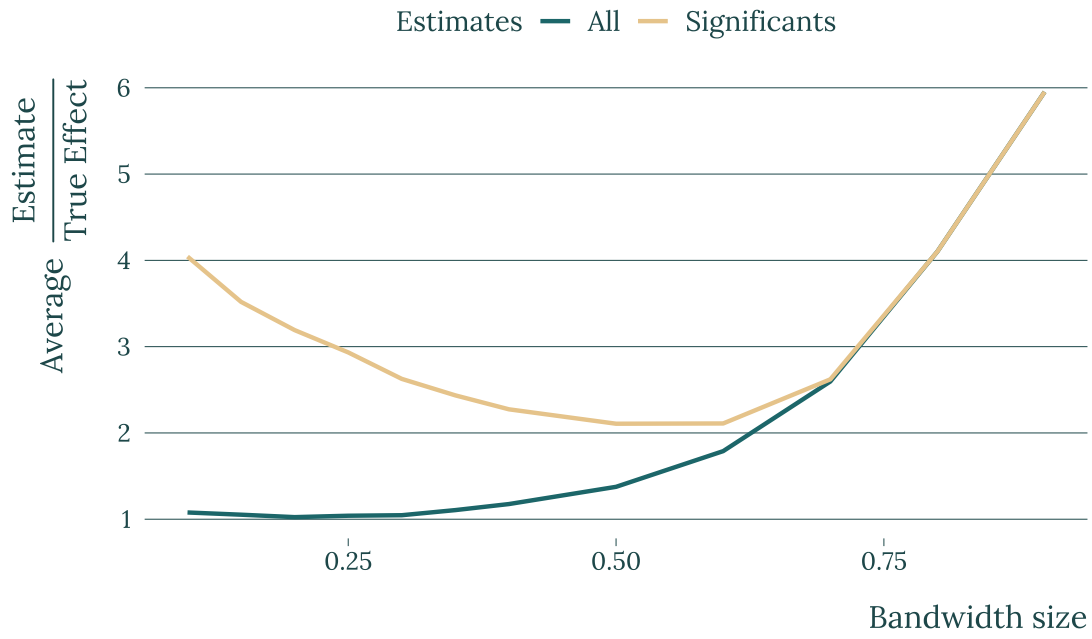
To illustrate this trade-off, we consider a standard application of the sharp RD design in economics of education in which students are offered additional lessons based on the score they obtained on a standardized test. Thistlethwaite and Campbell (1960) introduced the concept of RDD using a similar type of quasi-experiment. Students with test scores below a given threshold receive the treatment while those above do not. Since students far above and far below the threshold may differ along unobserved characteristics such as ability, a RDD estimates the effect of the treatment by comparing outcomes of students whose initial test scores are just below and just above this threshold.

To simplify, we only consider 4 variables in our simulation. We define an individual qualifying score as a non-linear function of a uniformly distributed individual ability and a random noise. A binary individual treatment status deterministically depends on the qualifying score. We then generate a final score as a non-linear function of the unobserved ability, the treatment status and a random noise.

We estimate the treatment effect by regressing the final score on the treatment status and the qualifying score for different bandwidth sizes. We reproduce this analysis by generating 1000 data sets.

Figure 2 displays the results of these simulations, describing how the ratio of the mean of the estimates over the true effect evolves with bandwidth size, conditional on significance. As for matching, in some settings, statistically significant estimates may never get close to the true effect. For large bandwidths, the omitted variable biases the effect while for small bandwidths, the small sample size creates power issues. The optimal bandwidth literature describes a similar trade-off but considers different consequences. They consider a bias/precision trade-off, we consider a omitted variable bias / exaggeration bias trade-off.

Figure 2: Evolution of Bias with Bandwidth Size in Regression Discontinuity Design, conditional on significativity.



Notes: The green line indicates the average bias for all estimates, regardless of their statistical significance. The beige line represents the inflation of statistically significant estimates at the 5% level. In this simulation, N = 10,000. The bandwidth size is expressed as the proportion of the total number of observations of the entire sample. Details on the simulation are available at this link.

1.3 Instrumental Variable Strategy

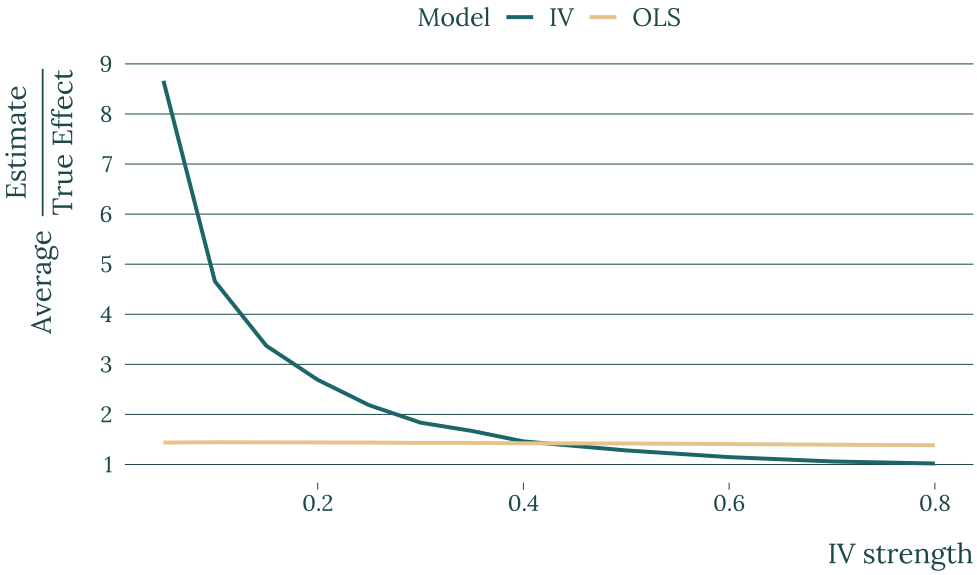
Using an instrumental variable enables to get rid of confounding by only considering the exogenous variation in the treatment. When this exogenous fraction of the variation is limited, the instrument can still successfully eliminate confounding on average. However, the IV estimator will be imprecise and power low. In the case of the IV, the confounding / exaggeration trade-off is therefore mediated by the "strength" of the instrument considered. The weaker the instrument, the more inflated statistically significant estimates will be.

To illustrate this trade-off, we consider the example of studies estimating the impact of voter turnout on election results. In this setting, to avoid the threat of confounding, one can to take advantage of exogenous factors that affect voter

turnout such as rainfall.

We draw the rainfall height from a Gamma distribution. We define turnout as a linear function of rainfall and of an unobserved variable drawn from a centered normal distribution. We call "IV strength" the value of the parameter linking rainfall to turnout. We construct vote share as a linear function of turnout and of the unobserved variable. We also add normally distributed error terms. We choose the values of the parameters to mimic real world distributions. We then regress the vote share on turnout instrumented by rainfall. For different values of IV strength, we reproduce this analysis by generating 1000 data sets and compare the performance of the IV to that of the "naive" OLS regression of vote share on turnout.

Figure 3: Evolution of Bias for Statistically Significant Estimates Against Intensity of the Instrument in Instrumental Variable Design.



Notes: The green line indicates the average bias for statistically significant IV estimates at the 5%. The beige line represents the bias of statistically significant OLS estimates at the 5% level. The strength of the instrumental variable is expressed as the value of the linear parameter linking rainfall to turnout. In this simulation, N = 10,000. Details on the simulation are available at this link.

Figure 3 displays, for different IV strengths, the average of statistically significant estimates scaled by the true effect size for both the IV and the OLS. When the

instrument is strong, the IV will recover the true effect, contrarily to the OLS. Yet, when the IV strength decreases, the exaggeration of statistical significant estimates skyrockets. Even if the intensity of the omitted variable bias is large, for limited IV strengths, the exaggeration ratio can become larger than the OVB. When the only available instrument is weak, using the "naive" OLS would, on average, produce statistically significant estimates that are closer to the true effect than the IV.

A large F-statistic does not necessarily shield against this problem. For the parameter values considered here, this phenomenon arises even in cases for which the F-statistic is substantially larger than the usually recommended threshold of 10, as illustrated in figure 5 in appendix.

1.4 Difference-in-Differences and Event Studies

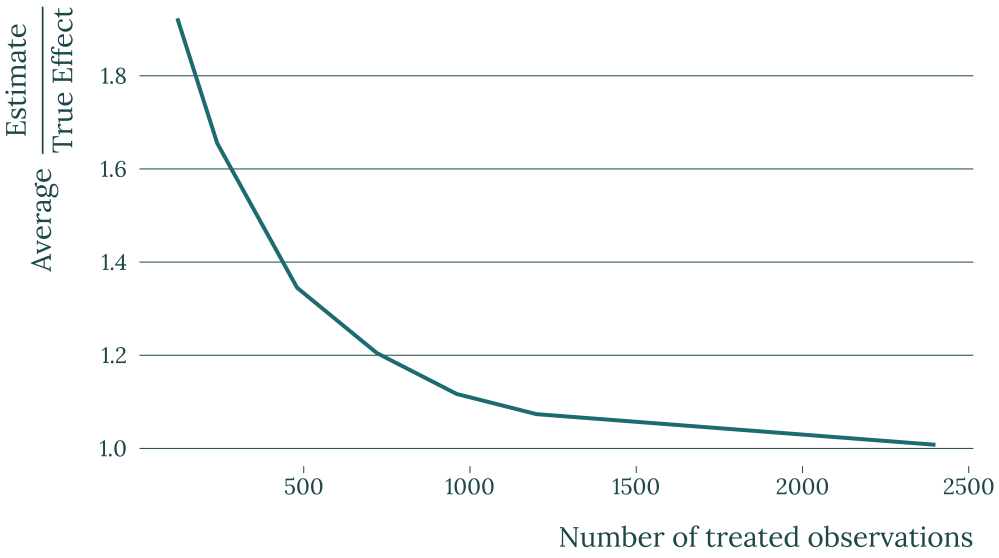
To avoid confounding, DiD event studies take advantage of exogenous shocks. In many settings, while the number of observations may be large, the number of events, their duration or the proportion of individuals affected might be limited. As a consequence, the number of treated observations is small and the variation available to identify the treatment is limited. In studies using discrete exogenous shocks, a confounding / exaggeration trade-off is thus mediated by the number of observations treated. It does not only concern DiD event studies but is particularly salient in this case.

This trade-off is linked to both questions of sample size and proportion of units treated. The larger the sample size, the larger the power. For a given sample size, power is maximized when the proportion of treated observations is equal to the proportion of untreated ones.

We simulate an analysis on the impacts of air pollution reduction on birth-weights. To avoid confounding, one can exploit exogenous shocks to air pollution such as plant closures, plant openings, strikes, creation of a low emission zone or of an urban toll, etc. We simulate our analysis at a zip code and monthly level and focus on the example of toxic plant closures. We consider that some zip codes experience a permanent plant closure over the study period and others do not. We

generate birthweights that depend on normally distributed zip code and time fixed effects, the treatment status and some noise. We set parameters values to emulate a realistic study. We then regress the average birthweight in zip code z and period t on the treatment status, adding fixed effects to the regression. We reproduce this analysis 1000 times for different numbers of treated observations but a fixed total number of observations.

Figure 4: Evolution of Bias for Statistically Significant Estimates against the Number of Treated Observations in Difference-in-Differences Event Study Design.



Notes: The green line indicates the average bias for statistically significant estimates at the 5%. In this simulation, N = 120,000. Details on the simulation are available at this link.

Even though the actual sample size is extremely large in our example, if the number of treated observations is small, the exaggeration can be important, as shown in figure 4. A very large number of observations does not necessarily prevent the exaggeration issue to arise.

1.5 Cross-cutting issues

Virtually any practice that leads to a reduction in precision increases the risk of falling into exaggeration problems. The issues discussed above are for the most

part specific to each identification strategy. Yet, some practices in causal observational studies are shared across all methods and can further create exaggeration.

Clustering is key to take into account an unspecified and potentially unobserved correlation structure in the error terms. By definition, it produces larger standard errors. If estimates that remain statistically significant when clustered at a high level are more likely to be published, published under-powered studies will on average overestimate the true effect.

As briefly discussed in subsection 1.4, precision and thus power are maximized when the proportion of treated is 0.5. This fact, widely discussed in the context of experiments, also applies to observational settings. For a given sample size, a large imbalance between the size of the control and treatment groups reduces power and may create an exaggeration issue.

Finally, in count models, when the number of occurrences in the output variable is limited, the variation available to identify the effect can be very limited. In such situations, power may be drastically low and exaggeration ratios extremely large.

2 Practical Recommendations

Now that we have shown that causal methods can produce inflated estimates when the study is under-powered, how can we address this problem? Even though it does not produce uninflated estimates, reporting power calculations enables to evaluate the risk of exaggeration for a study. In this section, we present a workflow to evaluate and report the power of a study before and after its implementation. We then discuss how changing our attitude towards statistical significance and replicating studies can help limit this issue.

2.1 Before Analyzing the Data

In randomized controlled trials, presenting statistical power calculations before running the experiment is not only an established practice but also a requirement (Duflo et al. 2007, McConnell and Vera-Hernandez 2015, Athey and Imbens 2016). In observational studies power is however rarely reported, despite the availability of specific power formulas for some causal inference methods (Freeman et al. 2013, Cattaneo et al. 2019). Two main reasons could explain this limited reporting of power calculations. First, we do not directly control the data collection process. Second, we may fear that available power formulas are not flexible enough to capture the complexity of their design. On top of these reasons, there is a lack of guidance on how to design well-powered observational studies. In causal inference textbooks, very few pages are devoted to the topic (Angrist and Pischke 2009; 2014, Imbens and Rubin 2015, Cunningham 2021). To the best of our knowledge, only two remarkable textbooks discuss the matter in depth (Shadish et al. 2002, Huntington-Klein 2021).

Simulating the design of an observational study is a solution to overcome these limits (Hill 2011, Gelman 2020, Black et al. 2021). Similarly to what we did in the previous section, the goal of this approach is to simulate the data generating process of the study from scratch. It requires thinking about both the distribution of the variables and their interconnections. External information found in previous studies can help guide the simulation process to make it more realistic. If the relationships among covariates are too complex to emulate, a second approach starts from an existing dataset to which a simulated treatment and potential outcomes are added[3].

When simulations indicate that statistical power is low, additional data could be collected or the statistical model could be expanded to increase precision. In any case, it should not stop from carrying out a research project. Simulation results

[3] In the future version of the paper, we will explain how we can easily simulate from scratch the study by Card (1993). For now, examples of simple simulations are available on our companion website.

rest on the way the data generation process was modeled and it can be difficult to gauge the amount of noise present in data before actually analyzing them. The two actual benefits of a prospective simulation procedure are to think about factors that affect power and not to be mislead by statistically significant estimates if power is low.

2.2 Once the Main Analysis is Completed

Once we have obtained a statistically significant estimate for the treatment of interest, we still need to think about the statistical power of the study to check whether the magnitude of our estimate is trustworthy. A *retrospective* power analysis helps evaluate whether the design of the study would produce uninflated statistically significant estimates if the true effect was smaller than the observed estimate (Gelman and Carlin 2014, Ioannidis et al. 2017, Stommes et al. 2021).

We illustrate how a retrospective analysis works by taking the example of Card (1993) on the relationship between human capital and income. He finds that an additional year of education, instrumented by the distance of growing near a four-year college, causes a 13.2% average increase in wage. The associated standard error is 5.5%. As noted by the author himself, the estimate is very imprecise: if the true causal effect was slightly smaller than the observed estimate, the study would very likely be under-powered. For instance, imagine that prior evidence suggests that the true effect could be to closer a 10% increase in wage. Computing the statistical power of the study only requires to draw many estimates from a normal distribution centered around the hypothesized true effect of 10% and with a standard deviation equal to the 5.5% standard error obtained in Card's study. Concretely, one proceeds as if they were able to replicate the study many times under the assumption that the true effect is different from the observed estimate. The proportion of sampled estimates that are statistically significant at the 5% level, 44% in this case, is the statistical power. The inflation of significant estimates is then computed as the average ratio of the values of statistically significant estimates over the assumed true effect size: these estimates would be 1.5 times too

large on average.

For a retrospective power analysis to be useful, it is necessary to make informed guesses about the range of plausible effect sizes. Such guesses can be based on results from meta-analyses or previous studies with a convincing design (e.g., a randomized controlled trial). When such information is not available, power calculations can be run for a range of smaller but credible effect sizes[4].

Results from power and exaggeration calculations would not only be highly informative but could also be reported very concisely in the robustness section of articles. `R` and `Stata` packages have been developed (Timm et al. 2019, Linden 2019) to easily implement retrospective power analyses.

2.3 Attitude Towards Statistical Significance and Replication

Higher level scientific practices could also limit the inflation of statistically significant estimates in under-powered studies. As shown in our simulations, if estimates were not filtered by their statistical significance, even in low power studies causal inference methods would on average recover the true effect. The publication bias arising from dichotomizing evidence according to p-values has long been criticized in many disciplines but has seen a revival with the recent replication crises in psychology, medicine and social sciences. Many researchers advocate abandoning statistical significance as a measure of a study's quality (McShane et al. 2019). This would essentially eliminate the trade-off described in this paper.

To be effective, this change in attitude towards statistical significance should be paired with an effort to replicate studies (Christensen and Miguel 2018). Repli-

[4]In a future version of this paper, we will investigate two complementary approaches that could help address the potential inflation of statistically significant estimates. In a series of articles, Zwet and Gelman (2021), Zwet and Cator (2021) and Zwet et al. (2021) present a Bayesian procedure to shrink statistically significant estimates based on a corpus of estimates from prior studies. The second approach consists in carrying out quantitative bias analyses to evaluate whether the threat of unobserved confounding requires a restrictive causal approach. Rosenbaum (2002), Oster (2019) and Cinelli and Hazlett (2020) have developed different methods to run sensitivity analyses. They could be paired with power calculations to better evaluate the hidden bias / power trade-off of competing research designs. For instance, if a sensitivity analysis reveals that a simple selection on observables strategy is very robust to omitted variable bias, one may avoid using an IV model since it has a higher chance to produce inflated estimates.

cations, even of low powered studies, would eventually enable to build the distribution of the causal estimand of interest. Subsequent meta-analyzes would reduce the uncertainty around the true value of the causal estimand by pooling estimates (Hernán 2021).

Finally, the inflation of statistically significant estimates can be limited by considering confidence intervals as compatibility intervals (Shadish et al. 2002, Amrhein et al. 2019, Romer 2020). The width of these intervals gives a range of effect sizes compatible with the data. Confidence intervals will be wide in underpowered studies signaling that point estimates should not be taken at face value, even if statistically significant.

3 Conclusion

Causal identification strategies have undoubtedly participated in making empirical analyses more credible (Angrist and Pischke 2010). To avoid confounding, they only exploit the exogenous portion of the variation. In this paper, we argue that the same aspect that makes causal identification strategies credible can create another type of bias. Not only the lack of precision makes it more difficult to precisely get a sense of the magnitude of the actual effect but it also increases the probability of published estimates to be inflated. The confounding / exaggeration trade-off we highlight in this paper manifests itself along different dimensions for each identification strategy. A systematic reporting of *pre* and *post* analysis power calculations in observational studies would help gauge the risk of falling into this low power trap.

References

Abadie, Alberto. Statistical Nonsignificance in Empirical Economics. *American Economic Review: Insights*, 2(2):193–208, June 2020. ISSN 2640-205X, 2640-2068. doi: 10.1257/aeri.20190252.

Amrhein, Valentin and Trafimow, David and Greenland, Sander. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, 73(sup1):262–270, March 2019. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.2018.1543137.

Andrews, Isaiah and Kasy, Maximilian. Identification of and Correction for Publication Bias. *American Economic Review*, 109(8):2766–2794, August 2019. ISSN 0002-8282. doi: 10.1257/aer.20180310.

Angrist, Joshua D. and Pischke, Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, 1 edition edition, January 2009. ISBN 978-0-691-12035-5.

Angrist, Joshua D. and Pischke, Jörn-Steffen. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2):3–30, June 2010. ISSN 0895-3309. doi: 10.1257/jep.24.2.3.

Angrist, Joshua D. and Pischke, Jörn-Steffen. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, December 2014. ISBN 978-1-4008-5238-3.

Athey, Susan and Imbens, Guido. The Econometrics of Randomized Experiments. *arXiv:1607.00698 [econ, stat]*, July 2016.

Black, Bernard S. and Hollingsworth, Alex and Nunes, Leticia and Simon, Kosali Ilayperuma. Simulated Power Analyses for Observational Studies: An Application to the Affordable Care Act Medicaid Expansion. SSRN Scholarly Paper ID 3368187, Social Science Research Network, Rochester, NY, March 2021.

Brodeur, Abel and Cook, Nikolai and Heyes, Anthony. Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, 110(11):3634–3660, November 2020. ISSN 0002-8282. doi: 10.1257/aer.20190687.

Button, Katherine S. and Ioannidis, John P. A. and Mokrysz, Claire and Nosek, Brian A. and Flint, Jonathan and Robinson, Emma S. J. and Munafò, Marcus R. Power failure: Why small sample size undermines the reliability of neuro-

science. *Nature Reviews Neuroscience*, 14(5):365–376, May 2013. ISSN 1471-0048. doi: 10.1038/nrn3475.

Camerer, Colin F. and Dreber, Anna and Forsell, Eskil and Ho, Teck-Hua and Huber, Jürgen and Johannesson, Magnus and Kirchler, Michael and Almenberg, Johan and Altmejd, Adam and Chan, Taizan and Heikensten, Emma and Holzmeister, Felix and Imai, Taisuke and Isaksson, Siri and Nave, Gideon and Pfeiffer, Thomas and Razen, Michael and Wu, Hang. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, March 2016. doi: 10.1126/science.aaf0918.

Card, David. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. Working Paper 4483, National Bureau of Economic Research, October 1993.

Cattaneo, Matias D. and Titiunik, Rocío and Vazquez-Bare, Gonzalo. Power calculations for regression-discontinuity designs. *The Stata Journal: Promoting communications on statistics and Stata*, 19(1):210–245, March 2019. ISSN 1536-867X, 1536-8734. doi: 10.1177/1536867X19830919.

Chang, Andrew C. and Li, Phillip. Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not". *Critical Finance Review*, 11, July 2022. ISSN 2164-5744, 2164-5760. doi: 10.1561/104.00000053.

Christensen, Garret and Miguel, Edward. Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, 56(3):920–980, September 2018. ISSN 0022-0515. doi: 10.1257/jel.20171350.

Cinelli, Carlos and Hazlett, Chad. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020. ISSN 1467-9868. doi: 10.1111/rssb.12348.

Cunningham, Scott. *Causal Inference: The Mixtape*. Yale University Press, January 2021. ISBN 978-0-300-25588-1 978-0-300-25168-5. doi: 10.2307/j.ctv1c29t27.

Deaton, Angus and Cartwright, Nancy. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, August 2018. ISSN 0277-9536. doi: 10.1016/j.socscimed.2017.12.005.

Dehejia, Rajeev H. and Wahba, Sadek. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999. ISSN 0162-1459. doi: 10.2307/2669919.

Duflo, Esther and Glennerster, Rachel and Kremer, Michael. In . , *Handbook of Development Economics*, volume 4, pages 3895–3962. Elsevier, January 2007. doi: 10.1016/S1573-4471(07)04061-2.

Freeman, G. and Cowling, B. J. and Schooling, C. M. Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *International Journal of Epidemiology*, 42(4):1157–1163, August 2013. ISSN 0300-5771, 1464-3685. doi: 10.1093/ije/dyt110.

Gelman, Andrew. *Regression and Other Stories*. Cambridge University Press, Cambridge New York, NY Port Melbourne, VIC New Delhi Singapore, 2020. ISBN 978-1-107-67651-0.

Gelman, Andrew and Carlin, John. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6): 641–651, November 2014. ISSN 1745-6916. doi: 10.1177/1745691614551642.

Hernán, Miguel A. Causal analyses of existing databases: No power calculations required. *Journal of Clinical Epidemiology*, page S0895435621002730, August 2021. ISSN 08954356. doi: 10.1016/j.jclinepi.2021.08.028.

Hernán, Miguel A and Robins, James M. *Causal Inference: What If*. Boca raton: Chapman & hall/crc edition, 2020.

Hill, Jennifer L. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, January 2011. ISSN 1061-8600, 1537-2715. doi: 10.1198/jcgs.2010.08162.

Huntington-Klein, Nick. *The Effect: An Introduction to Research Design and Causality*. Chapman and Hall/CRC, Boca Raton, first edition, November 2021. ISBN 978-1-00-322605-5. doi: 10.1201/9781003226055.

Imbens, Guido and Kalyanaraman, Karthik. Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies*, 79(3):933–

959, 2012. ISSN 0034-6527.

Imbens, Guido W. and Rubin, Donald B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751.

Ioannidis, John P. A. Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5):640–648, 2008.

Ioannidis, John P. A. and Stanley, T. D. and Doucouliagos, Hristos. The Power of Bias in Economics Research. *The Economic Journal*, 127(605):F236–F265, October 2017. ISSN 0013-0133. doi: 10.1111/ecoj.12461.

Kasy, Maximilian. Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It. *Journal of Economic Perspectives*, 35(3):175–192, August 2021. ISSN 0895-3309. doi: 10.1257/jep.35.3.175.

Linden, Ariel. RETRODESIGN: Stata module to compute type-S (Sign) and type-M (Magnitude) errors. Boston College Department of Economics, October 2019.

Lu, Jiannan and Qiu, Yixuan and Deng, Alex. A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*, 72(1): 1–17, 2019. ISSN 2044-8317. doi: 10.1111/bmsp.12132.

McConnell, Brendon and Vera-Hernandez, Marcos. Going beyond simple sample size calculations: A practitioner's guide. Technical report, Institute for Fiscal Studies, September 2015.

McShane, Blakeley B. and Gal, David and Gelman, Andrew and Robert, Christian and Tackett, Jennifer L. Abandon Statistical Significance. *The American Statistician*, 73(sup1):235–245, March 2019. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.2018.1527253.

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

Oster, Emily. Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, April 2019. ISSN 0735-0015, 1537-2707. doi: 10.1080/07350015.2016.1227711.

Ravallion, Martin. Should the Randomistas (Continue to) Rule? Working Paper

27554, National Bureau of Economic Research, July 2020.

Romer, David. In Praise of Confidence Intervals. *AEA Papers and Proceedings*, 110: 55–60, May 2020. ISSN 2574-0768, 2574-0776. doi: 10.1257/pandp.20201059.

Rosenbaum, Paul R. *Observational Studies*. Springer Series in Statistics. Springer New York, New York, NY, 2002. ISBN 978-1-4419-3191-7 978-1-4757-3692-2. doi: 10.1007/978-1-4757-3692-2.

Rosenthal, Robert. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979. ISSN 1939-1455. doi: 10.1037/0033-2909.86.3.638.

Schell, Terry L. and Griffin, Beth Ann and Morral, Andrew R. Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study. Technical report, RAND Corporation, December 2018.

Shadish, William R. and Cook, Thomas D. and Campbell, Donald Thomas. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2002. ISBN 978-0-395-61556-0.

Stommes, Drew and Aronow, P. M. and Sävje, Fredrik. On the reliability of published findings using the regression discontinuity design in political science. *arXiv:2109.14526 [stat]*, September 2021.

Thistlethwaite, Donald L. and Campbell, Donald T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960. ISSN 1939-2176(Electronic),0022-0663(Print). doi: 10.1037/h0044319.

Timm, Andrew and Gelman, Andrew and Carlin, John. Retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors, March 2019.

Young, Alwyn. Leverage, Heteroskedasticity and Instrumental Variables in Practical Application. page 43, June 2021.

Zwet, Erik and Gelman, Andrew. A Proposal for Informative Default Priors Scaled by the Standard Error of Estimates. *The American Statistician*, pages 1–9, July 2021. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.2021.1938225.

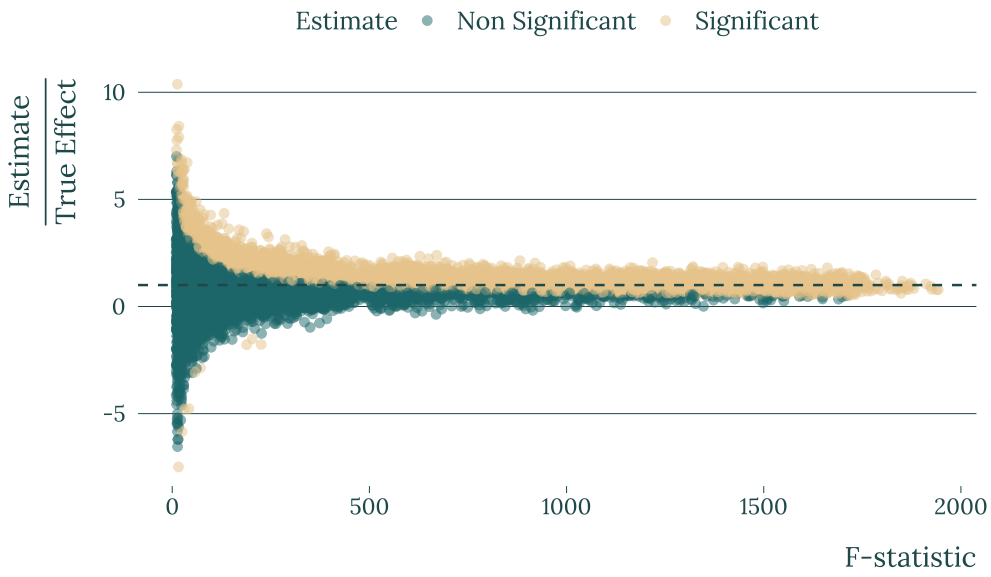
Zwet, Erik and Schwab, Simon and Senn, Stephen. The statistical properties of

RCTs and a proposal for shrinkage. *Statistics in Medicine*, 40(27):6107–6117, November 2021. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.9173.

Zwet, Erik W. and Cator, Eric A. The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, 75(4):437–452, November 2021. ISSN 0039-0402, 1467-9574. doi: 10.1111/stan.12241.

A Additional graphs

Figure 5: Bias in IV Simulations as a Function of the F-statistic, by Significance and Only for F-statistics Above 10



Notes: The green dots represent non statistically significant IV estimates. The beige one represent statistically significant ones. Details on the simulation are available at this link.