A Proposed Specification Check for P-Hacking

By Abel Brodeur, Nikolai Cook and Anthony Heyes*

There is growing evidence that some researchers engage in p-hacking – trying out several specifications and then selectively reporting those that produce statistically significant results (Brodeur et al. (2016); Brodeur, Cook and Heyes (2018)). This is an issue because studies that find a significant effect may be more likely to be published than studies with null results (Andrews and Kasy (2019)). Moreover, by choosing only estimates that are statistically significant, a researcher may paint an incomplete picture of the impact of a program or policy.

This paper proposes a specification check for p-hacking. More specifically, we advocate the reporting of t-curves and μ-curves - the t-statistics and estimated effect sizes derived from regressions using every possible combination of control variables from the researchers set - and introduce a standardized and accessible implementation. Our specification check allows researchers, referees and editors to visually inspect variation in effect sizes, significativity and sensitivity to the inclusion of control variables. We provide a Stata command which implements the specification check (available at the authors' webpages). Given the growing interest in estimating causal effects in the social sciences, the potential applicability of this specification check to empirical studies is very large.

As an illustrative example, we apply this specification check to study the returns to education using the NLSY-79 cohort.

* Brodeur: Department of Economics, University of Ottawa, 120 University, Social Sciences Building, Ottawa, Ontario K1N 6N5, Canada. E-mail: abrodeur@uottawa.ca. Cook: Department of Economics, University of Ottawa, 120 University, Social Sciences Building, Ottawa, Ontario K1N 6N5, Canada. E-mail: ncook@uottawa.ca. Heyes: Department of Economics, University of Ottawa, 120 University Private, Ottawa, Ontario, Canada, K1N 6N5 and University of Sussex, E-mail: anthony.heyes@uottawa.ca. Errors are ours.

The proposed methodology is related to meta-analysis and the detection of publication bias and p-hacking (Christensen and Miguel (2018); Doucouliagos and Stanley (2013); Leamer and Leonard (1983)). While meta-analysis focuses on the selective reporting of results *across* studies increasing overall research transparency, our method investigates selective reporting of results *within* paper, increasing the credibility of research. This paper also relates to methodologies testing how robust empirical results are to changes in possible controls (Simonsohn, Simmons and Nelson (2019); Young and Holsteen (2017)). See, for instance, Oster (2019) for a theory on coefficient movements after inclusion of controls and Imbens (2003) for an analysis of sensitivity in terms of partial R-squared values. Athey and Imbens (2015) also propose as a robustness measure to report the standard deviation of the point estimates over the set of models.

I. Specification Check

Suppose we are interested in estimating the causal effect of a policy or variable of interest in a setting with a large number of potential control variables. (We use the term variable of interest, treatment or intervention interchangeably and simplify the exposition by focusing on an example where there is only one treatment variable.) We observe N observations on an outcome variable, with one variable of interest (T) and n potential control variables X. In this setting, there are up to 2^n different regressions that the researcher might run, each defined by the inclusion of a distinct combination of controls.

A researcher may choose to present only a subset of specifications in their paper. P-hacking would occur if the researcher is more likely to present statistically significant estimates. P-hacking would also oc-

cur if the researcher collects or selects additional control variables until non-significant results become significant.

Our method computes a regression for each of the possible combinations of controls. For a horizon of n controls, the algorithm will compute 2^n regressions and record the associated treatment effect sizes and t-statistics. The program will present four graphs: (i) a t-curve, (ii) an effect size curve, (iii) the distribution of t-statistics by number of controls, and (iv) the distribution of effect sizes by number of controls.

The t-curve histogram displays the distribution of t-statistics of the estimated treatment effect. (A reference line at $p = 0.05$ or $z = 1.96$ is provided, indicating a customary significance threshold.) A t-curve that is wholly to the right of the threshold indicates that regardless of control variables included, the treatment effect remains statistically significant at conventional levels. The variance of the distribution displays how sensitive the statistical significance of the treatment effect is to control combination. Similarly, we present an estimated effect size curve. The dispersion of the estimates indicates how much the size of the treatment effect, rather than its statistical significance, varies by control combination.

The third and fourth graphs present a box plot for each control combination horizon for t-statistics and effect sizes, respectively. These plots illustrate the distribution of t-statistics (or effect size) by number of control variables included. The leftmost box plot shows the t-statistic (or effect size) for the coefficient of interest in the specification with no control variable. The rightmost box plot shows the t-statistic (or effect size) for the coefficient of interest in the specification with the full set of control variables. Any of the inner box plots illustrates the distribution of t-statistics (or effect size) for the n regressions with the specified number of control variables.

Why might this graphical specification check be an attractive way to go compared to, for example, showing that the results are robust to the exclusion or inclusion of specific control variables? One important reason is visual, with this specification check illustrating variation in effect sizes, significance and sensitivity to the inclusion of control variables. Within these figures, there is more information provided about the robustness of the treatment effect than in a comparably-sized results table - the whole set of possible inclusions and exclusions are displayed in a single figure. Second, the specification check could help researchers recognize inadvertent p-hacking in their own work. An example noted by Denton (1985) is if there are multiple authors using the same data set, each may be unaware of what the others are doing, and hence there may be an extent of (unintentional) collective p-hacking.

A. Limitations

One limitation is that the researcher still chooses how to code their variables, e.g., creating a binary variable for marriage versus separate variables for married, divorced, separated, widowed and single. Another limitation of this specification check is that the researcher may still strategically choose the controls that are ultimately included. Nonetheless, we believe that the specification check is helpful. First, editors and referees could ask researchers to include other relevant control variables in the specification check or to code the variables in a different way. Second, this specification check illustrates the statistical significance of all potential sequences that the controls could have been included. In other words, it shows the robustness (or lack thereof) of results to the order in which the control variables are sequentially included.

II. Example

A. Data and Control Variables

We rely on data from the NLSY-79 cohort to estimate the returns to education. We begin by considering the standard Mincer regression of log wages on educational attainment. We control for experience (defined as age minus education years minus 6) and experience-squared in all specifications. Our objective is to check the robustness of

the estimate to the inclusion of the following additional control variables: four dummies for region of residence, x dummies for race, x dummies for marital status, x dummies for mother's education, x dummies for father's education, x dummies for mother's occupation, x dummies for father's occupation, and number of siblings.

B. Figure and Results

In Figure 1, clockwise from top-left we present the t-curve, effect size curve, t-statistic by number of included controls, and effect size by number of included controls, for the Mincer equation. All statistics refer to the treatment effect – the estimated effect of an additional year of education on wages.

In the t-curve, we show that regardless of the control variable combination or horizon, the t-statistic associated with years of education ranges from 29 to 39, well above conventional statistical significance levels. In the effect curve, we show that regardless of the control variable combination or horizon, an additional year of education increases earnings between 9.6 to 11.6%.

In the bottom left panel, the first box plot (a singleton) shows that the t statistic of education when no additional controls are included is 38.98. (The right panel first box plot shows the associated effect size of 11.5%). The second box plot shows the distribution of the t-statistic of years of education when only one of the seven possible controls are included. The median t-statistic is 38, with a minimum of 32. (The right panel second box plot shows the associated median effect size of 11.3% and a minimum of 10.2%).

As the number of controls increases the median t-statistic steadily decreases. With the full set of controls, the statistical significance of $t = 29.3$ is below the 5th percentile of the 256 estimates. The median effect size behaves similarly.

In this case, the statistical significance and estimated effect size fall almost monotonically with additional controls, neither is in danger of becoming (statistically or economically) insignificant.

If it were instead the case that a researcher must choose between non and just significant specifications, the incentives become much stronger to perhaps selectively or strategically sequence the introduction of control variables.

The inclusion of Figure 1 alongside the preferred results of the researcher would provide persuasive evidence to a reader that the results were not substantially sensitive to researcher choices.

As a note, these graphs should not be viewed as providing a test for causality, but as an opportunity to increase research transparency.

III. General Discussion

There is an increasing recognition that publication bias and p-hacking may contribute to a weakening of the credibility and reproducibility of results. The proposed specification check thus has the potential to strengthen the credibility of empirical research on which policies are based. We discuss the conditions under which this specification check is appropriate and explain how it can be used by authors and referees to test for the presence of p-hacking in their analyzes. Moreover, we provide a software to implement the specification check (available at the authors' webpages).

Last, we demonstrate the applicability of the specification check by studying the returns to education using the NLSY-79 cohort.

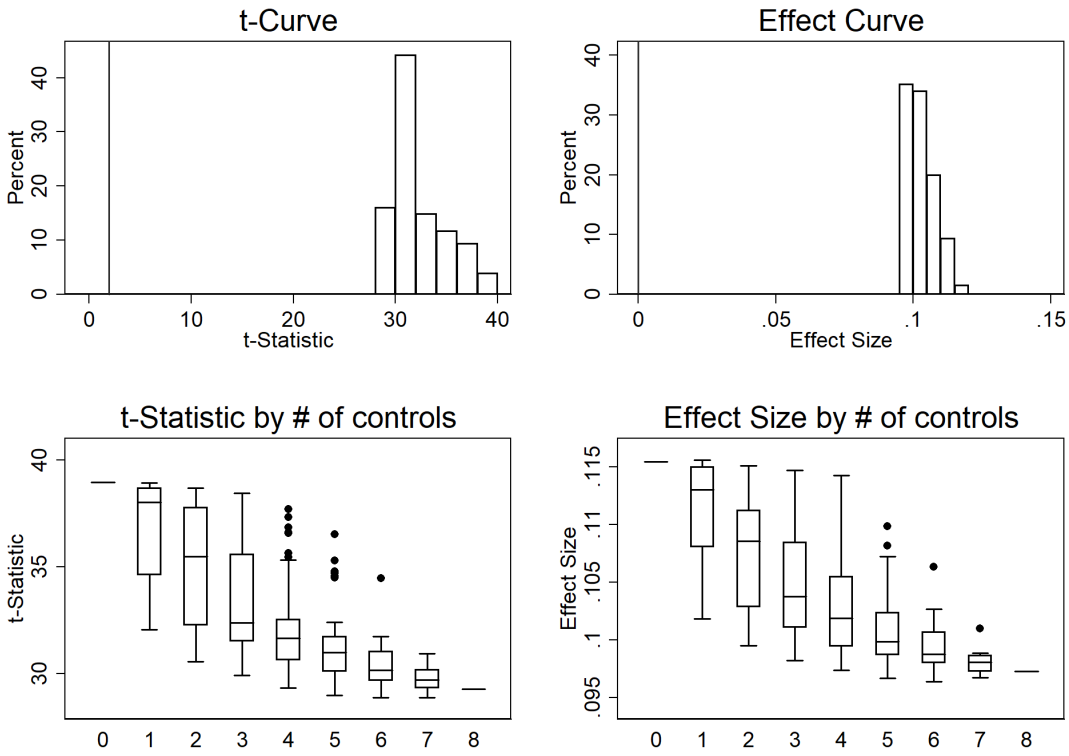
REFERENCES

Andrews, I., and M. Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review*, 109(8): 2766–94.

Athey, S., and G. Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review: Papers & Proceedings*, 105(5): 476–80.

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

FIGURE 1. SPECIFICATION CHECK



In the top left panel, we plot a histogram of t-statistics derived from the specification checker. In the top right, we display the distribution of the associated effect sizes. In the bottom left, we display box plots of the t-statistics by the number of coefficients included in the specifications. In the bottom right, we plot the associated effect sizes.

Brodeur, A., N. Cook, and A. Heyes. 2018. "Methods Matter: P-Hacking and Causal Inference in Economics." IZA Discussion Paper 11796.

Christensen, G., and E. Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature*, 56(3): 920–80.

Denton, F. T. 1985. "Data Mining as an Industry." *Review of Economics and Statistics*, 67(1): 124–27.

Doucouliagos, C., and T. D Stanley. 2013. "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity." *Journal of Economic Surveys*, 27(2): 316–339.

Imbens, G. W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *American Economic Review*, 93(2): 126–132.

Leamer, E. E., and H. Leonard. 1983. "Reporting the Fragility of Regression Estimates." *Review of Economics and Statistics*, 65(2): pp. 306–317.

Oster, E. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*, 37(2): 187–204.

Simonsohn, U., J. P. Simmons, and L. D. Nelson. 2019. "Specification Curve: Descriptive and Inferential Statistics on all Reasonable Specifications." SSRN 2694998.

Young, C., and K. Holsteen. 2017. "Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis." *Sociological Methods & Research*, 46(1): 3–40.