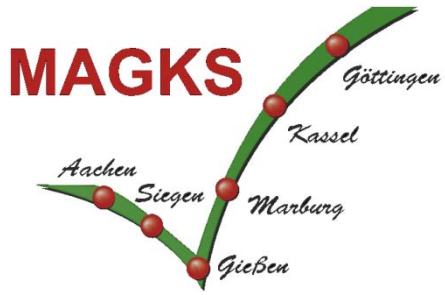

Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 03-2016

Stephan B. Bruns

The Fragility of Meta-Regression Models in Observational Research

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers/index_html%28magks%29

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

The Fragility of Meta-Regression Models in Observational Research

STEPHAN B. BRUNS

Meta-Research in Economics Group, University of Kassel, Nora-Platiel-Str. 5, 34109 Kassel, Germany.

(e-mail: bruns@uni-kassel.de)

Abstract

Many meta-regression analyses that synthesize estimates from primary studies have now been published in economics. Meta-regression models attempt to infer the presence of genuine empirical effects even if the authors of primary studies select statistically significant and theory-confirming estimates for publication. Meta-regression models were originally developed for the synthesis of experimental research where randomization ensures unbiased and consistent estimation of the effect of interest. Most economics research is, however, observational and authors of primary studies can search across different regression specifications for statistically significant and theory-confirming estimates. Each regression specification may possibly suffer from biases such as omitted-variable biases that result in biased and inconsistent estimation of the effect of interest. We show that if the authors of primary studies search for statistically significant and theory-confirming estimates, meta-regression models tend to systematically make false-positive findings of genuine empirical effects. The ubiquity of such search processes for specific results may limit the applicability of meta-regression models in identifying genuine empirical effects in economics.

Keywords: Meta-regression, meta-analysis, *p*-hacking, publication bias, omitted-variable bias, sampling variability, sampling error, Monte Carlo simulation

JEL classification: C12, C15, C40

Acknowledgements

I would like to thank Guido Bünstorf and David Stern for their helpful comments.

I. Introduction

Empirical research is often characterized by the selection of statistically significant results. It has been shown that published *p*-values cluster just below the widely used significance thresholds for the leading general-interest journals (Ridley, 2007), the top economics journals (Brodeur *et al.*, 2015), the top sociology journals (Gerber and Malhotra, 2008a), and the top political science journals (Gerber and Malhotra, 2008b). Meta-regression models try to address this selection bias by integrating the estimates from multiple primary studies in order to reveal the presence or absence of genuine effects. We show that meta-regression analyses of observational research suffer from a lack of robustness if primary authors search across different regression specifications for statistically significant and theory-confirming estimates. Therefore, using meta-regression models to make inferences on genuine effects in observational research may result in systematic false-positive findings of genuine effects.

We refer to experimental research if randomization is used to estimate an effect of interest, whereas observational research denotes research designs without randomization. While randomization ensures an unbiased and consistent estimate of the effect of interest, regression analyses based on observational data are characterized by a large analytical flexibility due to the multitude of potential regression specifications. Each regression specification may possibly suffer from biases such as omitted-variable biases resulting in a biased and inconsistent estimation of the effect of interest. This analytical flexibility was described as a key threat to the reliability of inferences in observational research (Hendry, 1980; Leamer, 1983; Sims, 1988).

There are many different labels to denote that the authors of primary studies may search for estimates that are theory-confirming and that provide a *p*-value that is below the common thresholds of 0.05 or 0.1. We follow the definition of Simonsohn *et al.* (2014) who coined the term "*p*-hacking" to denote the selection of statistically significant (or theory-confirming) estimates within each study while "publication bias" denotes the decreased publication rate of studies without statistically significant estimates (Rosenthal, 1979).

p-hacking is prevalent for both experimental and observational research and it is likely to be caused by the incentive system of academic publishing limiting the reliability of inferences that can be drawn from published empirical studies (Glaeser, 2006; Ioannidis, 2005). Empirical estimates have to be significant, but they also have to confirm the theory or hypothesis presented in the paper. Fanelli (2010) shows that the probability that a paper finds support for its hypothesis is high across all research disciplines. The pressure to provide significant and theory-confirming results is increased by declining acceptance rates in top journals and the need to publish in these journals in order to start or advance an academic career (Card and DellaVigna, 2013). As a result, Young *et al.* (2008) compare the publication process to the winner's curse in auction theory. The most spectacular or exaggerated results are rewarded with publication in the top journals, although in this case it is the scientific community rather than the author that is cursed.

In extreme cases, strong theoretical presumptions may lead authors to search for theory-confirming results (Card and Krueger, 1995). As soon as potentially false theories become established, empirical research may be characterized by the selection of results that meet the anticipated expectations of reviewers (Frey, 2003) rather than those that falsify the false theory. Null results may only be considered for publication if a series of articles previously established the presence of a genuine effect (De Long and Lang, 1992).

The combination of flexible observational research designs in economics and incentives to select for specific results may introduce severe biases in published empirical findings. Experimental sciences improve the reliability of inferences by using meta-analyses that integrate the evidence of multiple

studies while controlling for *p*-hacking (e.g. Sutton *et al.*, 2000). Such meta-analytic tools are increasingly being used to synthesize observational research in economics. The Precision-Effect Test (PET) that relates the *t*-value of an estimated regression coefficient to the precision of the estimate (Stanley, 2008) is commonly used (e.g. Adam *et al*., 2013; Efendic *et al.*, 2011). If a genuine effect is present, the coefficient's *t*-value and its precision are associated and this relation is used to test for the presence of genuine effects. However, such an association between a coefficient's *t*-value and its precision might also occur in the absence of a genuine effect due to omitted-variable bias.

Stanley's (2008) pioneering simulation results suggest that PET is largely robust to the presence of omitted-variable biases in primary studies and, consequently, genuine empirical effects are supposed to be reliably identified. This paper extends these simulation results and offers a note of caution regarding the reliability of meta-regression models in identifying genuine empirical effects in observational research. We show that inferences on genuine effects by PET are fragile in the presence of *p*-hacking based on omitted-variable biases, i.e. authors of primary studies search across different regression specifications for statistically significant and theory-confirming estimates. We discuss theoretically and show by means of simulations that PET provides systematically false-positive findings of genuine effects that are caused by omitted-variables biases in the primary literature.

PET can also be extended using dummy variables that measure the presence or absence of control variables in the primary studies to filter out omitted-variable biases. However, most meta-regression analyses in economics control for only a small degree of variation in primary regression specifications (e.g. Adam *et al*., 2013; Efendic *et al.*, 2011). One reason for this practice of meta-regression analysis may be the belief that PET is largely robust to omitted-variable biases based on the simulation results by Stanley (2008). More importantly, however, may be the high degree of heterogeneity of regression specifications across primary studies in empirical economics. Publication requires novelty and this is often achieved by modifying the set of control variables. This high degree of heterogeneity of regression specifications may often make it impossible for the meta-regression model to control for the variation of regression specifications.

Our findings suggest that the applicability of meta-regression models to identify genuine empirical effects in observational research in economics may be limited by the fragility of PET with respect to biases such as omitted-variable biases and the large heterogeneity of regression specifications in empirical economics that is difficult or even impossible to control for in a meta-regression model.

Section 2 presents *p*-hacking in observational research in economics and section 3 discusses the lack of robustness of meta-regression models in the presence of *p*-hacking that is based on omitted-variable biases. Section 4 provides evidence from a Monte Carlo simulation, section 5 discusses the limits of meta-regression analysis in economics, and section 6 concludes.

II. *p*-hacking in observational research

The majority of empirical economic research uses the regression framework to estimate conditional associations between variables that stem from observational data. This research design is characterized by a high degree of flexibility and, as a result, the corresponding range of obtainable estimates is wide (Hendry, 1980; Leamer, 1983; Sims, 1988). Sources of this flexibility include the choice of estimation techniques, functional forms, variable definitions, and, in particular, the sets of control variables included. Variations in the sets of control variables may introduce omitted-variable biases in the estimates of the effect of interest. This flexibility in research designs eases the search for statistically significant or theory-confirming results. To illustrate this, suppose a theory that states *x* causes *y* and the corresponding data generating process (DGP) is:

$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{Z}\boldsymbol{\delta}' + \boldsymbol{\epsilon} \quad (1)$$

where β is the coefficient of interest, $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_h]$ is a vector of h control variables with coefficients $\boldsymbol{\delta} = [\delta_1, \dots, \delta_h]$, and $E[\boldsymbol{\epsilon}|\boldsymbol{x}, \boldsymbol{Z}] = 0$. Let us define $\boldsymbol{Z}_q$ as a subsample of q variables from $\boldsymbol{Z}$ and $\boldsymbol{Z}_p$ as the complement to $\boldsymbol{Z}_q$ so that each variable of $\boldsymbol{Z}$ is either in $\boldsymbol{Z}_q$ or $\boldsymbol{Z}_p$. Let $\boldsymbol{\delta}_q$ and $\boldsymbol{\delta}_p$ be the corresponding coefficients of $\boldsymbol{Z}_q$ and $\boldsymbol{Z}_p$, respectively. Consider $s = 1, \dots, k$ independent studies estimating the following regressions:

$$\boldsymbol{y}_s = \hat{\alpha}_s + \hat{\beta}_s\boldsymbol{x}_s + \boldsymbol{Z}_{qs}\widehat{\boldsymbol{\delta}}'_{qs} + \hat{\boldsymbol{\epsilon}}_s, \quad (2)$$

where $\hat{\alpha}_s, \hat{\beta}_s$, and $\widehat{\boldsymbol{\delta}}_{qs}$ are the estimates of α, β, and $\boldsymbol{\delta}_{qs}$ for study s. The set of control variables $\boldsymbol{Z}_{qs}$ may be specific to study s as may be the utilized sample. Some studies may use the same $\boldsymbol{Z}_{qs}$ or the same sample, but in general we observe a variety of $\boldsymbol{Z}_{qs}$ and samples across studies.

Suppose primary authors search for positive and significant $\hat{\beta}_s$. In such a case, only a biased subset of $\hat{\beta}_s$ may be observable in the published literature, whereas a potentially large amount of $\hat{\beta}_s$ estimated in the process of conducting research remains unpublished. Let us define $\hat{b}_s = \left[\hat{\beta}_s, \widehat{\boldsymbol{\delta}}_{qs}\right]$ and $\boldsymbol{Q}_s = \left[\boldsymbol{x}_s, \boldsymbol{Z}_{qs}\right]$, then $\hat{b}_s$ in the presence of *p*-hacking is given by:

$$E\left[\hat{b}_s \middle| PH\right] = b + E\left[(\boldsymbol{Q}'_s\boldsymbol{Q}_s)^{-1}\boldsymbol{Q}'_s\boldsymbol{Z}_{ps}\boldsymbol{\delta}'_{ps} \middle| PH\right] + E[(\boldsymbol{Q}'_s\boldsymbol{Q}_s)^{-1}\boldsymbol{Q}'_s\boldsymbol{\epsilon}|PH] \quad (3)$$

where *PH* denotes *p*-hacking for positive and significant $\hat{\beta}_s$. Equation (3) illustrates two important sources of *p*-hacking in observational research. First, authors may vary the set of control variables, $\boldsymbol{Z}_{qs}$, and by this the set of omitted variables, $\boldsymbol{Z}_{ps}$. The specification searching results in potential omitted-variable biases for $\hat{b}_s$ that are given by $E\left[(\boldsymbol{Q}'_s\boldsymbol{Q}_s)^{-1}\boldsymbol{Q}'_s\boldsymbol{Z}_{ps}\boldsymbol{\delta}'_{ps} \middle| PH\right]$ in (3). We are primarily interested in its first entry, that is the potential omitted-variable bias of $\hat{\beta}_s$. Such a search across different regression specifications need not even be a deliberate manipulation of the estimate of β, but might result from naive and unconscious experimentation with the data or from dealing with limited data availability. Leamer (1983) highlights this search process as the key source of the low credibility of observational research. Omitted-variable biases result in biased and inconsistent estimation of β and we denote the use of omitted-variable biases to obtain statistically significant and theory-confirming estimates as *p*-hacking based on omitted-variable biases.

We focus here on *p*-hacking based on omitted-variable biases as varying the set of control variables is likely to represent an important source of bias when the authors of primary studies search for statistically significant and theory-confirming estimates. However, our results can be easily generalized as other types of bias also result in biased and inconsistent estimation of the effect of interest. We discuss this in Section V.

Second, authors may vary the utilized sample, e.g. by using subsamples or by deleting "outliers", to select positive and significant $\hat{\beta}_s$ from those estimates offered by sampling variability. Each sample implies a sampling error that may render $\hat{\beta}_s$ positive and significant by chance. If authors systematically use sampling errors to obtain positive and significant $\hat{\beta}_s$, the published $\hat{\beta}_s$ are characterized by an association between $\boldsymbol{x}$ and the true error, $\boldsymbol{\epsilon}$, in (1). As a result, we can expect the first entry of $E[(\boldsymbol{Q}'_s\boldsymbol{Q}_s)^{-1}\boldsymbol{Q}'_s\boldsymbol{\epsilon}|PH]$ in (3) to be positive.[1] In an extreme case, we may observe only

[1] If sampling errors are not systematically used to select positive and significant $\hat{\beta}_s$, we can expect that $E[(\boldsymbol{Q}'_s\boldsymbol{Q}_s)^{-1}\boldsymbol{Q}'_s\boldsymbol{\epsilon}] = \boldsymbol{0}$. Note that simultaneity also results in some entries of $E[(\boldsymbol{Q}'_s\boldsymbol{Q}_s)^{-1}\boldsymbol{Q}'_s\boldsymbol{\epsilon}]$ differing from zero.

those $\hat{\beta}_s$ that are positive and significant by chance, whereas the 95% insignificant $\hat{\beta}_s$ as well as the 2.5% negative and significant $\hat{\beta}_s$ remain in the file-drawer (Rosenthal, 1979). *p*-hacking based on sampling error is well discussed in meta-analyses of experimental studies. For these research designs randomization ensures unbiased and consistent estimation of the effect of interest and sampling variability may be the main source of bias. However, a major source of bias in observational research is likely to be omitted variables.

If a primary literature is distorted by *p*-hacking for statistically significant and positive $\hat{\beta}_s$, we may learn little about β by using a simple average of the published $\hat{\beta}_s$. Meta-regression models synthesize the published $\hat{\beta}_s$ and aim to identify the genuine β while controlling for *p*-hacking.

III. Meta-regression models

Basic Model

The basic meta-analysis model[2] is:

$$\hat{\beta}_s = \omega_B + u_s \tag{4}$$

where $\hat{\beta}_s$ are the estimated coefficients of interest of study $s = 1, \dots, k$ and $u_s = N(0, v_s^2)$ with v_s^2 as the sampling variance of β_s. The basic meta-analysis model is estimated by weighted least squares (WLS) where the weights are equal to the inverse variance of $\hat{\beta}_s$. This weighting procedure gives smaller weights to imprecisely estimated β_s and larger weights to more precisely estimated β_s (e.g. Sutton *et al.*, 2000). $H_0: \omega_B = 0$ tests for a non-zero weighted mean of $\hat{\beta}_s$. The Basic Model does not control for *p*-hacking. Therefore, if authors of primary studies opt for statistically significant or theory-confirming $\hat{\beta}_s$ by using sampling errors or omitted-variable biases, the weighted mean becomes biased.

An alternative to the Basic Model is the use of random-effects models (e.g. Sutton *et al.*, 2000). However, meta-regression analyses in economics are usually augmented by many control variables in the meta-regression. As random-effects models require strict exogeneity, most meta-regression analyses in economics focus on the use of fixed effects (e.g. Adam *et al.*, 2013; Efendic *et al.*, 2011).

Precision-Effect Test

The main meta-regression model that has been used in economics is the Egger et al. (1997) regression:

$$t_s = \delta + \omega_{PET} \frac{1}{\widehat{se}_s} + e_s \tag{5}$$

where t_s is the t-value of $\hat{\beta}_s$ and $\widehat{se}_s$ is the estimated standard error of $\hat{\beta}_s$. Stanley (2008) suggests $H_0: \omega_{PET} = 0$ to test for the presence of a genuine effect in a given primary literature and names this test "Precision-Effect Test" (PET).

If the correct primary regression specification is used $\hat{\beta}_s$ is an unbiased and consistent estimate of β; similar to the unbiased and consistent estimates obtained by randomization in experimental research

[2] The basic meta-analysis model is known in the meta-analysis literature as fixed-effects model. We refer to this model as the Basic Model to avoid confusion with the terminology from the econometrics of panel data.

designs. If the correct primary regression is used in all primary studies and a genuine effect is absent ($\beta = 0$), there should be no relation between t_s and $1/\widehat{se}_s$.

We can further expect that $\hat{\beta}_s$ is a precise estimate of β if the sample size is large and the estimated standard error ($\widehat{se}_s$) is small. On the contrary, if the sample size is small and $\widehat{se}_s$ is large, sampling variability yields a wide range of $\hat{\beta}_s$. *p*-hacking that is based on the use of sampling errors requires large $\hat{\beta}_s$ for large $\widehat{se}_s$ and small $\hat{\beta}_s$ for small $\widehat{se}_s$to ensure positive and statistically significant $\hat{\beta}_s$, i.e. $t_s \gtrsim 1.96$. Therefore, t_s is again unrelated to $1/\widehat{se}_s$ if *p*-hacking is only based on sampling errors.

If a genuine effect is present ($\beta \neq 0$) and the correct primary regression specification is used by all primary studies, t_s and $1/\widehat{se}_s$ are associated. As a result, $H_0: \omega_{PET} = 0$ can be used to infer the presence of genuine effects if all primary studies use the correct primary regression specification.

However, observational research is characterized by a large heterogeneity of regression specifications across primary studies potentially implying omitted-variable biases. Omitted-variable biases result in biased and inconsistent estimation of β and, thus, a relation between t_s and $1/\widehat{se}_s$ is introduced, exactly as it is for a genuine effect. Therefore, the relation between t_s and $1/\widehat{se}_s$ cannot distinguish between genuine effects and omitted-variable biases and $H_0: \omega_{PET} = 0$ can no longer serve as a test for the presence of genuine effects if the authors of primary studies use *p*-hacking based on omitted-variable biases . Figure 1 illustrates the properties of PET.

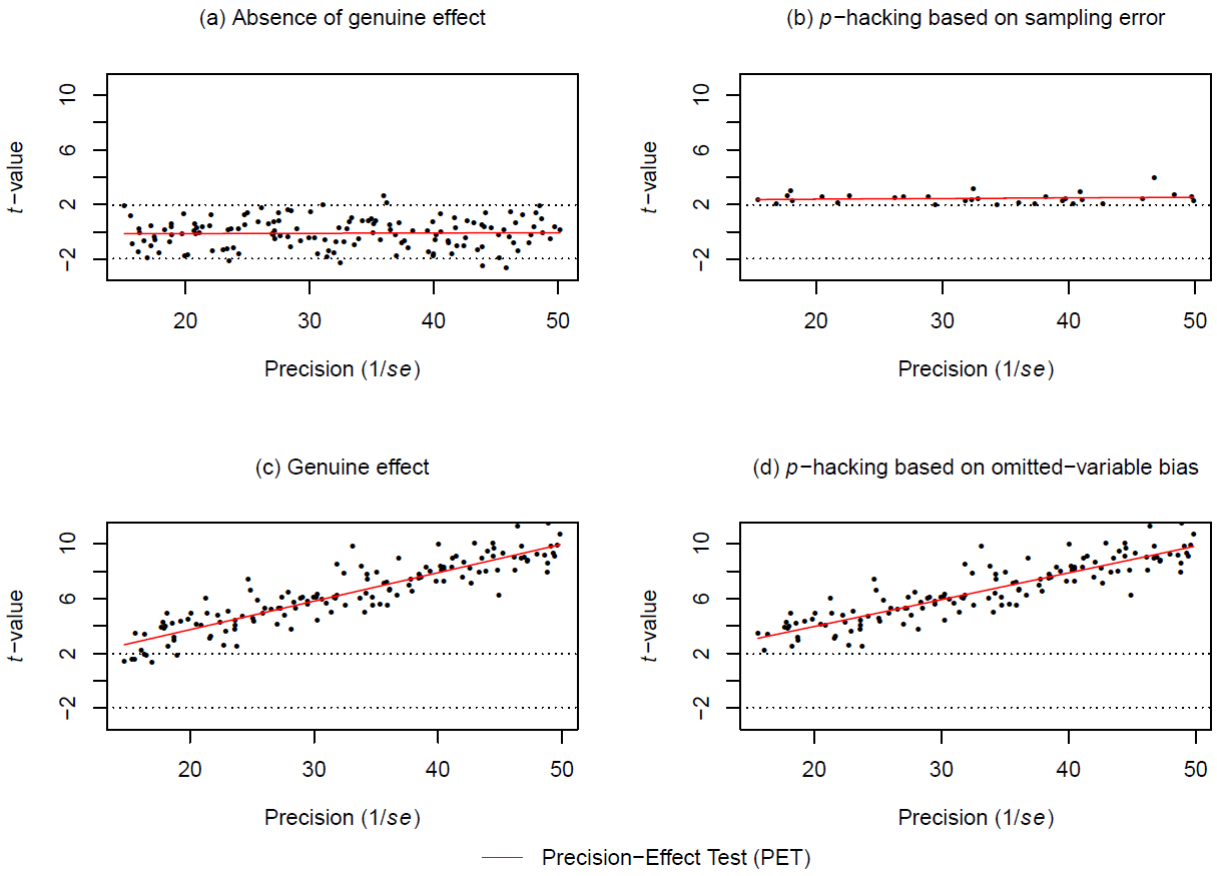

Figure 1. Properties of Precision-Effect Test. The dotted lines represent the (approximate) level of statistical significance (± 1.96). The dots are based on illustrative data and represent t-values of regression coefficients that can be obtained by primary studies (a) in the absence of a genuine effect and if all primary studies use the correct regression specification, (b) in the absence of a genuine effect and if all primary studies use the correct

regression specification but *p*-hacking based on sampling error, (c) in the presence of a genuine effect, and (d) in the absence of a genuine effect but in the presence of *p*-hacking based on omitted-variable bias.

In the presence of an omitted-variable bias, $\hat{\beta}_s$ is asymptotically biased and even with an infinitely small $\widehat{se}_s$ the resulting estimate would still be biased. PET captures biases that are associated with $\widehat{se}_s$ which is the case if sampling errors are used to select for statistically significant $\hat{\beta}_s$. In some cases, omitted-variable biases may also be associated with $\widehat{se}_s$ and, consequently, PET may reduce the false-positive findings of genuine effects in these cases compared to the Basic Model that does not control for any biases. Specifically, PET reduces false-positive findings of genuine effects if a positive omitted-variable bias goes in line with an increase in $\widehat{se}_s$ so that the change in $\hat{\beta}_s$ does not fully translate to a change in t_s.

In general, $\widehat{se}_s$ can increase or decrease if an omitted-variable bias is present. The variance of $\hat{\beta}_s$ for the correct regression specification is given by:

$$Var[\hat{\beta}_s|\boldsymbol{x}_s, \boldsymbol{Z}_s] = \frac{\hat{\sigma}^2_{\boldsymbol{Z}_s}}{(1 - \hat{R}_{\boldsymbol{x}_s\boldsymbol{Z}_s})\, S_{\boldsymbol{x}_s\boldsymbol{x}_s}} \tag{6}$$

where $\sigma^2_{\boldsymbol{Z}_s}$ is the estimated residual variance of primary study s, $\hat{R}^2_{\boldsymbol{x}_s\boldsymbol{Z}_s}$ is the R^2 of a regression of x_s on the set of control variables Z_s, and $S_{\boldsymbol{x}_s\boldsymbol{x}_s} = \sum(x_s - \bar{x}_s)$. If a control variable that is correlated with $\boldsymbol{x}_s$ and that has an own effect on y_s is dropped from the regression, the variance of $\hat{\beta}_s$ remains constant if $\hat{\sigma}^2_{\boldsymbol{Z}_s}$ and $(1 - \hat{R}_{x_s\boldsymbol{Z}_s})$ change proportionally. If $(1 - \hat{R}_{\boldsymbol{x}_s\boldsymbol{Z}_s})$ increases less (more) than proportionally to $\hat{\sigma}^2_{\boldsymbol{Z}_s}$, the variance of $\hat{\beta}_s$ increases (decreases). The change in the variance of $\hat{\beta}_s$ is based on the size of the coefficients and the covariances between the variables. We use Monte Carlo simulations to evaluate how PET performs if $\widehat{se}_s$ has a tendency to increase and decrease with the size of the omitted-variable bias.

IV. Monte Carlo simulation

Design

The Monte Carlo simulation analyses the robustness of PET and the Basic Model with respect to *p*-hacking that is based on omitted-variables biases and sampling errors. We consider the sample sizes of the meta-regression analyses as $k = 20, 40, 80, 160$ which is the number of primary studies that are synthesized by the meta-regression. These sample sizes reflect typical sample sizes that can be observed in meta-regression analyses in economics (Appendix I in Doucouliagos and Stanley (2013) gives an overview).

The primary study sample size of the sth study with $s = 1, \ldots, k$ is drawn from a gamma distribution with scale parameter equal to $\sigma^2/(\mu - 30)$ and shape parameter equal to $(\mu - 30)^2/\sigma^2$. Thus, μ denotes the mean of the primary study sample size distribution and σ^2 its variance. We round the obtained value for the primary study sample size to the next integer and add 30 so that 30 is the smallest primary study sample size. The choice of the scale and shape parameters allows us to vary μ and σ^2 independently. The use of a gamma distribution provides right-skewed primary study sample size distributions for small μ and increasingly symmetric ones for larger μ. We consider $\mu = 100, 200, 400$ and $\sigma^2 = 30^2, 60^2$. Figure 2 provides an overview of the sample size distributions mirroring primary literatures that are prevalent in empirical economic research.

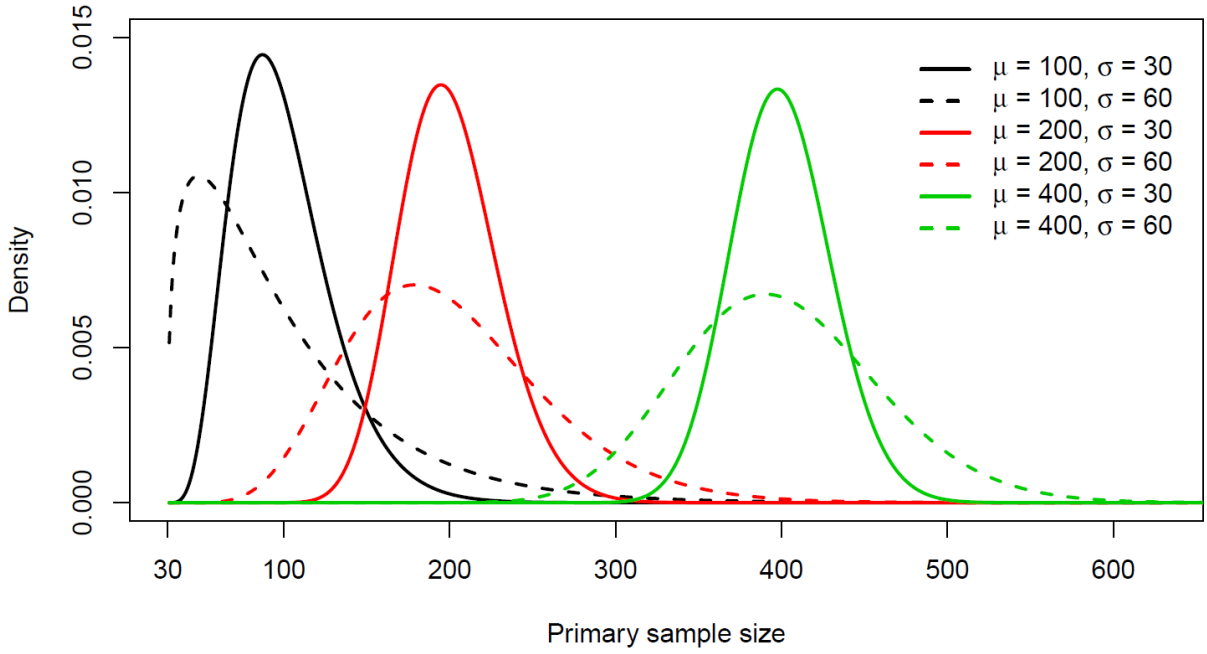

Figure 2. Primary sample size distributions.

The data-generating process (DGP) for each primary study s is given by:

$$y = \beta x + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \epsilon \quad (7)$$

where all right-hand side variables are $N(0,1)$ and $\beta = 0$ that is the coefficient of interest. We set $\gamma_1 = 0.2$, $\gamma_2 = 0.3$, and $\gamma_3 = 0.5$. Suppose primary authors intend to show that $\beta > 0$ at the 5% level of statistical significance using a two-sided t-test. The easiest way to search for an estimate of β that fulfils these criteria is to search across different regression specifications. We model this search for a positive and statistically significant β by first estimating the correct regression using x, z_1, z_2 and z_3. If the estimate of β is positive and significant by chance, the estimate is published. If these criteria are not fulfilled, the author of the primary study randomly drops z_1, z_2 or z_3 from the regression. If the estimate of β becomes positive and significant by using an omitted-variable bias, the estimate is published. If the thus obtained estimate is not positive and statistically significant, the author of the primary study adds the omitted variable and randomly drops one of the remaining two variables from the regression. If the estimate of β becomes positive and significant, the estimate is published. Finally, if the obtained estimate of β is not positive and statistically significant, the primary author adds the omitted variable again and drops the remaining control variable. If the estimate becomes positive and statistically significant, the estimate is published. If still no positive and significant estimate of β is obtained, the author of the primary study starts the same search across regression specifications for a different sample implemented by resampling all variables in equation (7).

The simulated process of p-hacking is based on a search across different regression specifications involving omitted-variable biases and sampling errors. We consider primary literatures where h% of the studies are affected by p-hacking, whereas $(100 - h)$% of the studies estimate the correct primary regression and publish the estimate of β irrespective of whether this estimate is positive and significant. We use $h = 10, 20, \ldots, 100$ to simulate primary literatures that range from the absence of

p-hacking to primary literatures that provide solely positive and statistically significant $\hat{\beta}$ though $\beta = 0$.

The simulated omitted-variable biases depend on the size of γ_1, γ_2 and γ_3 as well as on the covariance between the variables on the right-hand side in (7). We use two different cases to analyse the role of the covariance structure between the right-hand side variables in (7) on the robustness of PET and the Basic Model. Case I models a covariance structure between the variables that implies a variance of $\hat{\beta}$ that has a tendency to increase with the size of the omitted-variable bias:

$$Cov(x, z_1, z_2, z_3) = \begin{bmatrix} 1 & . & . & . \\ 0.5 & 1 & . & . \\ 0.5 & 0.2 & 1 & . \\ 0.5 & 0.2 & 0.2 & 1 \end{bmatrix}. \quad (8)$$

Case II models a covariance structure that implies a variance of $\hat{\beta}$ that has a tendency to decrease with the size of the omitted-variables bias:

$$Cov(x, z_1, z_2, z_3) = \begin{bmatrix} 1 & . & . & . \\ 0.4 & 1 & . & . \\ 0.5 & 0.2 & 1 & . \\ 0.6 & 0.2 & 0.2 & 1 \end{bmatrix}. \quad (9)$$

We analyse the rejection rate of $H_0: \omega_B = 0$ for the Basic Model and $H_0: \omega_{PET} = 0$ for PET for all 480 scenarios $\left(\#k * \#\mu * \#\sigma^2 * \#h * \#Cov(x, z_1, z_2, z_3)\right)$. As a genuine effect is absent ($\beta = 0$) these rejection rates are the type I errors of testing for genuine empirical effects in the presence of *p*-hacking.

Finally, Case III analyses the robustness of PET and the Basic Model with respect to *p*-hacking that is only based on sampling errors. For this scenario we use the simulation design described above, but the authors of primary studies always estimate the correct regression specification that is a regression of y on x, z_1, z_2 and z_3. If $\hat{\beta}$ is not positive and significant for the h% of studies that search for a positive and significant $\hat{\beta}$, the author of the primary study re-estimates the correct primary regression specification for a new sample. A new sample is generated by resampling all variables in equation (7) and the author continues to estimate the correct regression specification for new samples until a positive and significant $\hat{\beta}$ is obtained by chance. There are no omitted-variable biases involved in this case and *p*-hacking is only based on sampling errors. We focus on smaller primary sample size distributions in Case III as obtaining statistically significant estimates of β by chance becomes computationally intensive for large primary sample sizes.

Results

Type I errors of $H_0: \omega_B = 0$ and $H_0: \omega_{PET} = 0$ for Case I that uses the covariance structure (8) implying a tendency of increasing $\widehat{se}_s$ with the size of the omitted-variable bias are presented in Figure 3. The Basic Model neither controls for *p*-hacking that is based on sampling errors nor for *p*-hacking that is based on omitted-variable biases and, consequently, this model provides highly inflated type I errors even if the degree of *p*-hacking in the primary literature is small. PET also provides inflated type I errors, though they are smaller compared to the Basic Model. The type I errors increase with the primary sample sizes as sampling errors play a larger role for smaller primary sample sizes and PET controls for this type of bias. The type I errors of PET often show an inverted U-shape that is induced

by the decreasing number of statistically insignificant observations for lower levels of precision of the estimate of interest as the degree of *p*-hacking increases.

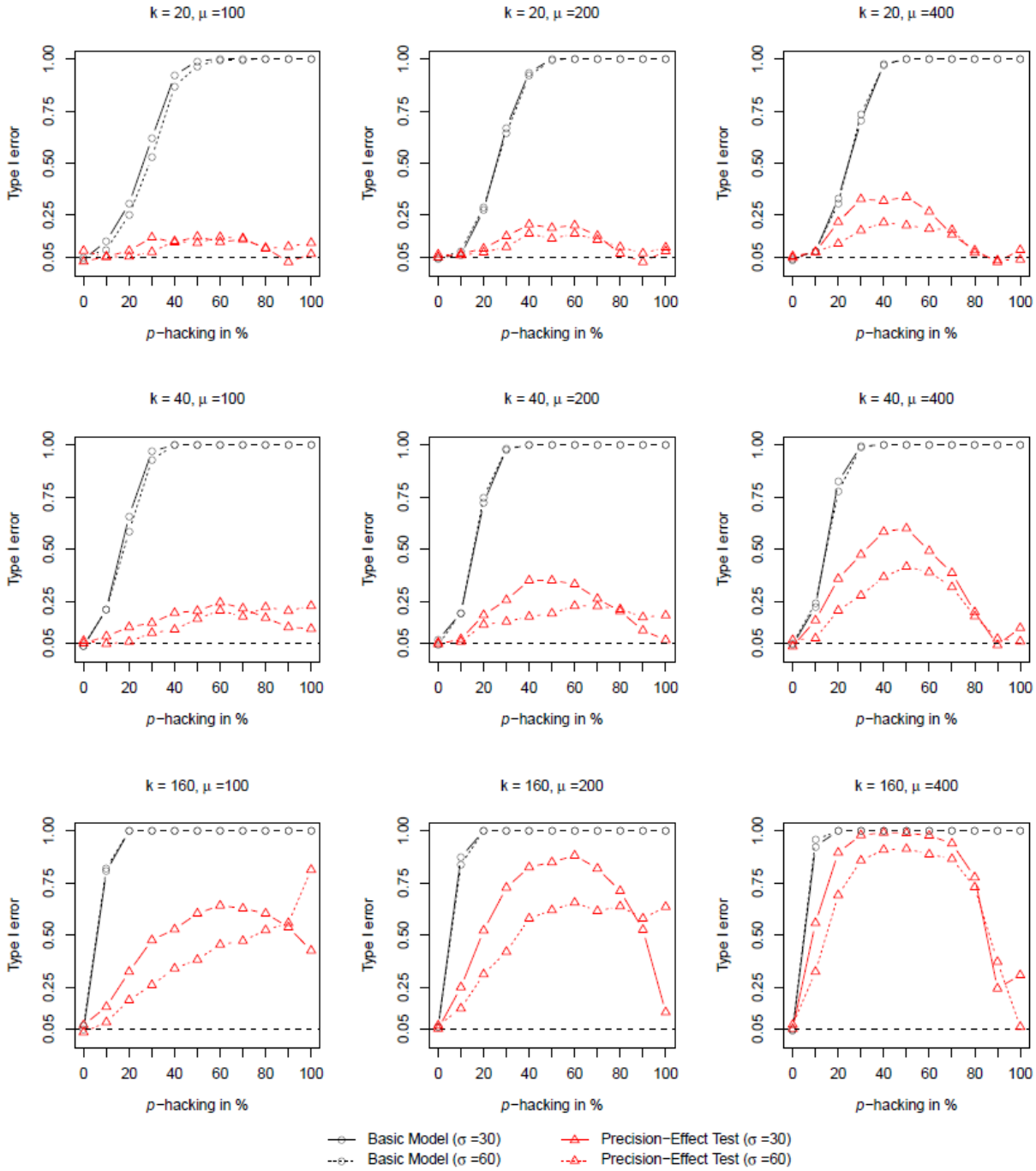

Figure 3. Type I errors of $H_0: \omega_B = 0$ and $H_0: \omega_{PET} = 0$ for Case I with a tendency of increasing $\widehat{se}_s$ with the size of the omitted-variable bias are presented for small to large meta sample sizes ($k = 20, 40, 160$) in combination with different primary sample size means $\mu = 100, 200, 400$ and primary sample size variances $\sigma^2 = 30^2, 60^2$.

Type I errors of $H_0: \omega_B = 0$ and $H_0: \omega_{PET} = 0$ for Case II that uses the covariance structure (9) implying a tendency of decreasing $\widehat{se}_s$ with the size of the omitted-variable bias are presented in Figure 4. The type I errors of PET are systematically larger than for Case I indicating that the positive association between $\widehat{se}_s$ and the size of omitted-variable biases that is modelled in Case I indeed

reduces the type I errors of PET. If the meta-sample size becomes large and the primary sample sizes are large, the type I errors of PET become as inflated as the type I errors of the Basic Model.

Finally, type I errors of $H_0: \omega_B = 0$ and $H_0: \omega_{PET} = 0$ for Case III that uses *p*-hacking based only on sampling errors are presented in Figure 5. In Case III the correct regression specification is estimated in all primary studies leading to unbiased and consistent estimates of β. PET can largely filter out this type of *p*-hacking whereas the Basic Model provides highly inflated type I errors.

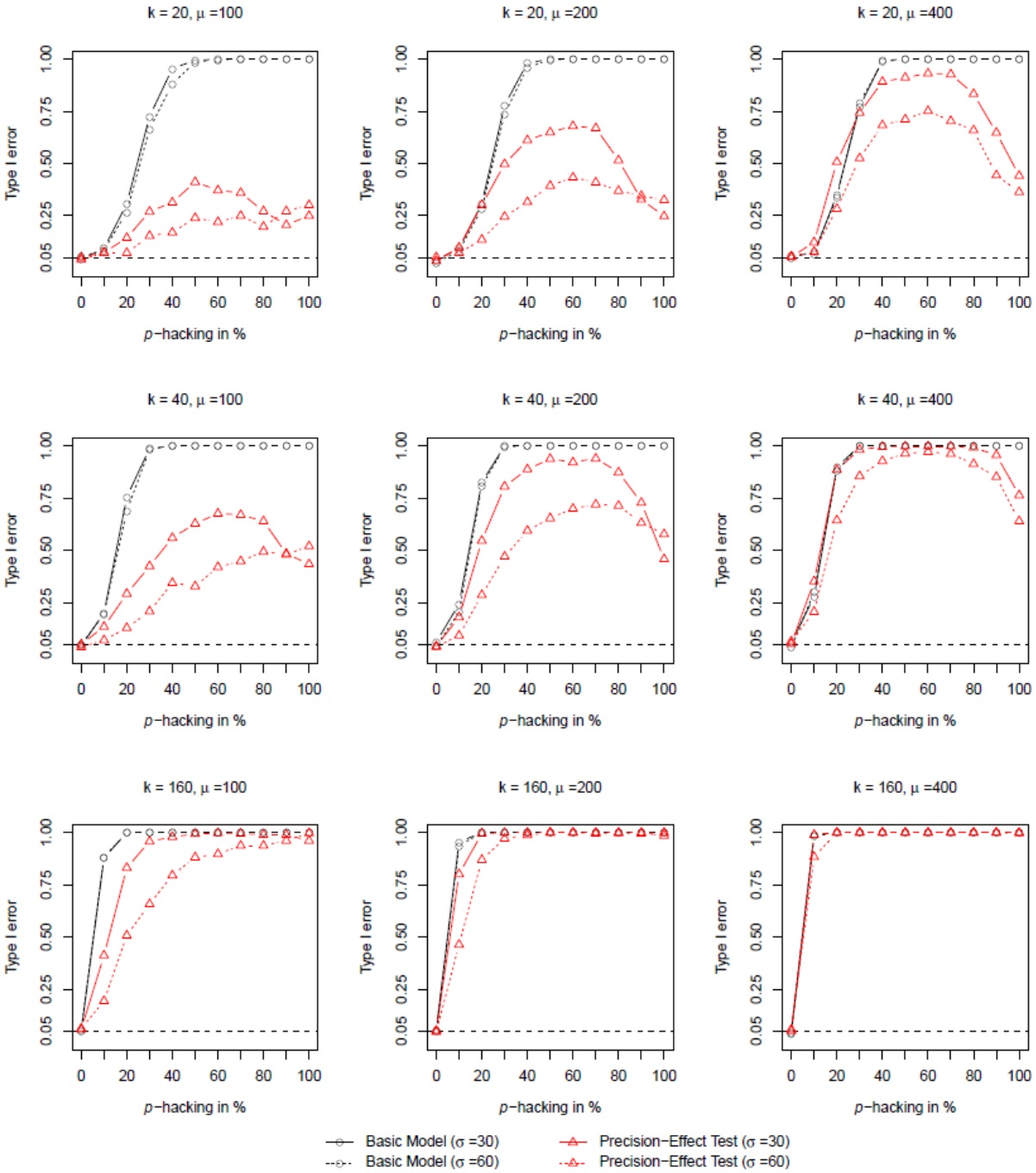

Figure 4. Type I errors of $H_0: \omega_B = 0$ and $H_0: \omega_{PET} = 0$ for Case II with a tendency of decreasing $\widehat{se}_s$ with the size of the omitted-variable bias are presented for small to large meta sample sizes $(k = 20, 40, 160)$ in combination with different primary sample size means $\mu = 100, 200, 400$ and primary sample size variances $\sigma^2 = 30^2, 60^2$.

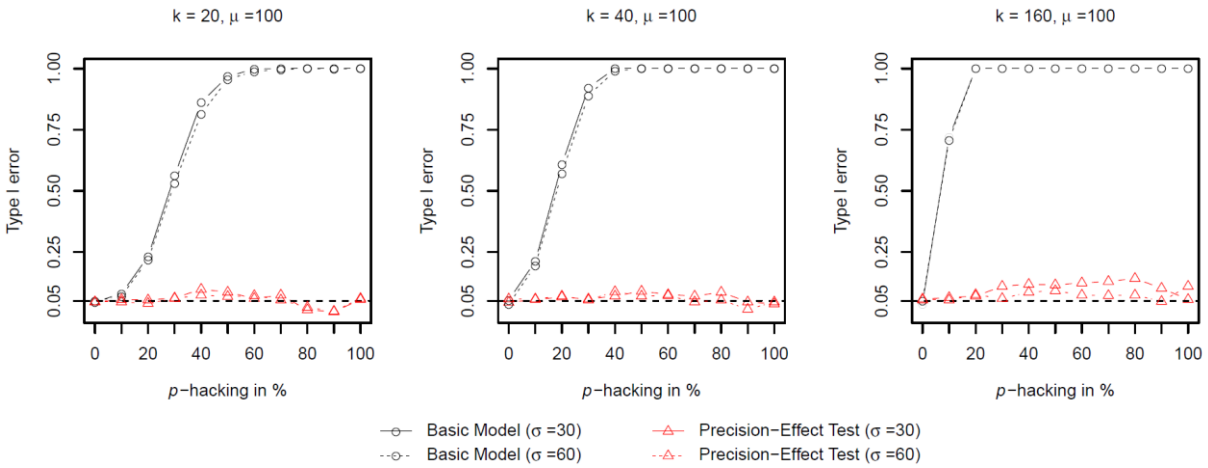

Figure 5. Type I errors of $H_0: \omega_B = 0$ and $H_0: \omega_{PET} = 0$ for Case III using *p*-hacking based only on sampling errors are presented for small to large meta sample sizes $(k = 20, 40, 160)$ in combination with a primary sample size mean of $\mu = 100$ and primary sample size variances $\sigma^2 = 30^2, 60^2$.

V. Discussion

We discuss theoretically and show by means of simulations that both the Basic Model and PET provide inflated type I errors if primary authors search across regression specifications for positive and statistically significant estimates. If the primary literature suffers from *p*-hacking that is only based on sampling errors implying the use of the correct regression specification for all primary studies, PET provides adequate type I errors and the Basic Model suffers from inflated type I errors. However, *p*-hacking in observational research in economics is more likely to be characterized by specification searching rather than by estimating the same regression specification for different samples. Therefore, even if the standard error of the estimate of interest increases with the size of the omitted-variable bias introducing an association between bias and standard error, PET provides inflated type I errors, though the type I errors are reduced compared to the Basic Model that does not control for any biases.

These findings extend the pioneering simulation results of Stanley (2008) suggesting that PET is largely robust to *p*-hacking that is based on omitted-variable biases and sampling errors. His simulation design is based on a DGP with two variables where the coefficient of the first variable is the coefficient of interest and the coefficient of the second variable is drawn from a normal distribution. Omitted-variable biases are simulated by primary studies estimating a regression with the first variable while the second variable is omitted from the regression. If the estimate of the first coefficient is not positive and statistically significant, all variables of the DGP are resampled and a new second coefficient is drawn that generates a new omitted-variable bias.

Our simulation design extends this simulation design in two respects that are important for the robustness of PET in testing for the presence of genuine effects. First, Stanley (2008) simulates *p*-hacking based on omitted-variable biases by resampling not only the second coefficient of the DGP to generate a new omitted-variable bias but also by resampling all variables of the DGP. If the variance of the normal distribution from which the second coefficient is selected is small, it requires multiple resamplings of the second coefficient to generate an omitted-variables bias that provides a positive and significant estimate of interest. However, each resampling of the second coefficient goes in line with a resampling of all variables of the DGP. This implies that primary studies that face an insignificant estimate of interest not only choose a new regression specification but simultaneously use a new sample that may by chance result in a positive and significant estimate of interest. Hence, in Stanley's

simulation design *p*-hacking based on omitted-variable biases that should actually mirror specification searching for desired estimates may be dominated by *p*-hacking that is based on sampling errors. Case III of our simulation illustrates that PET is robust to *p*-hacking that is only based on sampling errors. Consequently, a dominance of sampling errors over omitted-variable biases in the process of *p*-hacking is likely to result in adequate type I errors of PET. We model *p*-hacking as a search across different regression specifications and only if all of these specifications fail to provide a positive and significant estimate of interest does the primary study chose a new sample.

Second, Stanley (2008) generates omitted-variable biases by drawing the second coefficient of the DGP from a normal distribution with mean zero and by omitting the second variable from the estimated primary regressions. This implies potentially large positive and negative omitted-variable biases in the primary studies. As discussed above, a small variance of the normal distribution from which the second coefficient is drawn may result in a dominance of sampling errors over omitted-variable biases. If the variance of the normal distribution becomes larger, large positive and negative omitted-variable biases are present in the primary literature. Given that Stanley considers degrees of publication selection of 0%, 25%, 50%, and 75%, at least 25% percent of primary studies may have (large) negative omitted-variable biases. These negative and potentially significant estimates reduce the type I errors of PET. Such a primary literature is characterized by opposing estimates of the effect of interest rather than by a selection for theory-confirming results. In the presence of *p*-hacking for estimates that confirm a dominant theory, it may be unlikely to observe large contradicting estimates in the published literature. Even if two opposing theories are present in the literature, it is more likely that one implies a genuine effect and the other no effect rather than two genuine effects with opposing signs. We simulate empirical literatures that suffer from *p*-hacking for positive and significant estimates and those primary studies that are not affected by *p*-hacking do not publish large contradicting negative findings.

Both extensions of the simulation design result in what may be a more realistic simulation of *p*-hacking for specific estimates, i.e. theory-confirming and statistically significant estimates. Authors of primary studies may estimate many regression specifications and consciously or unconsciously choose those that fit their theoretical presumptions. In these cases the Basic Model and PET provide systematically false-positive findings of genuine effects.

Meta-regression analyses in economics widely use PET to integrate estimates of various primary studies. As the heterogeneity of study characteristics is often large, meta-analysts extend PET by adding dummy variables that are interacted with precision. These dummy variables may mirror various study characteristics including differences in variable definitions, considered time periods or subjects of investigation, estimation techniques, data sources, and the sets of primary control variables. Koetse *et al.* (2010) show that if the correct primary regression specification is known, using a dummy variable to control for an omitted variable can filter out the omitted-variable bias. Probably due to the large heterogeneity of regression specifications across different primary studies and the belief that PET has a certain robustness to omitted-variable biases, meta-regressions usually only add some dummy variables that control for omitted variables (e.g. Adam *et al.*, 2013; Efendic *et al.*, 2011). However, if the primary literature suffers from *p*-hacking for theory-confirming and statistically significant estimates, PET is likely to false-positively identify a genuine effect if the meta-regression model does not capture the heterogeneity introduced by different regression specifications.

In general, even if we assume that the meta-analyst knows the correct primary regression specification, dummy variables for all primary control variables that are omitted at least in one study may be required. If heterogeneity in the estimate of interest caused by omitted-variable biases remains, PET is likely to suffer from inflated type I errors. Given the large heterogeneity of regression specifications in

observational research in economics, it may be difficult or even impossible to cover this heterogeneity by dummy variables. This may limit the applicability of meta-regression analyses for the purpose of identifying genuine effects in observational research in economics.

Statisticians are well aware of the limits of meta-analysis in observational research. Becker and Wu (2007) and Wu and Becker (2012) present approaches to integrate regression slopes across primary studies. But these approaches require either the covariances of the regression coefficients of each primary study or the covariances of the variables used in the regression of each primary study. As both pieces of information are usually not available across primary studies in economics, these approaches are difficult to apply in empirical economic research.

We focus in this article on *p*-hacking that is based on omitted-variable biases to analyse the fragility of meta-regression models in identifying genuine empirical effects. However, our results can be easily generalized as the increased rate of false-positive findings of genuine effects stems from biased and inconsistent estimates in the primary literature. Such biased and inconsistent estimates introduce an association between the *t*-value of the estimate of interest and the precision of the estimate of interest that is interpreted by PET as evidence for a genuine effect. Omitted-variable biases are a prominent source of biased and inconsistent estimates, but other types of biases such as misspecifications of the functional form, simultaneity, and measurement errors may also result in biased and inconsistent estimates in the primary studies. If the authors of primary studies include these other sources of bias in their search process for statistically significant and theory-confirming estimates, meta-regression models also provide systematically false-positive findings of genuine effects as demonstrated for the case of omitted-variable biases in this article.

In experimental research, randomization ensures - at least in theory - unbiased and consistent estimation of the effect of interest and *p*-hacking is likely to rely more on chance rather than on systematically biased and inconsistent estimates of the effect of interest. If *p*-hacking relies more on chance, the bias is likely to be positively associated with the estimated standard error as large biases are unlikely to occur for large sample sizes. Therefore, meta-analysis is likely to play a promising role in identifying genuine effects and in adjusting for *p*-hacking in experimental research in economics including field experiments and quasi-experiments. Vivalt (2015), for example, demonstrates how meta-analytic tools may be utilized to synthesize research in the field of impact evaluation in development economics.

Further research is needed to better understand if and how estimates stemming from highly heterogeneous primary studies can be synthesized by meta-regression models to improve the reliability of inferences in observational research.

VI. Conclusions

Meta-regression models are increasingly being utilized in economics to integrate estimates from observational research designs. We show by means of theory and Monte Carlo simulations that the Precision-Effect Test (PET) that aims to test for the presence of genuine effects provides highly inflated type I errors if the authors of primary studies use *p*-hacking based on omitted-variable biases, i.e. they search for statistically significant and theory-confirming estimates across different regression specifications. Our findings extend previous pioneering simulation results that suggest a robustness of PET with respect to omitted-variable biases in primary studies (Stanley, 2008).

Our findings cast doubt on whether recent meta-regression analyses can help to improve the reliability of inferences on genuine effects in observational research in economics. Meta-regression models may

add dummy variables to filter out omitted-variable biases if the meta-analyst is willing to assume which primary regression specification is best to estimate the effect of interest. However, many meta-regression analyses in economics control only for some variation of the control variables in the primary studies (e.g. Adam *et al.*, 2013; Efendic *et al.*, 2011). This practice of meta-regression analysis may be caused by the belief that PET is largely robust to omitted-variable biases. More importantly, however, it may be difficult or even impossible in many cases to control for the high degree of heterogeneity of regression specifications that can be observed in primary literatures in economics.

Given the uncertainty whether a primary literature suffers from *p*-hacking, the difficulties to control for *p*-hacking may limit the applicability of meta-regression models for the purpose of identifying genuine effects in observational research in economics.

References

Adam, A., Kammas P. and Lagou, A. (2013). 'The effect of globalization on capital taxation: What have we learned after 20 years of empirical studies?', *Journal of Macroeconomics*, Vol. 35, pp. 199-209.

Becker, B. J. and Wu, M.-J. (2007). 'The synthesis of regression slopes in meta-analysis', *Statistical Science*, Vol 22, pp. 414–429.

Brodeur, A., Le, M., Sangnier, M. and Zylberberg, Y. (2015). 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics* (forthcoming).

Card, D. and DellaVigna, S. (2013). 'Nine Facts about Top Journals in Economics', *Journal of Economic Literature*, Vol. 51, pp. 144-461.

Card, D. and Krueger, A. B. (1995). 'Time-series minimum-wage studies: A meta-analysis', *American Economic Review*, Vol. 85, pp. 238–243.

De Long, J. B. and Lang, K. (1992). 'Are all economic hypotheses false?', *Journal of Political Economy*, Vol. 100, pp. 1257-1272.

Doucouliagos, C. and Stanley, T. D. (2013). 'Are all economic facts greatly exaggerated? Theory competition and selectivity', *Journal of Economic Surveys*, Vol. 27, pp. 316-339.

Efendic, A., Pugh, G. and Adnett, N. (2011). 'Institutions and economic performance: A meta-regression analysis', *European Journal of Political Economy*, Vol. 27, pp. 586-599.

Egger, M., Smith, G. D., Schneider, M. and Minder, C. (1997). 'Bias in meta-analysis detected by a simple, graphical test', *British Medical Journal*, Vol. 315, pp. 629–634.

Fanelli, D. (2010). '"Positive" results increase down the hierarchy of the sciences', *PLoS ONE*, Vol. 5, e10068.

Frey, B. S. (2003). 'Publishing as prostitution? - Choosing between one's own ideas and academic success', *Public Choice*, Vol. 116, pp. 205–223.

Gerber, A. S. and Malhotra, N. (2008a). 'Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?', *Sociological Methods & Research*, Vol. 37, pp. 3-30.

Gerber, A. S. and Malhotra, N. (2008b). 'Do statistical reporting standards affect what is published? Publication bias in two leading political science journals', *Quarterly Journal of Political Science*, Vol. 3, pp. 313-326.

Glaeser, E. L. (2006). 'Researcher Incentives and Empirical Methods', *NBER Technical Working Paper 329*.

Hendry, D. F. (1980). 'Econometrics - alchemy or science?', *Economica*, Vol. 47, pp. 387–406.

Ioannidis, J. P. A. (2005). 'Why most published research findings are false', *PLoS Medicine*, Vol. 2, e124.

Koetse, M. J., Florax, R. J., and De Groot, H. L. (2010). 'Consequences of effect size heterogeneity for meta-analysis: a Monte Carlo study', *Statistical Methods and Applications*, Vol. 19, pp. 217-236.

Leamer, E. E. (1983). 'Let's take the con out of econometrics', *American Economic Review*, Vol. 73, pp. 31–43.

Ridley, J., Kolm, N., Freckelton, R. P. and Gage, M. J. G. (2007). 'An unexpected influence of widely used significance thresholds on the distribution of reported p-values', *Journal of Evolutionary Biology*, Vol. 20, pp. 1082–1089.

Rosenthal, R. (1979). 'The file drawer problem and tolerance for null results', *Psychological Bulletin*, Vol. 86, pp. 638–641.

Simonsohn U, Nelson, L. D. and Simmons, J. P. (2014). 'P-curve: A key to the file-drawer', *Journal of Experimental Psychology: General*, Vol. 143, pp. 534-547.

Sims, C. A. (1988). 'Uncertainty across models', *American Economic Review*, Vol. 78, pp. 163–167.

Stanley, T. D. (2008). 'Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection', *Oxford Bulletin of Economics and Statistics*, Vol 70, pp. 103–127.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A. and Song, F. (2000). *Methods for meta-analysis in medical research*, J. Wiley, New York, USA.

Vivalt, E. (2015). 'How much can we generalize from impact evaluations?' *Working paper*.

Wu, M. J. and Becker, B. J. (2013). 'Synthesizing regression results: a factored likelihood method', *Research Synthesis Methods*, Vol. 4, pp. 127-143.

Young, N. S., Ioannidis, J. P. A. and Al-Ubaydli, O. (2008). 'Why current publication practices may distort science', *PLoS Medicine*, Vol. 5, e201.