DISCUSSION PAPER SERIES

DP17927

SPURIOUS PRECISION IN META-ANALYSIS

Zuzana Irsova, Pedro R. D. Bom, Tomas Havranek and Heiko Rachinger

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

ISSN 0265-8003

SPURIOUS PRECISION IN META-ANALYSIS

Zuzana Irsova, Pedro R. D. Bom, Tomas Havranek and Heiko Rachinger

Discussion Paper DP17927
Published 21 February 2023
Submitted 16 February 2023

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

SPURIOUS PRECISION IN META-ANALYSIS

Abstract

Meta-analysis upweights studies reporting lower standard errors and hence more precision. But in empirical practice, notably in observational research, precision is not given to the researcher. Precision must be estimated, and thus can be p-hacked to achieve statistical significance. Simulations show that a modest dose of spurious precision creates a formidable problem for inverse-variance weighting and bias-correction methods based on the funnel plot. Selection models fail to solve the problem, and the simple mean can beat sophisticated estimators. Cures to publication bias may become worse than the disease. We introduce an approach that surmounts spuriousness: the Meta-Analysis Instrumental Variable Estimator (MAIVE), which employs inverse sample size as an instrument for reported variance.

JEL Classification: C15, C26, C83

Keywords: N/A

Zuzana Irsova - zuzana.irsova@ies-prague.org
Charles University, Prague

Pedro R. D. Bom - pedro.bom@deusto.es
University of Deusto

Tomas Havranek - t.havranek@gmail.com
Charles University, Prague and CEPR

Heiko Rachinger - heiko.rachinger@gmail.com
University of the Balearic Islands

Spurious Precision in Meta-Analysis*

Zuzana Irsova[a], Pedro R. D. Bom[b], Tomas Havranek[a,c,d], and Heiko Rachinger[e]

[a]Charles University, Prague
[b]University of Deusto, Bilbao
[c]Centre for Economic Policy Research, London
[d]Meta-Research Innovation Center, Stanford
[e]University of the Balearic Islands, Palma

February 16, 2023

Abstract

Meta-analysis upweights studies reporting lower standard errors and hence more precision. But in empirical practice, notably in observational research, precision is not given to the researcher. Precision must be estimated, and thus can be p-hacked to achieve statistical significance. Simulations show that a modest dose of spurious precision creates a formidable problem for inverse-variance weighting and bias-correction methods based on the funnel plot. Selection models fail to solve the problem, and the simple mean can beat sophisticated estimators. Cures to publication bias may become worse than the disease. We introduce an approach that surmounts spuriousness: the Meta-Analysis Instrumental Variable Estimator (MAIVE), which employs inverse sample size as an instrument for reported variance.

Keywords: Publication bias, p-hacking, selection models, meta-regression, funnel plot, inverse-variance weighting

JEL Codes: C15, C26, C83

*Corresponding author: Zuzana Irsova, `zuzana.irsova@ies-prague.org`. Replication files are available in an online appendix at `meta-analysis.cz/maive`. The Meta-Analysis Instrumental Variable Estimator (MAIVE), a novel bias-correcting technique introduced in this paper, can be implemented using the `maive` package in R. Irsova, Bom, and Rachinger acknowledge support from the Czech Science Foundation grant "Spurious Precision in Meta-Analysis of Social Science Research" (#23-05227M). Havranek acknowledges support from the Institute for Research on the Socioeconomic Impact of Diseases and Systemic Risks (#LX22NPO5101) funded by the European Union—Next Generation EU.

1 Introduction

Inverse-variance weighting reigns in meta-analysis.[1] More precise studies, or rather those seemingly more precise based on lower reported standard errors, get upweighted explicitly or implicitly. The weight is explicit in traditional summaries: the fixed-effect model (assuming a common effect) and the random-effects model (allowing for heterogeneity).[2,3] These models are weighted averages, the weight diluted in random effects by a heterogeneity term. The weight is also explicit in publication bias correction models based on the funnel plot.[4–12] In funnel-based models, precision is particularly important because the weighted average gets reinforced by assigning more importance to supposedly less biased (nominally more precise) studies. The weight is implicit in selection models estimated using the maximum likelihood approach,[13–18] which often reduce to the random-effects model in the absence of publication bias.

The tacit assumption behind all these techniques is that the reported, nominal precision represents the true, underlying precision. The standard error, inverse of precision, is given to the researcher by her data and methods. It is fixed and cannot be manipulated, consciously or unconsciously. The assumption is plausible in experimental research, for which most meta-analysis methods were developed. But in observational research, where thousands of meta-analyses are produced each year, the derivation of the standard error is often a key part of the empirical exercise. Consider a regression analysis with longitudinal data: explaining the health outcomes of patients treated by different physicians and observed over several years. Individual observations are not independent, and standard errors need to be clustered.[19] But how? At the level of physicians, patients, or years? Should one use double clustering[20] or perhaps wild bootstrap[21]? It is complicated, and with a different computation of confidence intervals the researcher will report different precision for the same estimated effect size.

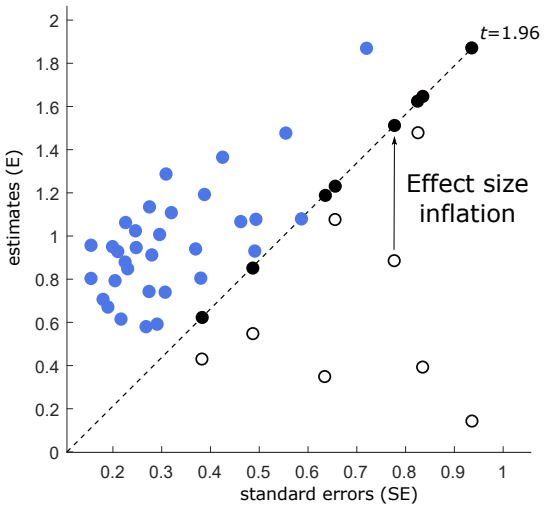
Spurious precision can arise in many contexts other than longitudinal data analysis. Ordinary least squares, the workhorse of observational research, assume homoskedasticity of residuals. The assumption is often violated, and in these cases researchers should use heteroskedasticity-robust standard errors,[22] typically larger than plain vanilla standard errors. If researchers ignore heteroskedasticity, they report precise estimates, but the precision is spurious. Similar problems may arise due to nonstationarity in time series[23] and a myriad of other issues. When a study with exaggerated precision enters meta-analysis, it gets too much impact because of inverse-variance weighting. If a meta-analyst spots the methodological problem, she can exclude the study or add a corresponding control. Either way, the weighting problem is not properly addressed. And spotting misspecification is hard, because the tweak lifting reported precision can be hidden within a complex model.

Spurious precision can also arise due to cheating. For economics journals, quasi-experimental evidence shows that obligatory data sharing reduced the reported t-statistics.[24] Prior to the introduction of data sharing some authors had probably cheated by manipulating data or results. Pütz and Bruns[25] find hundreds of reporting errors in top economics journals; when they ask authors to explain the errors, the authors are four times more likely to admit a mistake in the standard error than in the estimated effect size. But cheating, mistakes, and other issues

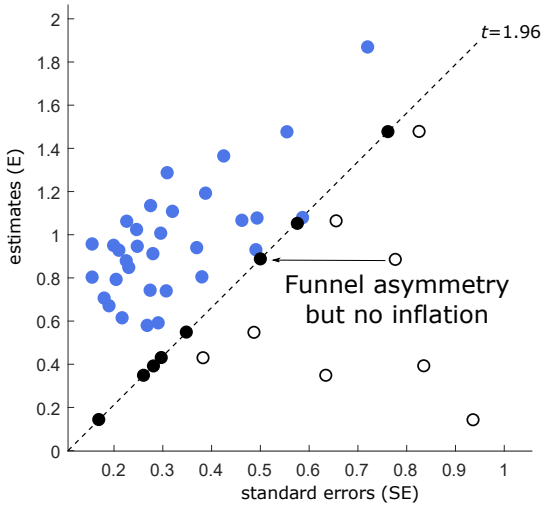
that can affect the standard error independently of the estimated effect size are unnecessary to produce spurious precision. A realistic mechanism is p-hacking, in which the researcher adjusts the entire model to produce statistically significant results. After adjusting the model, both the effect size and standard error change, and both can jointly contribute to statistical significance. We examine, by employing Monte Carlo simulations, the consequences of cheating and the more realistic p-hacking behavior, of which spurious precision is a natural result.

Figure 1: Two flavors of selection and repercussions for conventional meta-analysis

(a) Selection on estimates (E-selection)



(b) Selection on standard errors (SE-selection)



Notes: Blue-filled circles (lighter in grayscale) denote estimates statistically significant at the 5% level. The significant estimates are reported. Hollow circles denote insignificant estimates, which are not reported but p-hacked to yield statistical significance (black-filled circles). In the left-hand panel the resulting mean of reported estimates is biased upwards, and inverse-variance weighting helps mitigate the bias. In the right-hand panel the resulting mean is unbiased, and inverse-variance weighting introduces a downward bias. A realistic scenario of p-hacking combines both types of selection, so the p-hacked estimates move not strictly north or west (as in the figure) but northwest. The resulting bias direction due to inverse-variance weighting is ex ante unclear.

Figure 1 gives intuition on the cheating/clustering/heteroskedasticity/nonstationarity simulation. For brevity we call it a cheating scenario. Researchers crave statistically significant estimates and to that effect manipulate effect sizes or standard errors at will, but not both at the same time. The scenario is simplistic, and we start with it because it allows for a clean separation of selection on estimates (conventional in the literature) and selection on standard errors (our focus). The separation is not so clean in the p-hacking scenario, but can be mapped back to the cheating scenario. The mechanism of the left-hand panel of Figure 1 is analogous to the Lombard effect in psychoacoustics:[26–28] speakers increase their vocal effort in response to noise. Here researchers increase their selection effort in response to noise in data or methods, noise that produces imprecision and insignificance. When researchers so cheat with effect sizes, the results are consistent with funnel-based models of publication bias: funnel asymmetry arises, the most precise estimates remain close to the true effect, and inverse-variance weighting helps

mitigate the bias—aside from improving the efficiency of the aggregate estimate, the original rationale for using the weights.[29,30]

The right-hand panel of Figure 1 paints a different picture. Here the mechanism is analogous to Taylor's law in ecology:[31,32] variance can decrease with a smaller mean (originally describing population density for various species). When researchers achieve statistical significance by lowering the standard error, we again observe funnel asymmetry. But this time no bias arises in the reported effect sizes: the black-filled circles and the hollow circles denote the same effect size, only precision changes. The simple unweighted mean of reported estimates is unbiased, and inverse-variance weighting creates a downward bias. The bias increases when we use a correction based on the funnel plot: effectively, when we estimate the size of a hypothetical infinitely precise estimate, the intercept of a regression curve.

In practice, as noted, selection on estimates and standard errors arises simultaneously. We generate this quality in simulations by allowing researchers to replace control variables in a regression context, a mechanism that also creates sizable heterogeneity. Control variables are correlated with the main regressor of interest (treatment), and their replacement affects both the estimated treatment effect and the corresponding precision. Then p-hacked estimates move not strictly north or west, as in the figure, but northwest. Even spuriously large estimates can be spuriously precise. The resulting bias direction due to inverse-variance weighting is unclear. Our simulations suggest that an upwards bias in conventional estimators is plausible.

Does any technique yield little bias and good coverage rates in the case of panel B of Figure 1, or at least with a small ratio of selection on standard errors? We examine 7 current estimators: simple unweighted mean, fixed effects (weighted least squares, FE/WLS),[33] precision-effect test and precision-effect estimate with standard errors (PET-PEESE),[9] endogenous kink (EK),[11] weighted average of adequately powered estimates (WAAP),[10] the selection model by Andrews and Kasy,[17] and p-uniform*[18]. The first two are basic summary statistics, the next three are correction methods based on the funnel plot, and the last two are selection models. The choice of estimators is subjective, but the three funnel-based techniques are commonly used in observational research.[34–41] The two selection models are also used often[42–49] and represent the latest incarnations of models in the tradition of Hedges[13–16] and their simplifications[50–55].

None of these 7 estimators work well with even a sprinkle of spurious precision. The simple unweighted mean is often the best, but still no good. The reader might expect selection models to beat funnel-based models, because of the latter's heavier reliance on precision. This is generally not the case, and even selection models are often beaten by the simple mean when selection on standard errors is non-negligible (about 1:5 and more compared to selection on estimates). We propose a straightforward adjustment of funnel-based techniques, the meta-analysis instrumental variable estimator (MAIVE), which corrects most of the bias and restores valid coverage rates. MAIVE replaces, in all meta-analysis contexts, reported variance with the portion of reported variance that can be explained by inverse sample size used in the primary study. We justify the idea by starting with a version of the Egger regression:[4]

$$\hat{\alpha}_i = \alpha_0 + \beta SE(\hat{\alpha})_i^2 + v_i, \tag{1}$$

where $\hat{\alpha}$ denotes effects estimated in primary studies and SE their standard errors. This is the PEESE model due to Stanley and Doucouliagos, but for simplicity without additional inverse-variance weights—since the model searches for the effect conditional on maximum precision, it already features an implicit, built-in weight. In panel A of Figure 1, the quadratic regression would fit the data quite well,[9] and estimated α_0 would lie close to the mean underlying effect. In panel B, however, the regression fails to recover the underlying coefficients. The regression fails because it assumes a causal effect of the standard error on the estimate: a good description of panel A (Lombard effect), but not of panel B (Taylor's law). In panel B, the standard error sometimes depends on the estimated effect size, and is thus correlated with the error term, v_i. The resulting estimates of α_0 (true effect) and β (intensity of selection) are biased.

The problem is the correlation between SE and v_i, which can arise for three reasons: First, selective reporting on standard errors, which we simulate. Second, measurement error in SE. This issue was mentioned in 2005 by Tom Stanley,[56] who was the first to instrument the standard error in meta-analysis. Nevertheless, Stanley did not discuss the adjustment of weights nor did he pursue the idea further as a bias-correction estimator. We do not consider this source of correlation. Third, the correlation can be caused by unobserved heterogeneity: some method choices affect both estimates and standard errors, and some standardized meta-analysis effects feature a mechanical correlation between both quantities.[39] (A careless meta-analyst may also mix estimates measured in different units.[57]) Our p-hacking simulation only partly addresses this mechanism by allowing researchers to change control variables, which affect both estimates and standard errors at the same time—a combination of panel A and panel B of Figure 1. In other words, we model only some of the mechanisms which give rise to spurious precision.

The statistical solution to the problem, often called endogeneity, is to find an instrument for the standard error. A valid instrument is correlated with the standard error, but not with the error term (and thus unrelated to the three sources of endogeneity mentioned above). While it is often challenging to find good instruments, here the answer beckons. By definition, SE^2 is a linear function of the inverse of the sample size used in the primary study. The sample size is plausibly robust to selection, or at least it is more difficult to collect more data than to p-hack the standard error in order to achieve significance. The sample size is not estimated, and so does not suffer from measurement error. The sample size is typically not affected by changing methodology, certainly not by changing control variables. Some endogeneity may remain if researchers correctly expecting smaller effects design larger experiments.[44] But, at least in observational research, authors often use as much data as available from the start. Indeed, the sample size, unlike the standard error, is often given to the researcher: the very word *data* means things given.

We regress the squared reported standard errors on the inverse sample size and plug the fitted values instead of the variance to the right-hand side of Equation 1. Thence we obtain the baseline MAIVE estimator. For the baseline MAIVE we choose the instrumented version of PEESE without additional inverse-variance weights, because it works well in simulations. The version with additional adjusted weights (again, using fitted values instead of reported precision)

has often similar performance—but is more complex, so we prefer the former, parsimonious solution. In principle, any funnel-based technique (and the funnel plot itself) can be adjusted by the procedure described above: just replace the standard error with the square root of the fitted values. The adjustment helps the fixed-effect, WAAP, and endogenous kink model to typically defeat both the simple unweighted mean and selection models in the presence of spurious precision. MAIVE can be easily applied using our `maive` package in R.

Why not to simply replace variance with inverse sample size?[39,58–60] While the replacement would also address spurious precision, the instrumental approach has many advantages, discussed in detail in Section 3. One advantage is flexibility: the instrumental approach can incorporate other aspects of study design, on top of sample size, that affect standard errors. Sample size will rarely form a perfect proxy for precision, and MAIVE can be extended by adding instruments to improve the fit. Moreover, the instrumental approach remains statistically valid even if the correlation between reported variance and inverse sample size is small.

The remainder of the manuscript is structured as follows. In Section 2 we briefly describe how conventional estimators in meta-analysis use reported standard errors. Section 3 introduces the meta-analysis instrumental variable estimator. Section 4 presents a simple simulation based on cheating. Section 5 presents a more realistic simulation based on p-hacking. Section 6 concludes the paper. Appendix A and Appendix B provide additional simulation results.

2 Benchmark Estimators

We select 7 estimators, described in Table 1, to illustrate the impact of spurious precision. The choice of estimators is inevitably subjective. More could be used, but that would clutter simulation figures. We believe the 7 estimators represent the breadth of meta-analysis approaches. We always try to include the latest correction method that corresponds to a particular line of research—e.g., selection models in the tradition of Hedges—and has already been examined by simulations and applications. As economists we are biased towards techniques frequently used in economics meta-analyses. We include the Andrews and Kasy model among other reasons (including performance in previous simulations[61]) because Isaiah Andrews won the 2021 John Bates Clark Medal, the most prestigious award in economics after the Nobel Prize, explicitly also for his work on publication bias correction models.[17]

The simple average is the only examined estimator unaffected by precision. The FE/WLS estimator is the inverse-variance weighted version of the simple average. One could add random effects, but since there the weight is diluted by the heterogeneity term, the results would always lie between the simple average and FE, so random effects are less interesting for our simulations. A worry about spurious precision will reinforce the case of random effects versus unadjusted FE as a summary statistic, a choice otherwise depending on the nature of heterogeneity.[62,63]

The simple average and FE/WLS are summary statistics not primarily meant to correct biases—though researchers note that FE/WLS helps attenuate publication bias.[3,33,69] Of the 5 bias-correcting estimators, 3 are based on the funnel plot (PET-PEESE, EK, WAAP), and 2 are

Table 1: The role of the standard error in 7 benchmark estimators

Estimator	Weight	Regressor	Identification
Simple average			
FE/WLS	✓		
PET-PEESE	✓	✓	
EK	✓	✓	✓
WAAP	✓		✓
Andrews & Kasy	✓		✓
p-uniform*	✓		✓

Notes: Simple average = unweighted mean. FE/WLS = fixed effect/weighted least squares: mean weighted by inverse variance.[33] PET-PEESE = precision-effect test and precision-effect estimate with standard errors: selection is a quadratic function of SE when true effect $\neq 0$.[9] EK = endogenous kink: selection is a linear function of SE for imprecise estimates, no selection for precise ones.[11] WAAP = weighted average of adequately powered estimates: only estimates with at least 80% power included.[10] Andrews & Kasy = a selection model in the tradition of Hedges.[17] p-uniform* = a simplified selection model based on the principle that p-values should be uniformly distributed at the true effect size.[18] For tractability we do not consider promising new approaches that combine other estimators, such as RoBMA.[64–68]

selection models (Andrews and Kasy, p-uniform*). PET-PEESE is a classical funnel technique and can be understood as an extension of the Egger regression; it was shown to work well relative to other estimators when compared to pre-registered replications.[70] Regressions based on the funnel plot effectively invoke inverse-variance weighting twice: first as an explicit weight, second implicitly via the search for the mean effect conditional on maximum precision. Intuitively, such an approach will be highly sensitive to spurious precision. EK is a model that can be understood as a hybrid of funnel-based and selection model approaches[71] because it estimates a threshold for which estimates are selected no more. Since it follows funnel plot intuition and depends on the same assumption, we group it together with funnel-based techniques. EK is the technique theoretically most dependent on correctly reported precision: the standard error is used as a weight, regressor, and identification threshold.

The last funnel-based technique, WAAP, is not a regression. It is based on FE/WLS, but also uses the standard error to exclude studies that have low power (indirectly, that are too imprecise). We understand WAAP as another way to estimate the top of the funnel, the mean study implied by maximum precision. So WAAP, EK, and PET-PEESE share the same underlying assumptions: 1) In the absence of publication bias, there is no correlation between estimates and standard errors. 2) If there is publication bias, it works on the reported effect size, not the standard error. The standard error is exogenous (given to the researcher) and more precise (less noisy) estimates are less biased, as in the Lombard effect analogy laid out in the Introduction. Note that, in addition to publication bias, the funnel-based techniques also allow for p-hacking of effect sizes—if the p-hacking is a direct response to imprecision.

The selection models we consider, Andrews and Kasy and p-uniform*, allow for joint selection on both estimates and standard errors, as long as the final selection criterion is the p-value.

Andrews and Kasy is the latest selection model in the tradition of Hedges, while the p-uniform* is the latest version of selection model simplifications that are more parsimonious but less flexible. Even more flexible selection models exist that allow for separate selection on estimates and standard errors, irrespective of the p-value,[72–74] but these are difficult to estimate in meta-analyses of typical sizes and have so far been rarely used outside sensitivity analysis.[75,76] From the description above it might seem that selection models are immune to spurious precision. Alas, not necessarily so: they are not robust to p-hacking. If individual results are not only selectively chosen for publication but actively p-hacked, they are not individually unbiased.[77] Reported (and thus potentially spurious) precision may add to the bias, because it is used as a weight in maximum likelihood selection models. Both the Lombard effect and Taylor's law intuitions would then be inconsistent with selection models.

3 Meta-Analysis Instrumental Variable Estimator

The methodological recommendation of this paper is to replace the reported standard error with the error's portion explainable by sample size. That is, we only use variation in errors that can be linked to variation in sample sizes. Because in most contexts the sample size is harder to increase than the standard error is to p-hack, the adjusted measure is likely to better capture the underlying precision. The variation in standard errors unrelated to variation in sample sizes is more susceptible to p-hacking. Conventional estimators can be rid of the unrelated variation, and the result is the meta-analysis instrumental variable estimator (MAIVE). As noted, we choose the adjusted version of PET-PEESE as the baseline MAIVE. PEESE uses squared standard errors, and so we instrument the reported *variance* with inverse sample size:

$$SE(\hat{\alpha})_i^2 = \psi_0 + \psi_1(1/N_i) + \nu_i, \tag{2}$$

where $\hat{\alpha}$ is the effect size reported in a primary study, ψ_0 is a constant, N_i is the sample size of the primary study, and ν_i is an error term that soaks up, among other things, the spurious elements of the reported standard error related to p-hacking. Note that the approach is statistically valid even if the link between $SE(\hat{\alpha})^2$ and $(1/N_i)$ is weak, as long as there is any correlation. The instrumented variance, to be used for adjusting current meta-analysis estimators, equals $\hat{\psi}_0 + \hat{\psi}_1(1/N_i)$. In addition to PEESE, squared standard errors are also used in conventional inverse-variance weighting, so we find the quadratic specification natural. In estimators that require the standard error without squaring (such as PET, the first stage of PET-PEESE), we use $\sqrt{\hat{\psi}_0 + \hat{\psi}_1(1/N_i)}$ for the adjustment of the standard error. Regressing in Equation 2 the reported standard error on the inverse of the square root of sample size would cause only minor quantitative changes; it is largely a matter of taste and aesthetics.

The idea of using an instrument for the standard error in meta-analysis goes back to Tom Stanley in 2005.[56] He notes that meta-regression techniques can suffer from attenuation bias, the "iron law of econometrics,"[48,78] because the reported standard errors on the right-hand side are estimated standard deviations. When a regressor is measured with random noise, the

estimated slope coefficient is biased downwards. Stanley uses the instrumental variable approach to explore the robustness of funnel asymmetry tests, but does not employ it as an estimator of the underlying effect nor does he adjust the weights—in his influential 2005 *Journal of Economic Surveys* paper or any subsequent work. Because the foundational idea is due to Stanley, and also the PEESE estimator we prefer for adjustment was developed by Stanley (together with Doucouliagos),[9] the MAIVE technique could just as well be called the Stanley estimator.

We believe that the attenuation problem is not important in meta-regression. Consider the basic Egger regression, a linear regression of reported effect sizes on the reported standard errors—for simplicity without weights. To our knowledge, this regression was first used by Card and Krueger in their 1995 *American Economic Review* meta-analysis of the effect of minimum wage on employment,[79] a part of long-term research effort for which David Card won the 2021 Nobel Prize in economics. For attenuation bias to appear, the underlying relation needs to hold between estimates and standard deviations. Then the right-hand variable in meta-regression, the standard error, is a noisy version of the standard deviation. But the underlying relation really holds for estimates and reported standard errors. What matters is the resulting t-statistic, the ratio of estimates and standard errors. It follows that in the basic Card-Krueger-Egger regression there should be no attenuation bias. Measurement error bias is perhaps possible via inverse-variance weighting, or in models that combine meta-regression and selections models, such as endogenous kink[11]. We do not consider this mechanism.

Instead, we focus on reverse causality (selection on standard errors, a possibility noted in economics by Olken[80]) and omitted variables (resulting in p-hacking on both estimates and standard errors). These two issues were not discussed by Stanley, but have since been raised in footnotes by the co-authors of the present manuscript. Partial and imperfect versions of MAIVE, with limited and scattered justification, have appeared as robustness checks in our applied meta-analyses.[23,48,81] The estimator we prefer, MAIVE version of PET-PEESE without additional inverse-variance weights, has not been mentioned or used previously. Aside from Stanley's papers, we were inspired by the work of Egger et al.[82], Schmidt et al.[62], Hansen[83], and Nakagawa et al.[39,84], who explicitly or implicitly refer to the possibility of spurious precision in meta-analysis. Consider, for example, the following quote from Hansen, 2016, p. 1920:[83]

> The weighting of studies according to estimated precision is particularly problematic because the most unreliable estimates are also those with the least precise standard errors, and some of them will be treated as reliable simply due to error in their standard errors.

An obvious response to spurious precision could be to simply replace precision by a function of sample size, a replacement mentioned by several researchers[39,58–60]. While this would also remove spurious precision, the instrumental approach has 7 advantages. First, the underlying causal relationship in Equation 1 is one between estimates and standard errors, not sample size. The t-statistic is important. Second, the optimal meta-analysis weight is based on inverse variance, not on sample size.[29] Third, the correlation between precision and sample size is not perfect. By using the instrumental variable approach, the confidence intervals can take into

Table 2: Estimators and their MAIVE variants considered in simulations

Estimator	Variants
Simple average	(1) Unadjusted
FE/WLS	(1) Unadjusted (2) Adjusted weights
PET-PEESE	(1) Unadjusted (2) Adjusted weights (3) Instrumented SEs (4) Adjusted weights and instrumented SEs **(5) Instrumented SEs and no weights (MAIVE baseline)**
EK	(1) Unadjusted (2) Adjusted weights (3) Instrumented SEs (4) Adjusted weights and instrumented SEs (5) Instrumented SEs and no weights
WAAP	(1) Unadjusted (2) Adjusted weights and SEs
Andrews & Kasy	(1) Unadjusted (2) Adjusted SEs
p-uniform*	(1) Unadjusted (2) Adjusted SEs

Notes: SE = standard error. MAIVE = meta-analysis instrumental variable estimator. Simple average = unweighted mean. FE/WLS = fixed effect/weighted least squares: mean weighted by inverse variance. PET-PEESE = precision-effect test and precision-effect estimate with standard errors: selection is a quadratic function of the standard error.[9] EK = endogenous kink: selection is a linear function of the standard error for imprecise estimates, no selection for precise estimates.[11] WAAP = weighted average of adequately powered estimates: only estimates with at least 80% power are included.[10] Andrews & Kasy = a selection model in the tradition of Hedges.[17] p-uniform* = a simplified selection model using the statistical principle that p-values should be uniformly distributed at the true effect size.[18]

account that imperfect correlation in the first stage. Fourth, under classical assumptions, the funnel plot with estimate size and precision is symmetrical in the absence of publication bias. The assumptions are stronger for funnels that use sample size instead of precision.[85]

Fifth, the standard error is affected not only by sample size but also by estimation context. Variables related to methodology can be used as instruments alongside sample size in MAIVE. The technique is thus more flexible, though it may be challenging to find instruments uncorrelated with the meta-regression error term. Sixth, when the correlation between precision and sample size is small, which may happen in observational research contexts (or in experimental research with poor randomization), the instrument can be weak. Statistical methods have been developed to deal with weak instruments and ensure valid inference,[86–89] an outcome not guaranteed when we simply replace variance with inverse sample size in Equation 1. Seventh, MAIVE automatically recomputes sample size to the units of precision reported by primary

studies. This task can be achieved without the instrumental variable approach, but here the process is automatic. In consequence, with MAIVE all meta-analysis techniques can proceed as usual, now with a corrected measure of precision that is more likely than the reported, nominal precision to reflect the underlying, true precision. The message of our paper is not that inverse-variance weighting and funnel-based correction methods are wrong, but that they can be made considerably more robust with little cost.

In Table 2 we list the versions of the 7 estimators employed in simulations. For the simple average there is just one version, because the unweighted mean does not use the standard error at all. For the remaining 6 estimators we adjust the inverse-variance weight using the fitted value from the first-stage instrumental variable regression described above, creating the MAIVE variants of these estimators. For FE/WLS, WAAP, Andrews and Kasy, and p-uniform* we have just one version of adjustment: the standard error in these models is simply replaced with the square root of the fitted value. For regression-based models (PET-PEESE and EK) we have in total 5 variants: i) no adjustment, ii) adjusted weights, iii) instrumented standard errors in the meta-regression but unadjusted weights, iv) adjusted weights and instrumented standard errors, and v) instrumented standard errors and omitted weights. Meta-regression techniques thus allow us to easily separate the effect of spurious precision on weighting and identification. The separation is more difficult for selection models; the adjusted versions that we use for selection models aggregate both effects and thus have weak statistical justification. This is an important issue we return to in the Conclusion and highlight for future research.

4 Selection Based on Cheating

4.1 Simulation Setup

We simulate a meta-analysis environment where the object of interest is a regression coefficient. Striving for statistical significance, researchers engage in questionable research practices,[90] which give rise to selection on estimates or standard errors. We consider two alternative selection mechanisms, one based on cheating and one based on p-hacking. In the cheating selection environment, researchers unsatisfied with statistically insignificant results simply replace the obtained estimates or standard errors by fake values that are just enough to make the estimate statistically significant. (We call this cheating but, as noted, the lower standard error can be achieved via a different treatment of clustering, heteroskedasticity, or nonstationarity.) Although admittedly simplistic, this selection mechanism allows us to conveniently control the relative degree of selection entering through estimates and standard errors. Acknowledging the limitations of this selection scenario as a description of researchers' actual behavior, we use it merely to highlight the *qualitative* implications of selection on standard errors for the performance of the various estimators considered.

The second mechanism, discussed in the next section, is based on p-hacking, whereby the researcher engages in a continuous search of statistical significance by trying many different control variables. Although more realistic, this scenario features significantly less control over

the relative degree of selection on standard errors. One important advantage of p-hacking selection is its ability to generate selection of the two aforementioned flavors while, at the same time, causing heterogeneity in the reported estimates. But first back to cheating.

4.1.1 Generation of Primary Data

The primary data in the cheating scenario are generated according to

$$Y = \alpha_0 + \alpha_1 X + u, \tag{3}$$

where $\alpha_0 = 0$ (without loss of generality), $X \sim U(0,1)$ and $u \sim N(0, \sigma_u^2)$. The parameter of interest to meta-analysis is α_1. Let i index each primary study and let there be M such primary studies, so that $i = 1, 2, \ldots, M$. Each primary study obtains random samples of size N_i for variables Y and X, estimates the regression model specified by Equation 3, and reports the OLS estimate of α_1 and its corresponding standard error.

4.1.2 Selection

Researchers prefer positive and statistically significant estimates of α_1. A fraction π of the researchers are potential cheaters, who are willing to cheat on either the reported estimates (E-selection) or on the standard errors (SE-selection) in order to inflate the statistical significance of their findings. They do so only when obtaining a positive but statistically insignificant estimate. Hence, when the obtained estimate is either negative or positive but statistically significant, the results are reported truthfully. If, on the contrary, the obtained estimate is positive but statistically insignificant, the researcher cheats on the reported findings with probability π.

With probability ϕ, a cheating researcher engages in SE-selection, replacing the obtained standard error by a fake value that is just enough to achieve statistical significance at the 5% level; that is, the reported standard error is $\hat{\alpha}_1/1.96$. With probability 1-ϕ the researcher engages instead in E-selection, replacing the obtained estimate by a fake value that is just enough to achieve statistical significance at the 5% level; that is, the reported estimate is $SE(\hat{\alpha}_1) \times 1.96$.

In this environment, therefore, the overall magnitude of publication selection is measured by π and the relative importance of SE-selection versus E-selection is measured by ϕ. Note that, to keep perfect control on ϕ as measuring the relative importance of the two types of selection, we assume that researchers do not engage in both types simultaneously. This is also the reason why we assume that negative estimates are not subject to selection; otherwise, a negative estimate would have to become positive, which would necessarily involve E-selection. This restriction will be removed in the p-hacking scenario of the following section, so we ask the reader to tolerate it while we build toward a more realistic simulation.

4.1.3 Parameter Values and Distributions

We implement this type of selection by means of the parameter values and distributions summarized in Table 3. The number of studies in a meta-analysis is $M = 80$ in line with previous

related simulations.[9,11] We assume that primary sample sizes are drawn from a uniform distribution over (30, 1000); in the next section we will calibrate the sample size distribution based on 436 published meta-analyses. We consider three alternative values of α_1: zero, one, and two. We interpret these values as representing no effect, a moderate effect, and a large effect, respectively. We assume that the probability of potentially engaging in cheating is $\pi = 0.5$ and let ϕ vary from 0 to 1 in steps of 0.25. Note that $\phi = 0$ corresponds to pure E-selection, whereas $\phi = 1$ corresponds to pure SE-selection. Finally, we calibrate $\sigma_u^2 = 3.3$ in order to generate similar effective incidences of selection for $\alpha_1 = 0$ and $\alpha_1 = 2$, which is about 24% in both cases; for $\alpha_1 = 1$ it is a bit larger, at about 32%. (The effective incidence of selection is the overall fraction of findings subject to selection. Note that, in this scenario, effective selection incidence has a hump-shaped profile when graphed against α_1. This is because of the assumption of no selection on negative findings. Hence, when $\alpha_1 = 0$, selection incidence is not very high because approximately half of the estimates are negative. It gets higher for $\alpha_1 = 1$, because less estimates are then negative. And it gets lower again for $\alpha_1 = 2$ because more estimates become significantly positive naturally, even without selection.)

Table 3: Parameter values and distributions in the cheating selection scenario

Parameter/Variable	Description	Values/distribution
X	Regressor of the primary model	$\sim U(0,1)$
u	Error term of the primary model	$\sim N(0,\sigma_u^2)$
M	Number of studies/estimates	80
N_i	Sample size of the primary study	$\sim U(30,1000)$
α_1	Size of the true effect	0, 1, 2
π	Fraction of potential cheaters	0.5
ϕ	Fraction of selection on standard errors	0, 0.25, 0.5, 0.75, 1
σ_u^2	Variance of the error term	3.3

Notes: See text for explanation of the chosen values and distributions. When possible, in calibration we follow the tradition of previous simulations built for meta-analysis in the context of regression estimates.[9,11,61]

4.1.4 Replications and Statistics

To study the performance of the 7 baseline estimators and their MAIVE variants, we set the number of replications to $R = 2000$. We compute the bias and the mean squared error (MSE) of each estimator by averaging the estimation errors and the squared estimation errors over R, respectively. Hence, for a generic estimator z, the two statistics are given by

$$\begin{aligned} \text{Bias}(z) &= \frac{1}{R}\sum_{i=1}^{R}(z_i - \alpha_1), \\ \text{MSE}(z) &= \frac{1}{R}\sum_{i=1}^{R}(z_i - \alpha_1)^2. \end{aligned}$$

In addition, we also compute the coverage rates of each estimator by counting the number of confidence intervals that contain the true value of α_1 as a fraction of the total number of replications. Because of space and ease of exposition considerations, in the main text we present only the results on bias and coverage rates, with MSE results reported in Appendix B.

4.2 Results

Figures 2-5 in the main text show simulation results for the case of cheating selection. Because the results for a large true effect are very similar to the results for a moderate effect, we relegate the former to Appendix A. We do not show the results for Andrews and Kasy's estimator and p-uniform*, since they are not suited for this particular flavor of selection, giving huge biases and low coverage rates. We discussed the issue with the authors of p-uniform*, who confirmed that the estimator is not suitable to our cheating scenario. Thus it would be unfair to include these estimators now and compare them with the agnostic funnel-based techniques. All estimators are included in the more realistic p-hacking scenario of the following section.

The figures follow the same structure: panel (a) displays the unadjusted estimators, panels (b)-(e) show, one by one, the effect of adjusting FE/WLS, PET-PEESE, EK, and WAAP, and panel (f) compares the best versions of the adjusted estimators. Figures 2-3 show the bias and coverage rates for the case of no effect ($\alpha_1 = 0$), then Figures 4-5 display the same results for the case of a moderate effect ($\alpha_1 = 1$), and, finally, Figures 10-12 in Appendix A do the same for the case of a large effect ($\alpha_1 = 2$). Results for MSE are shown in Appendix B.

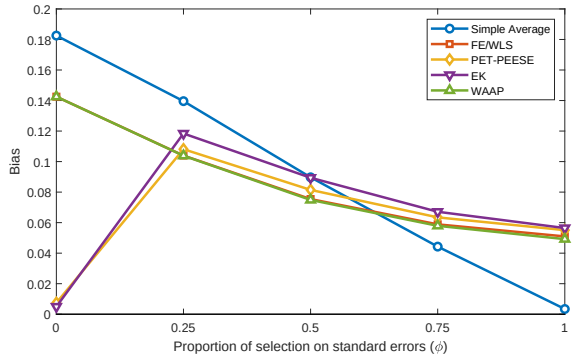
The most important results from Figure 2 are the following. When selection operates fully on estimates ($\phi = 0$), the simple average of the reported estimates shows a large bias, partly corrected by FE/WLS and WAAP, and entirely corrected by PET-PEESE and EK, as expected. When selection is fully on standard errors ($\phi = 1$), however, only the simple average is unbiased; FE/WLS, WAAP, PET-PEESE, and EK all show a similar positive bias. This implies that selection on standard errors is fundamentally a weighting problem. The bias arises from assigning too much weight to positively-selected estimates. Accounting for the correlation between estimates and standard errors (PET-PEESE and EK) does not solve the problem.

Because selection on standard errors is a weighting problem, adjusting the weights in FE/WLS naturally makes it unbiased. The same applies for WAAP, which is virtually equal to FE/WLS in this case. But adjusting the weights by itself does not solve the problem if the standard errors are also included as a regressor (PET-PEESE and EK) and ϕ is large; in fact, it makes the bias larger. Not including the standard errors as a regressor, on the other hand, does not correct the bias if ϕ is small. Because ϕ is, in practice, unknown, the standard error should be included in the regression but instrumented (either in PET-PEESE or in EK). Instrumenting makes the bias much smaller.

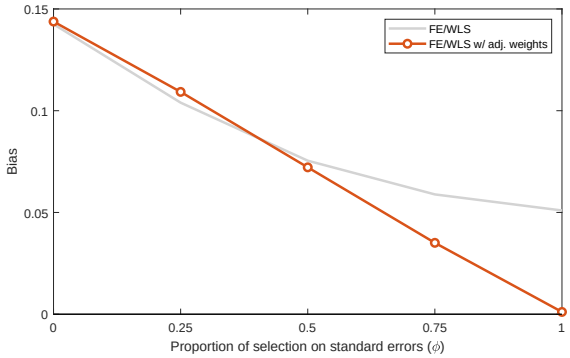
What about instrumenting the SEs and at the same time adjusting the weights in PET-PEESE and EK? It turns out that this does not change much the bias relative to the case of only instrumenting the SEs. As we will see below, however, it does improve on other metrics (coverage, in particular) and for nonzero true effects. Dropping the weights when including and

Figure 2: Bias: cheating selection, no effect ($\alpha_1 = 0$)

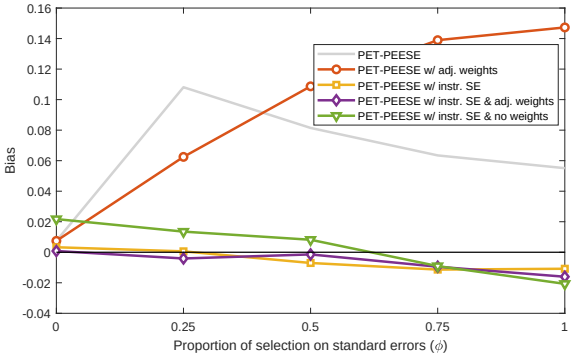
(a) Unadjusted estimators



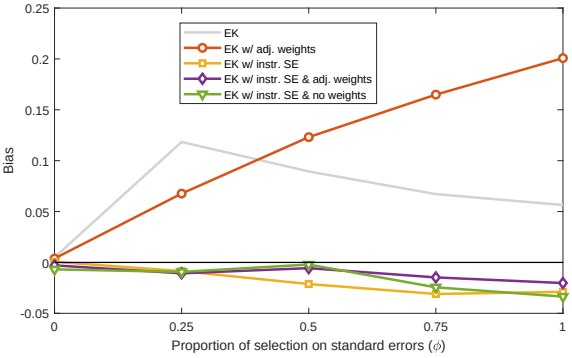
(b) Adjusting FE/WLS



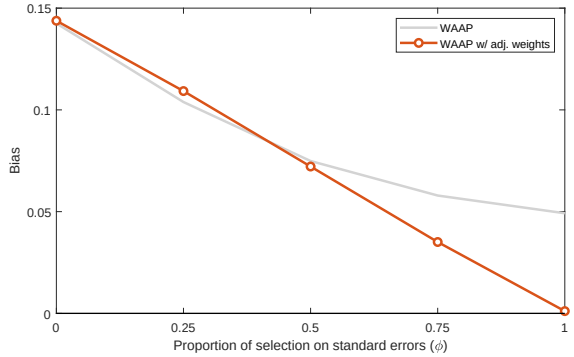
(c) Adjusting PET-PEESE



(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators

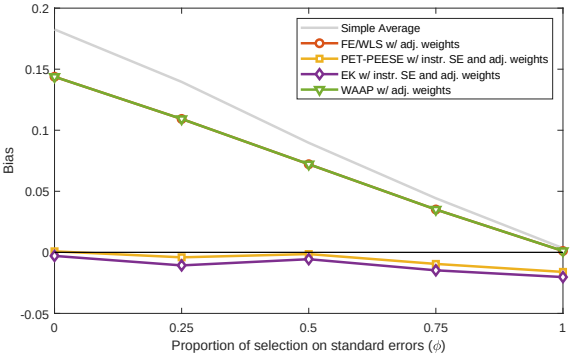
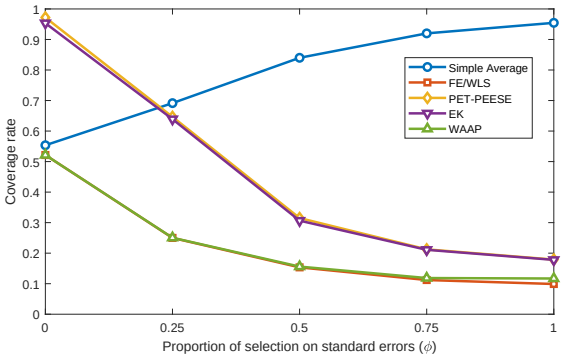
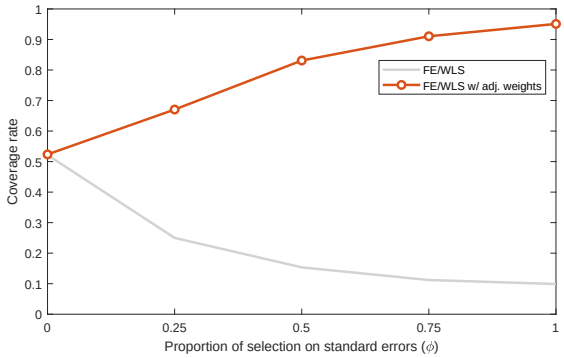


Figure 3: Coverage: cheating selection, no effect ($\alpha_1 = 0$)

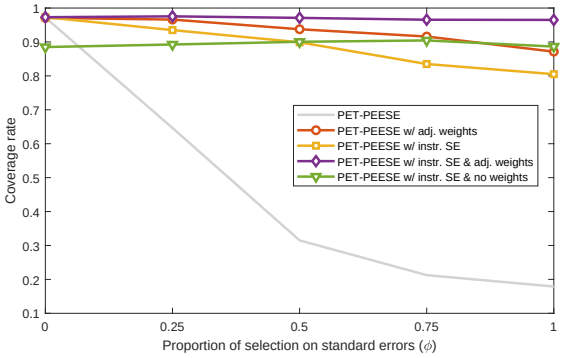
(a) Unadjusted estimators



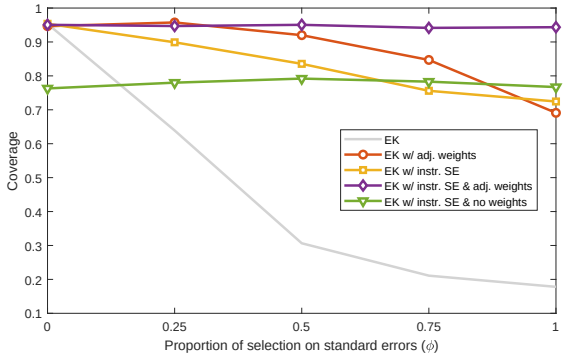
(b) Adjusting FE/WLS



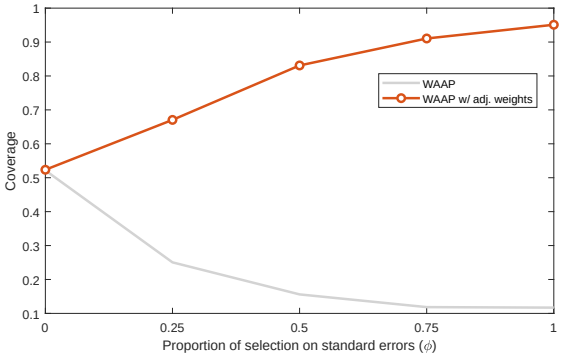
(c) Adjusting PET-PEESE



(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators

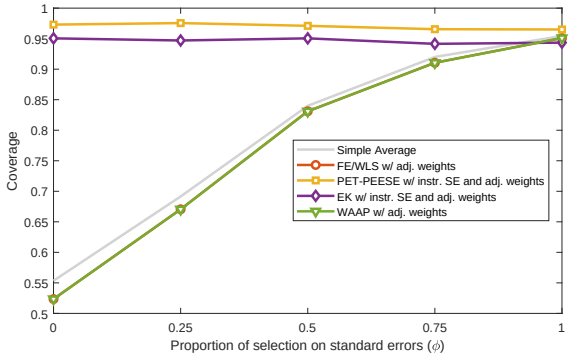
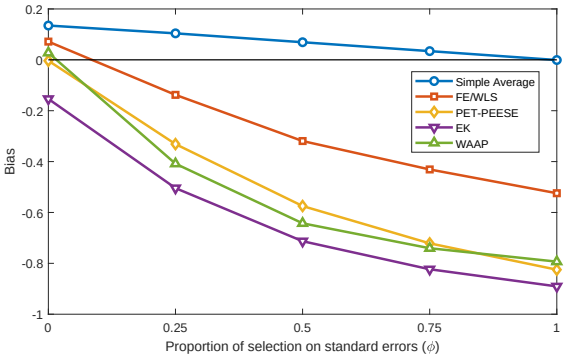
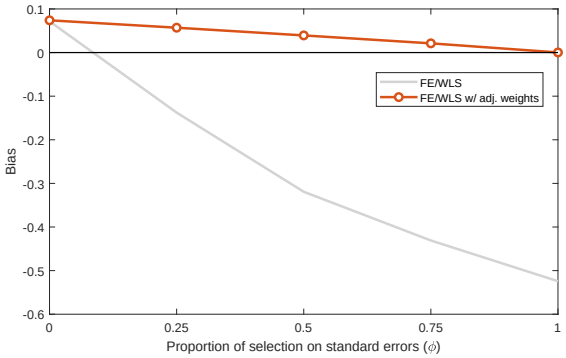


Figure 4: Bias: cheating selection, moderate effect ($\alpha_1 = 1$)

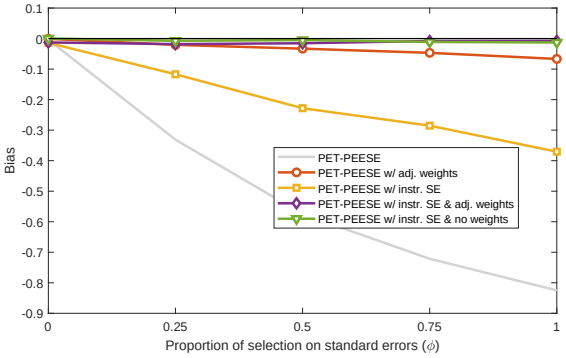
(a) Unadjusted estimators



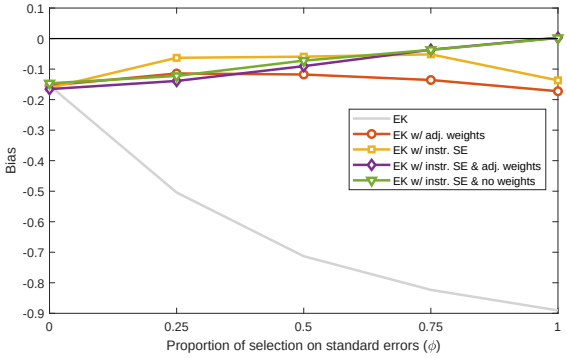
(b) Adjusting FE/WLS



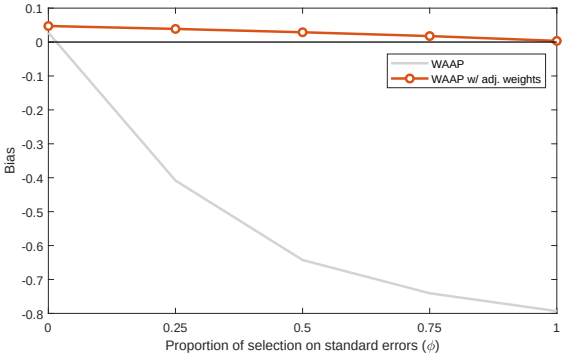
(c) Adjusting PET-PEESE



(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators

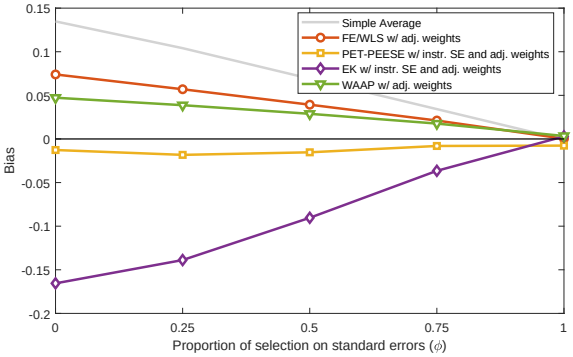
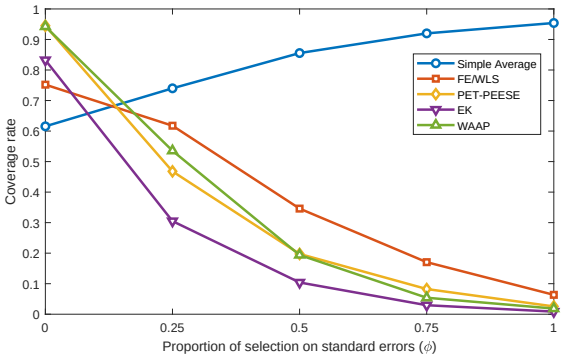
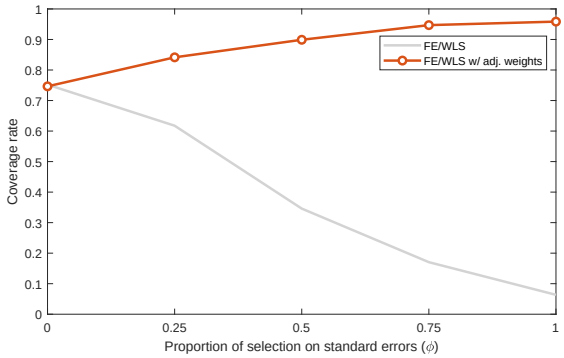


Figure 5: Coverage: cheating selection, moderate effect ($\alpha_1 = 1$)

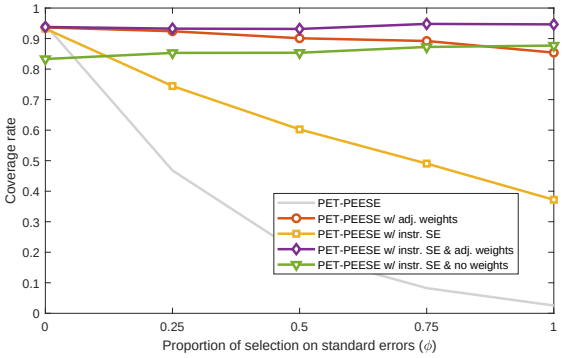
(a) Unadjusted estimators



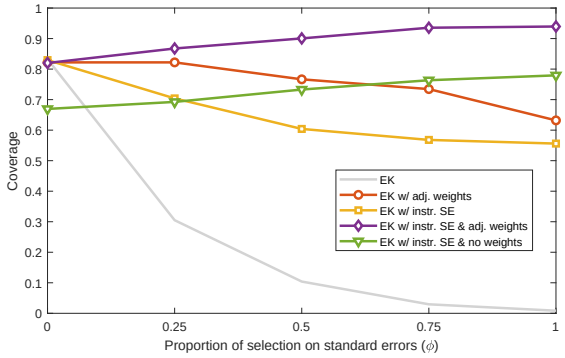
(b) Adjusting FE/WLS



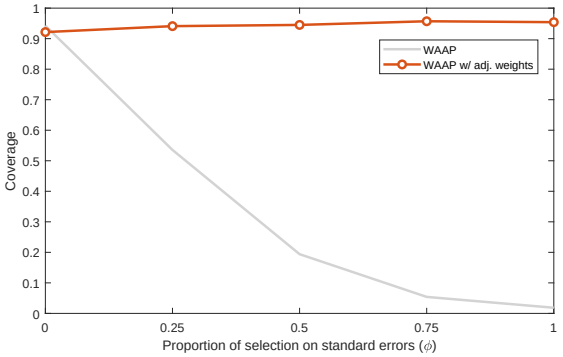
(c) Adjusting PET-PEESE



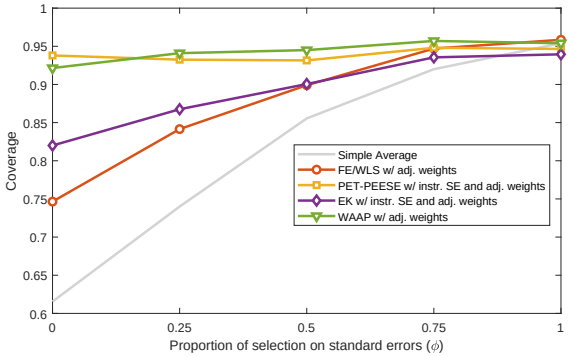
(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators



instrumenting the SEs in the regression also works reasonably well. The bottom line is that instrumenting SEs and adjusting or dropping the weights in PET-PEESE or EK reduces the selection bias for any value of ϕ.

Regarding the mean squared error, shown in Figure 13 in Appendix B, adjusting the weights or instrumenting the SEs, despite decreasing the estimators' bias, often increases the estimators' variance and hence their MSE. But the lower variance of the unadjusted estimators is partly spurious and, coupled with large biases, results in low coverage. Hence, this increase in variance is a necessary adjustment to restore the nominal coverage of confidence intervals, as we will see below. For all following simulations, MAIVE adjustments also help improve MSE.

Results for coverage rates are shown in Figure 3. Like the bias, the coverage rates of the unadjusted estimators deteriorate rapidly as the degree of selection on standard errors (ϕ) increases. For as little as $\phi = 0.25$, the coverage rate drops from about 95% to about 60% for PET-PEESE and EK, and from 50% to 25% for FE/WLS and WAAP. Only adjusting the weights or only instrumenting the SEs improves the coverage rate substantially but this rate falls with ϕ. But instrumenting the SEs while adjusting the weights keeps the coverage rate at the nominal level in PET-PEESE and EK, irrespective of ϕ. Dropping instead of adjusting the weights leads to lower coverage. (Note that this is generally not the case in the more realistic p-hacking scenario of the next section, when dropping the weights often performs better than adjusting them.) Adjusting the weights in FE/WLS or WAAP improves coverage, but only sufficiently so for very large values of ϕ.

The key results from Figures 4-5, which focus on a moderate true effect, are the following. For a positive true effect ($\alpha_1 = 1$), the unadjusted estimators, and in particular those that explicitly address publication selection (PET-PEESE, EK, and WAAP), become extremely downward biased as the fraction of selection on standard errors (ϕ) increases. The reason is that selection on standard errors, like selection on estimates, gives rise to a positive correlation between estimates and standard errors. The solution, as we saw above, should be readjusting the weights. But PET-PEESE and EK attribute this correlation to inflation in the estimates, and correct it down, causing a large negative bias for large ϕ. WAAP, likewise, is negatively biased, because it assigns too much weight to small estimates with artificially small standard errors.

As in the case of $\alpha_1 = 0$, adjusting the weights in FE/WLS and WAAP eliminates the bias for large ϕ. In PET-PEESE and EK, simultaneously instrumenting the SEs and adjusting (or dropping) the weights also eliminates the bias for large ϕ while keeping it low for small ϕ. However, only adjusting the weights or only instrumenting the SEs leads to larger negative biases for large ϕ. The bias due to inflated precision is much larger than the bias due to inflated estimates. Overall, the fully adjusted PET-PEESE shows the smallest bias across values of ϕ. In contrast to $\alpha_1 = 0$, the MSEs of the unadjusted estimators now increase with ϕ (Figure 14). Moreover, the adjusted estimators now show smaller MSEs. The adjusted estimators, especially PET-PEESE and WAAP, attain a coverage rate very close to the nominal level.

The results for a large underlying effect are available in Appendix A. The pattern of bias is very similar to the case of $\alpha_1 = 1$, with unadjusted estimators showing large negative biases for

large ϕ. Adjusting the weights in FE/WLS and WAAP eliminates the bias for large ϕ. Fully adjusting PET-PEESE and EK also leads to low biases across the different values of ϕ. In terms of MSEs and coverage rates, the case of $\alpha_1 = 2$ is also similar to the case of $\alpha_1 = 1$: the adjusted estimators always attain lower MSEs and better coverage, very close to the nominal level in all cases for any value of ϕ.

In a nutshell, selection on standard errors causes conventional estimators to be biased. The bias can be huge and negative for positive true effects. Coverage rates deteriorate substantially in the presence of this type of selection. Adjusting or omitting the weights and instrumenting the standard errors in PET-PEESE and EK is effective in reducing the bias caused by selection on standard errors. The instrumental adjustment also restores the coverage rates of confidence intervals to their nominal levels. Adjusting the weights in FE/WLS and WAAP improves these models' performance, but they remain less effective than the MAIVE version of PET-PEESE in correcting the bias caused by selection on estimates.

5 Selection Based on *p*-Hacking

5.1 Simulation Setup

5.1.1 Generation of Primary Data

In the more realistic p-hacking simulation the data generating process for primary studies includes not one but two regressors, X_1 and X_2:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u, \tag{4}$$

where, again, $\alpha_0 = 0$ (without loss of generality), $X_1 \sim U(0,1)$, and $u \sim N(0, \sigma_u^2)$. The second regressor, X_2, is a convex combination of X_1 and an independent random term $\epsilon \sim N(0,1)$; i.e., $X_2 = \psi X_1 + (1-\psi)\epsilon$, where $\psi \in (0,1)$. Hence, X_1 and X_2 are positively correlated by construction, this correlation being governed by ψ. The parameter of interest to meta-analysis is α_1. The M primary studies each report an OLS estimate and a corresponding standard error of α_1 using a sample of size N_i. The numerical values of the parameters depend on the selection mechanism assumed and are discussed below.

5.1.2 Selection

In this selection scenario, some researchers engage in questionable research practices by manipulating the specification of the model. In particular, we assume that primary studies start by estimating the correctly specified model (Equation 4). If the obtained estimate of α_1 is not positive and statistically significant in the correctly specified model, then, with probability π, the dissatisfied authors of such a primary study replace the true control variable X_2 by a different control variable, X_3. They try many such variables until they find one that 'works,' in the sense of turning the estimate of α_1 positive and statistically significant. (We implement

this idea by first uniformly drawing a correlation coefficient between X_2 and X_3, constrained to be positive and less than 0.8. We then generate variable X_3 to match this correlation with X_2. The maximum correlation of 0.8 is imposed just to save on computing time, since very high correlations do not help the cause of getting statistical significance.)

Replacing X_2 by a related but weaker control variable X_3 helps achieving statistical significance through both E-selection and SE-selection. E-selection works through the bias it causes on the estimate of α_1. Because X_1 and X_2 are positively correlated, dropping X_2 in fact biases upwards the estimate of α_1 (omitted-variable bias), making statistical significance more likely. The bias increases with the correlation between X_1 and X_2 and with the value of α_2. The bias is somewhat mitigated by the inclusion of X_3. Note that, by inducing biases in the reported estimates, p-hacking causes not only publication selection but also excess variation in the reported findings (i.e., heterogeneity), a feature that characterizes most meta-analyses in observational research. In economics, for example, heterogeneity—rather then the unconditional mean—is often the focus of applied meta-analyses.

The p-hacking process also causes SE-selection. This is because replacing X_2 by a weaker control decreases the amount of multicollinearity in the model, thus artificially decreasing the standard error of the estimate of α_1 relative to the corresponding standard error in a correctly specified model. SE-selection also increases with the correlation between X_1 and X_2, governed by ψ, but proportionally more so than E-selection. (To see this, first note that E-selection depends on the bias of the estimate of α_1. At most, in case X_2 is dropped or replaced by an irrelevant control, this bias is given by $\psi\alpha_2$, and thus increases linearly with ψ. But SE-selection increases more than linearly with ψ. This is because the standard error of $\hat{\alpha}_1$ can be written as $c\sqrt{1/(1-\psi^2)}$, where c depends on the variances of Y and X_1, on the R^2 of the regression, and on the sample size, but is invariant to ψ. So the standard error increases more than linearly with ψ, approaching infinity as ψ approaches one.) Hence, this scenario still allows us to control the relative magnitude of SE-selection versus E-selection, albeit indirectly and imperfectly. It is not possible, however, to fully decouple the two flavors of selection in this simulation environment, unlike in the cheating scenario.

On a technical note, we limit the number of control variables attempted by p-hackers to H. If, at the H-th attempt, the estimate of α_1 remains negative or statistically insignificant, the p-hacker gives up the p-hacking search and resorts instead to an entirely new dataset, starting the process all over again. We do so because it may be extremely difficult (and time-consuming, from a computational perspective) in some cases to p-hack a very negative estimate into a significantly positive one. This is especially the case when α_1 is very small, which, by sampling error alone, may generate substantially negative estimates in some datasets.

5.1.3 Parameter Values and Distributions

We implement p-hacking selection using the parameter values and distributions summarized in Table 4. A key parameter in the simulations is ψ, the correlation between X_1 and X_2, since it governs the relative degree of SE-selection versus E-selection. The higher its values, the

Table 4: Parameter values in the p-hacking selection scenario

Parameter/Variable	Description	Values/distribution
X_1	Main regressor of the primary model	$\sim U(0,1)$
X_2	Control variable of the primary model	$\psi X_1 + (1-\psi)\epsilon$
ϵ	Independent stochastic component of X_2	$\sim N(0,1)$
ψ	Correlation coefficient between X_1 and X_2	0.5, 0.6, 0.7, 0.8, 0.9
α_1	Size of the true effect	0, 1
α_2	Slope coefficient of X_2	2
u	Error term of the primary model	$\sim N(0,\sigma_u^2)$
σ_u^2	Variance of the error term	5.06
N_i	Sample size of the primary study	$\sim \Gamma(a,b)$
a	Parameter of the gamma distribution	0.65
b	Parameter of the gamma distribution	731
M	Number of studies/estimates	80
π	Fraction of potential p-hackers	0.5
H	p-hacking attempts before drawing new data	50

Notes: See text for explanation of the chosen values and distributions. When possible, in calibration we follow the tradition of previous simulations built for meta-analysis in the context of regression estimates.[9,11,61,91]

larger the relative degree of SE-selection. (We can quantify, for each value of ψ, the relative importance of SE-selection that would correspond to parameter ϕ in the cheating selection case; see below.) We let ψ take on values from 0.5 to 0.9 in steps of 0.1, using the middle value of 0.7 as the baseline value for the calibration of other parameters. In line with related simulation studies,[9,11] we assume a meta-analysis of $M = 80$ studies and a probability of engaging in publication selection (in this context, the fraction of potential p-hackers) of $\pi = 50\%$. The maximum number of p-hacking attempts before drawing new data is set at $H = 50$.

Regarding the true effect, we consider the cases where it is nil ($\alpha_1 = 0$) and where it is positive ($\alpha_1 = 1$). We do not separately consider a case of a "large effect" as in the cheating scenario, because once again the results would be qualitatively identical to $\alpha_1 = 1$, so we have just one value for the positive effect. We set the remaining parameters (especially σ_u^2) so that $\alpha_1 = 1$ is neither too small nor too large an effect. If σ_u^2 is too large, $\alpha_1 = 1$ effectively represents a small effect. Conversely, when σ_u^2 is too small, $\alpha_1 = 1$ effectively represents a large effect. Moreover, the larger the effective size of α_1, the smaller the effective incidence of publication selection, eventually dropping to zero. For $\alpha_1 = 0$, the effective incidence of selection is about 49%. (This assumes that primary studies test the null hypothesis that $\alpha_1 = 0$ using a two-sided test at the 5% level. Hence, the probability of not finding a significantly positive value is 97.5%. Because only half of the studies engage in publication selection, the effective selection incidence is half of this rate, that is, 48.75%.) We then choose σ_u^2 so that the effective selection incidence for $\alpha_1 = 1$ is half of the incidence for $\alpha_1 = 0$—that is, 24%. The implied value of σ_u^2 is 5.06.

The p-hacking scenario generates heterogeneity. Given σ_u^2, α_1, and ψ, the main parameter determining the degree of parameter heterogeneity is α_2. Based on the typical findings of

applied meta-analyses, simulation studies[9,11] often assume values of I^2 of at least 70%. By setting $\alpha_2 = 2$, we arrive at an I^2 of about 73% for $\alpha_1 = 0$ (for $\alpha_1 = 1$, the I^2 is about half). The sample size of a primary study, N_i, is drawn from a truncated gamma distribution $\Gamma(a, b)$. Note that the mean of this distribution is given by ab and the variance by ab^2. We choose the values of a and b to match the research record. Using a database of 436 meta-analyses in economics provided to us by Chris Doucouliagos,[10] we find the medians of the mean and variance of the sample sizes within the individual meta-analyses to be 473 and 588^2, respectively; using these as target values, we find the required gamma parameters to be $a = 0.65$ and $b = 731$. We truncate the distribution from below, so that a sample size is never smaller than 30.

5.1.4 Implied Relative Degree of SE-Selection

As mentioned above, we control the relative degrees of selection on estimates and selection on standard errors indirectly through the parameter ψ. In the cheating selection scenario, this relative degree was controlled directly through ϕ. Although we cannot control ϕ directly here, we can nevertheless infer its size for each value of ψ. To do so, start by denoting, for the set of selected estimates, the observed (post-selection, hacked) t-statistic of $\hat{\alpha}_1$ by $t = \hat{\alpha}_1/\text{SE}(\hat{\alpha}_1)$ and the original (pre-selection, unhacked) t-statistic by $t^* = \hat{\alpha}_1^*/\text{SE}(\hat{\alpha}_1)^*$. Of course, the objective of selection is to increase the size of the t-statistic, so $t > t^*$. E-selection implies $\hat{\alpha}_1 > \hat{\alpha}_1^*$ and SE-selection implies $\text{SE}(\hat{\alpha}_1) < \text{SE}(\hat{\alpha}_1)^*$. In the p-hacking scenario, however, both types usually occur simultaneously and ϕ measures the relative importance of each. Because $t/t^* = (\hat{\alpha}_1/\hat{\alpha}_1^*) \times (\text{SE}(\hat{\alpha}_1)^*/\text{SE}(\hat{\alpha}_1))$, it follows that

$$\ln\left(\frac{t}{t^*}\right) = \ln\left(\frac{\hat{\alpha}_1}{\hat{\alpha}_1^*}\right) + \ln\left(\frac{\text{SE}(\hat{\alpha}_1)^*}{\text{SE}(\hat{\alpha}_1)}\right),$$

which decomposes the amount of publication selection in selected estimates (percent change of the t-statistic) into its E-selection component (given by the first term, the percent increase of $\hat{\alpha}_1$ after selection) and its SE-selection component (given by the second term, the percent decrease in $\text{SE}(\hat{\alpha}_1)$ after selection). Hence, the relative importance of SE-selection can be approximated by the relative size of the second term:

$$\phi = \frac{\ln(\text{SE}(\hat{\alpha}_1)^*/\text{SE}(\hat{\alpha}_1))}{\ln(t/t^*)}. \tag{5}$$

On a technical note, we need to impose some restrictions to ensure that $0 \leq \phi \leq 1$. If, for a particular selected estimate, $\text{SE}(\hat{\alpha}_1) > \text{SE}(\hat{\alpha}_1)^*$, then selection must have occurred entirely through the estimates and we set $\phi = 0$. If, on the other hand, $\hat{\alpha}_1 < \hat{\alpha}_1^*$, then selection must have occurred through the standard errors, and we set $\phi = 1$. Table 5 shows the values of ϕ corresponding to the various values of ψ. Clearly, the relative importance of SE-selection increases with ψ.

Table 5: The fraction of SE-selection corresponding to correlation between regressors

True effect (α_1)	Correlation (ψ)				
	0.5	0.6	0.7	0.8	0.9
0	0.009	0.015	0.041	0.095	0.216
1	0.026	0.040	0.102	0.207	0.370

Notes: The table shows the fraction of selection on standard errors relative to selection on estimates (ϕ in the cheating scenario) mapped to the correlation coefficient between X_1 and X_2 (ψ in the p-hacking scenario) and the true effect. For example, if the correlation is 0.9 and the true effect is 1, the implied fraction of SE-selection is 0.37.

5.2 Results

In the p-hacking selection environment we simulate all estimators in all versions described earlier in Table 2. The results are reported following the same structure as in the cheating simulation scenario. Figures 6-7 show the bias and coverage for no effect ($\alpha_1 = 0$) and various values of ψ. Figures 8-9 show the results for a positive effect ($\alpha_1 = 1$). Results showing MSE are available in Appendix B. Note that the horizontal axis now does not measure the relative degree of SE-selection, but can be recomputed to that degree using Table 5.

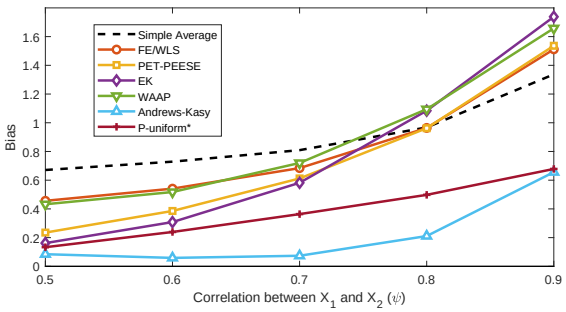
Regarding Figure 6, the simple average measures the effect size inflation caused by E-selection. It increases with ψ, because E-selection increases with ψ: the larger its value, the larger the (positive) biases in the individual studies that are subject to selection. All results should thus be evaluated relative to the simple average. SE-selection also increases with ψ, and more than proportionally so (see Table 5). For $\psi = 0.5$, SE-selection is fairly low and all methods correct a large chunk of the selection bias. As ψ goes up, so does the relative importance of SE-selection. As a consequence, the capacity of these methods to correct for the increasing bias deteriorates rapidly. For most methods, the biases due to spurious precision eventually surpass the selection bias itself; i.e., the methods actually worsen the existing publication bias. Notable exceptions are Andrews-Kasy and p-uniform*, which always correct part of the selection bias.

Our proposed adjustment improves all methods except Andrews-Kasy and p-uniform*—where, as we have noted, the adjustment has weaker statistical justification. In EK and PET-PEESE, it works better when instrumenting the SEs in the meta-regression and using no weights. Thus corrected, these estimators tend to perform the best, especially for large ψ. The profiles of MSE in Figure 15 in Appendix B are very similar to the profiles of bias in Figure 6.

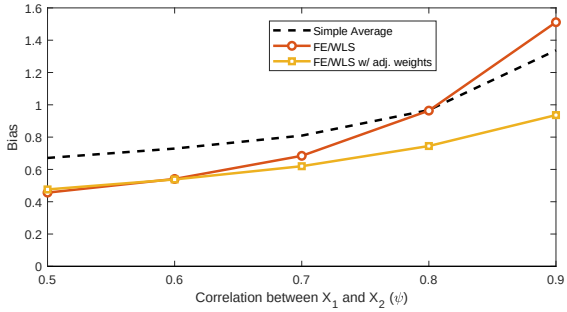
Regarding Figure 7, the coverage rates also decrease markedly with ψ. Only Andrews-Kasy manages to sustain acceptable coverage for ψ up to 0.75. At this value of ψ, all the other estimators show very low coverage rates of 20% or even lower. Our adjustment works relatively well in PET-PEESE and EK, especially when no weights are used, with coverage rates above 80% (still below the 95% nominal level, though). It's important to emphasize the role of the weights as ψ increases. Only instrumenting the SEs in the meta-regression does not work.

Figure 6: Bias: p-hacking selection, no effect ($\alpha_1 = 0$), various values of ψ

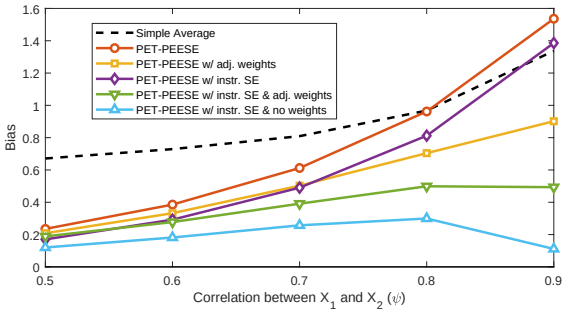
(a) Unadjusted estimators



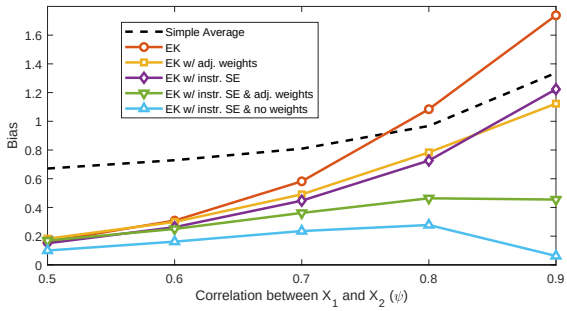
(b) Adjusting FE/WLS



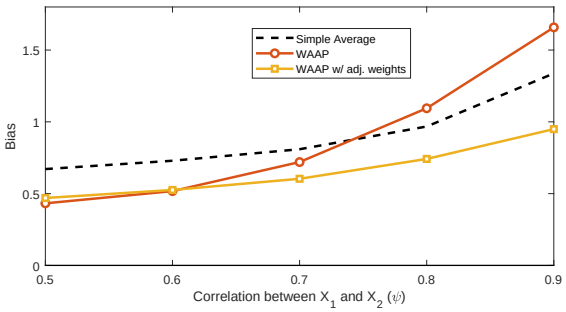
(c) Adjusting PET-PEESE



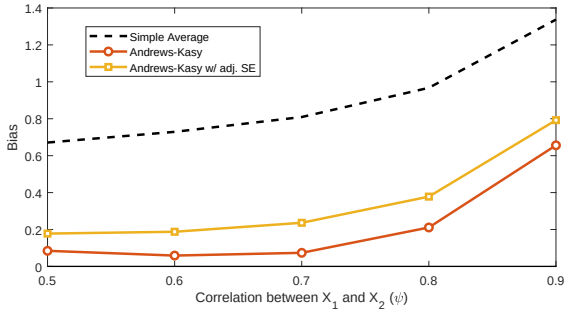
(d) Adjusting EK



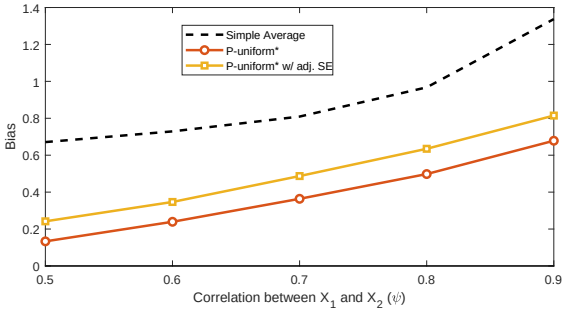
(e) Adjusting WAAP



(f) Adjusting Andrews-Kasy



(g) Adjusting p-uniform*



(h) Comparison of adjusted estimators

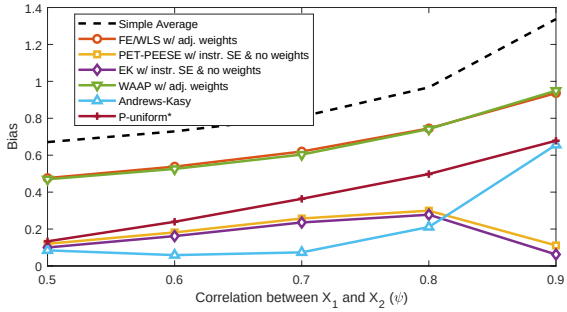
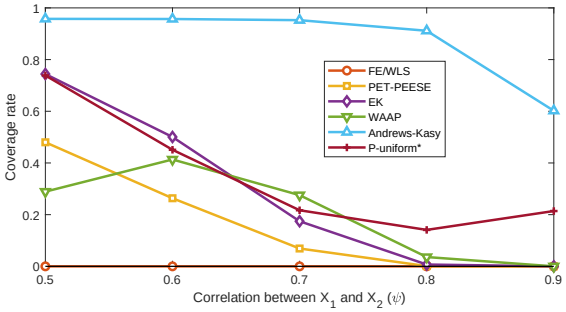
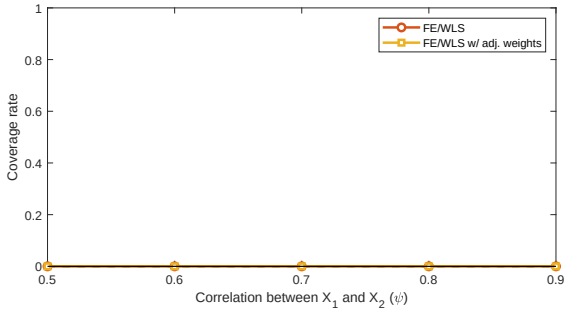


Figure 7: Coverage: p-hacking selection, no effect ($\alpha_1 = 0$), various values of ψ

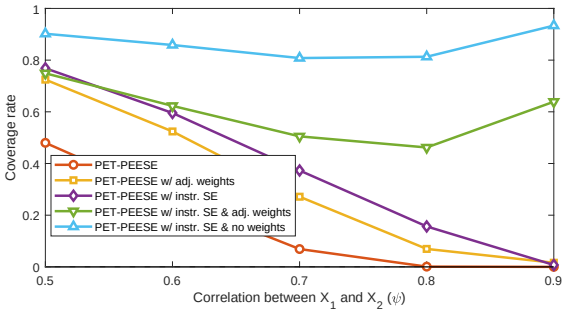
(a) Unadjusted estimators



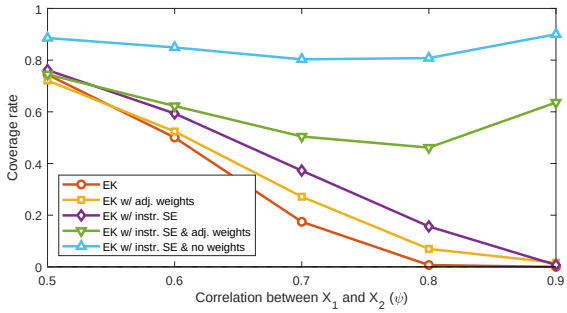
(b) Adjusting FE/WLS



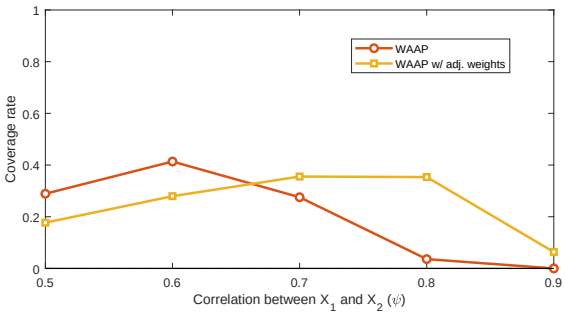
(c) Adjusting PET-PEESE



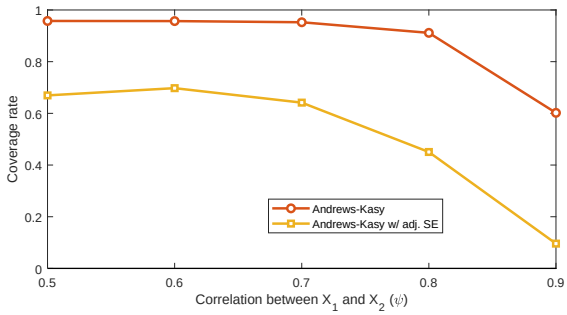
(d) Adjusting EK



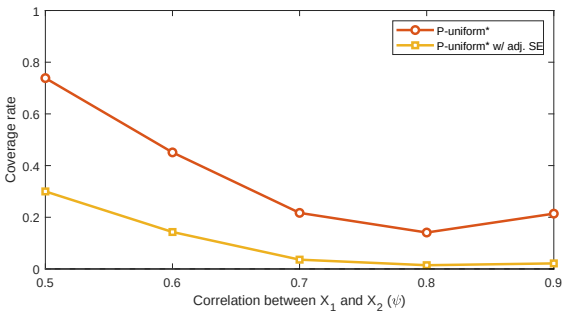
(e) Adjusting WAAP



(f) Adjusting Andrews-Kasy



(g) Adjusting p-uniform*



(h) Comparison of adjusted estimators

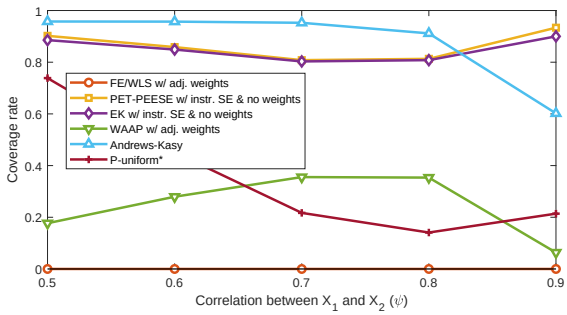
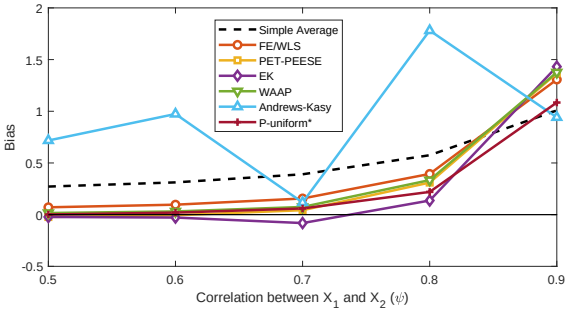
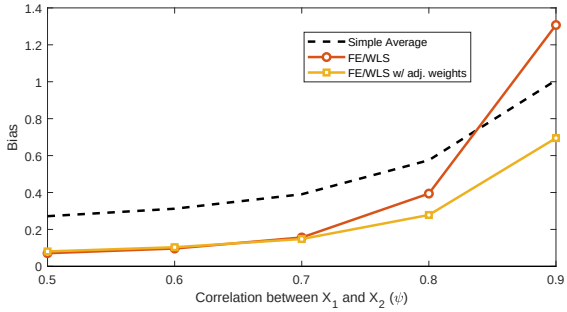


Figure 8: Bias: p-hacking selection, positive effect ($\alpha_1 = 1$), various values of ψ

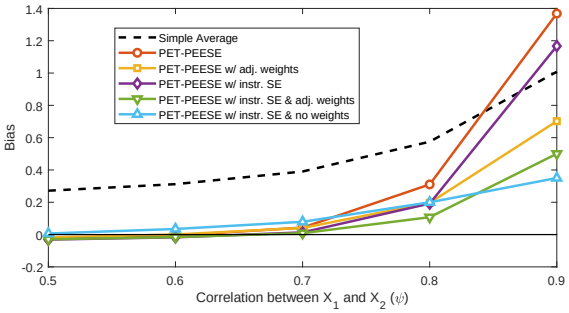
(a) Unadjusted estimators



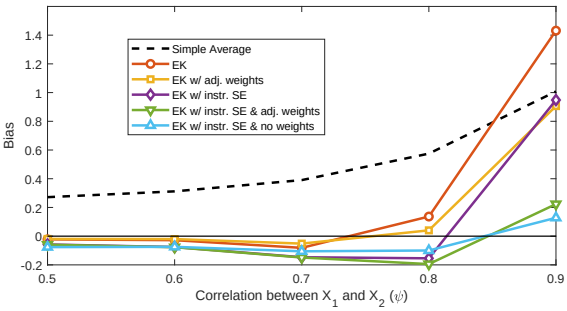
(b) Adjusting FE/WLS



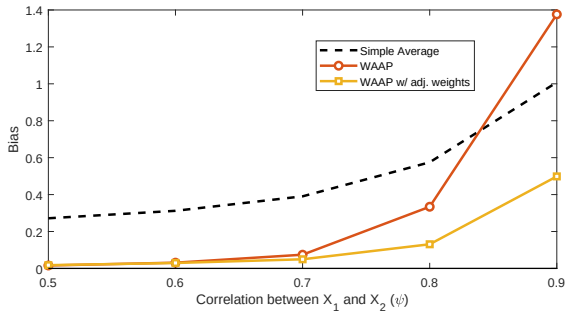
(c) Adjusting PET-PEESE



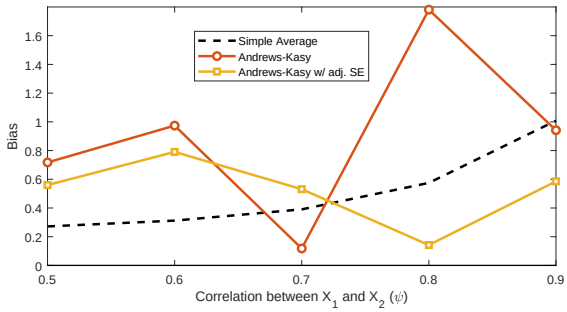
(d) Adjusting EK



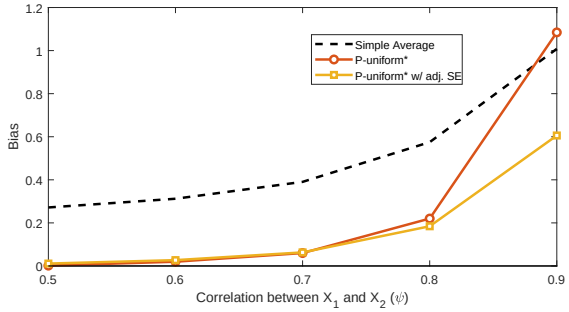
(e) Adjusting WAAP



(f) Adjusting Andrews-Kasy



(g) Adjusting p-uniform*



(h) Comparison of adjusted estimators

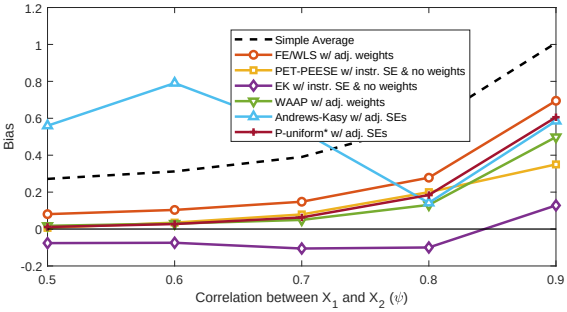
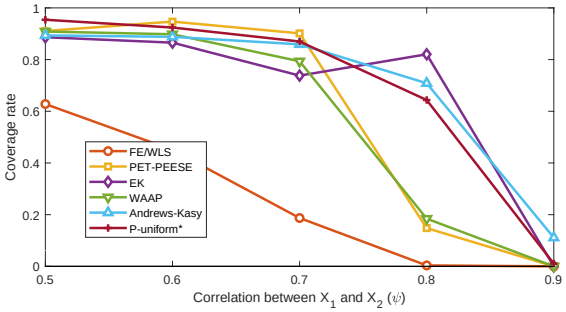
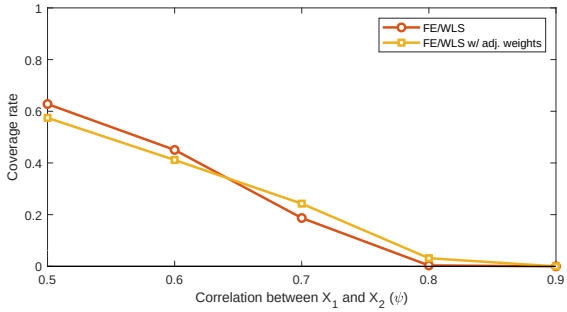


Figure 9: Coverage: p-hacking selection, positive effect ($\alpha_1 = 1$), various values of ψ

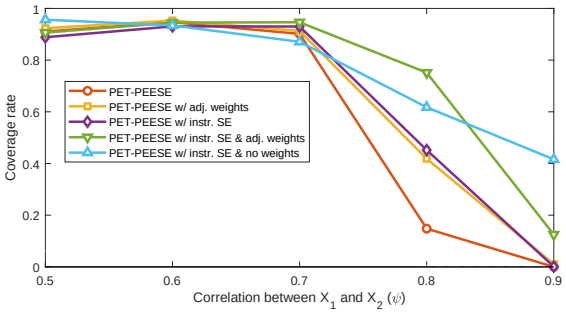
(a) Unadjusted estimators



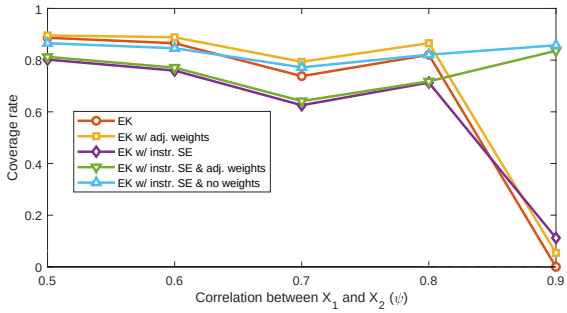
(b) Adjusting FE/WLS



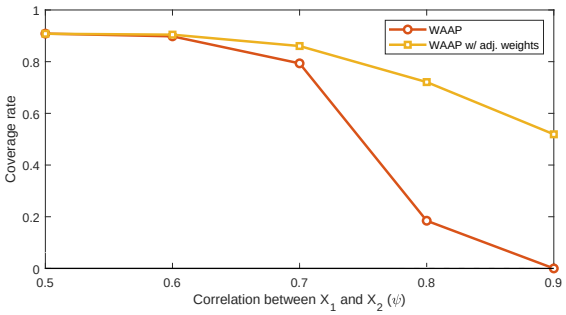
(c) Adjusting PET-PEESE



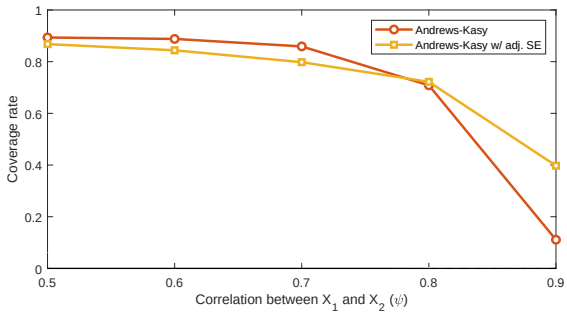
(d) Adjusting EK



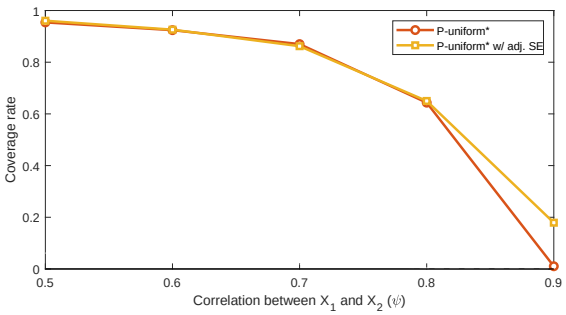
(e) Adjusting WAAP



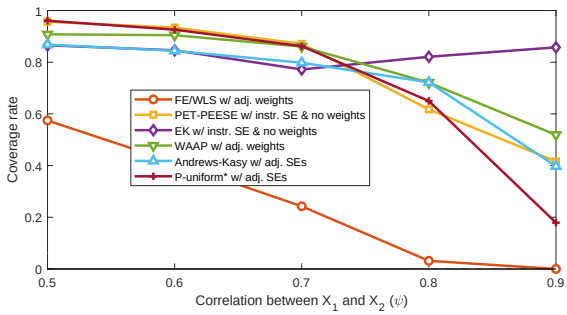
(f) Adjusting Andrews-Kasy



(g) Adjusting p-uniform*



(h) Comparison of adjusted estimators



Better, though still insufficient, is to also adjust the weights; much better is to drop them altogether. For Andrews-Kasy and p-uniform*, our adjustment does not work at all.

Regarding Figures 8-9, the results for $\alpha_1 = 1$ are qualitatively similar to those seen above for $\alpha_1 = 0$. Quantitatively, the biases and MSEs are now typically smaller and coverage is better; this is because a larger true effect reduces the overall size of publication bias (less p-hacking is needed). All methods still show a decreasing capacity to correct for publication bias as ψ increases. Eventually, the biases of the correction methods become larger than publication bias itself (measured by the simple average), although this happens at a higher ψ than before. The results of Andrews and Kasy's model in particular are volatile, which in our experience is a common feature of selection models in meta-analyses of typical size. Our adjustment now improves the performance of all methods, including selection models. In PET-PEESE, adjusting the weights is comparable to dropping them, and a similar observation applies to EK. WAAP with adjusted weights and SEs now does almost as well as PET-PEESE.

The reader may be surprised that the bias of unadjusted methods is now positive, while in the cheating scenario the bias was negative (both for $\alpha_1 = 1$). The reason is that in the p-hacking scenario, selection on estimates and selection on standard errors (or, analogously, the Lombard effect and Taylor's rule) interact. In the cheating scenario the bias is always downwards if $\alpha_1 > 0$, because small estimates are assigned small standard errors to be significant, which gives small estimates too much weight. In the p-hacking scenario estimates can be too large and too precise at the same time, which creates the upward bias shown in Figure 8.

In a nutshell, the more realistic p-hacking simulation implies that the MAIVE version of PET-PEESE without additional inverse-variance weights is more robust to spurious precision than available alternatives. When spurious precision is important, the method clearly dominates unadjusted estimators, including selection models. When the degree of spurious precision is negligible, the performance of MAIVE is similar to unadjusted methods, though it can be sometimes (especially for a zero true effect) beaten by selection models. Based on the simulation results and Table 5, the simple unweighted mean can often beat sophisticated unadjusted estimators for a ratio of SE-selection as low as 0.2 relative to total amount of selection. Cures to publication bias can be worse than the disease.

6 Conclusion

We do not argue that spurious precision is common. We argue that it can plausibly arise in observational research. Even in experimental settings, randomization can fail,[92] and authors often use regressions to control for pre-treatment covariates or make other adjustments[93] that can yield spurious precision. When it arises, a small dose can render the simple mean more reliable than sophisticated correction techniques. The Meta-Analysis Instrumental Variable Estimator (MAIVE) solves the problem by using inverse sample size as an instrument for reported variance. That is, we regress the reported squared standard errors on the inverse of the number of observations used in the primary study. The fitted values from this regression are then used

instead of reported variance in the PEESE meta-regression. Standard weighted means, funnel plots, and funnel-based methods can be adjusted similarly to make them robust to spurious precision. The entire meta-analysis toolkit can be salvaged with this modification.

The instrumental approach has seven benefits over using sample size as a proxy for precision, and we explain them in Section 3. There are at least two costs as well, both compared to the proxy approach and the classical one that relies on reported precision. First, MAIVE is more complex since it involves an additional regression and computation of fitted values and valid confidence intervals. But the instrumental approach is available in most statistical programs. We create the `maive` package for R, which makes estimation easy for meta-analysts unfamiliar with instrumental variables. Second, the additional regression makes MAIVE noisier compared to conventional techniques. When a meta-analyst is sure there can be no spurious precision in her data, using reported precision without instruments will yield unbiased and more efficient meta-analysis estimates. The lack of spurious precision can be tested approximately by employing the Hausman specification test:[94] if the coefficients estimated in MAIVE are far from those of an unadjusted PEESE, spurious precision is likely an issue.

A discussion is in order regarding the application of MAIVE—pronounced, by the way, as the Irish name Maeve. The instrument is the overall sample size, not degrees of freedom, because the latter depends on clustering units. We prefer the MAIVE version of PEESE without weights (after testing with unweighted MAIVE-PET whether the true effect is nonzero). This parsimonious specification intuitively fits both panels of Figure 1 in the Introduction. The `maive` package allows for optional adjusted weights. Researchers may choose a MAIVE version of another estimator, such as endogenous kink. The package also runs the Hausman test. Because PEESE is heteroskedastic by definition and we prefer not to use inverse-variance weights, the package produces heteroskedasticity-robust standard errors by default. When studies report multiple estimates, standard errors in MAIVE—and any meta-analysis estimator—should be clustered at the study level, again a default option. With fewer than 30 studies we recommend wild bootstrap.[21] It is a good idea to include study-level dummies (econometric fixed effects) to filter out study-specific idiosyncrasies related to unobserved heterogeneity. The package also reports a robust F-statistic of the first-stage regression. If the F-statistic is below 10, the instrument is weak and MAIVE results should be treated with caution.[89] Researchers may want to use confidence intervals robust to weak instruments.[86–88]

The reader will object that our simulation is unfair to correction methods. The methods were designed to counter publication bias; we simulate p-hacking. Individual estimates and standard errors get biased, which is why selection models do not work well here—though they do not assume, as funnel methods assume, that selection works only on estimates (the Lombard effect discussed in the Introduction). The distinction between publication bias and p-hacking is clear in theory, but in practice both are often observationally equivalent to the meta-analyst. (But p-hacking likely predominates.[95]) As long as we believe our p-hacking environment is broadly realistic, we need a technique that corrects the resulting bias. MAIVE is the only such technique. One can design p-hacking scenarios in which misspecifications make it almost

impossible for meta-analysis methods to uncover the true unconditional mean.[92,96] If that is a realistic description of observational research, unconditional meta-analysis means are meaningless.[97] MAIVE can be extended to allow for observed heterogeneity and deliver context-specific means via incorporation into Bayesian model averaging meta-regression approaches addressing model uncertainty.[45–49]

We leave questions open regarding spurious precision. How common is it in practice? How does measurement error influence the relative performance of MAIVE? What happens when method heterogeneity explicitly affects both estimates and their precision? Does spurious precision help explain why meta-analyses exaggerate the true effect in comparison to multilab preregistered replications?[70,98,99] How to correctly adjust selection models for spuriousness? The last is perhaps the most important question for future research, because many meta-analysts prefer selection models over funnel-based techniques.[76] The adjustment of selection models is not straightforward since here precision has two intertwined roles: identification and weighting. For identification, we need the reported, nominal precision, which determines statistical significance. But for weights we need the underlying, true precision. The maximum likelihood approach has to be modified to allow a different measure of precision for each role.

The bottom line is that spurious precision, while plausibly destructive, is surmounted by adjusting funnel-based methods.

Highlights

What is already known

- In meta-analysis it is optimal to give more weight to more precise studies.
- Inverse-variance weighting maximizes efficiency and may attenuate publication bias.
- Inverse-variance weighting is used by all common estimators.

What is new

- If reported precision exaggerates real precision, inverse-variance weighting creates a bias.
- With enough spurious precision, cures to publication bias are worse than the disease.
- Spurious precision arises naturally in observational research via p-hacking.
- Meta-Analysis Instrumental Variable Estimator (MAIVE) corrects for spurious precision.

Potential impact

- Meta-analysts should use MAIVE if they suspect p-hacking.
- The difference between MAIVE and unadjusted estimators can measure spurious precision.
- MAIVE substantially improves the robustness of the current meta-analysis toolkit.

References

1. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. *Nature* 2018; 555: 175–182.
2. Borenstein M, Hedges L, Higgins J, Rothstein H. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* 2010; 1(2): 97–111.
3. Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine* 2015; 34(13): 2116-2127.
4. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; 315(7109): 629–634.
5. Duval S, Tweedie R. Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56(2): 455–463.
6. Stanley TD. Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection. *Oxford Bulletin of Economics and Statistics* 2008; 70(1): 103–127.
7. Stanley TD, Jarrell SB, Doucouliagos H. Could It Be Better to Discard 90% of the Data? A Statistical Paradox. *The American Statistician* 2010; 64(1): 70–77.
8. Stanley TD, Doucouliagos H. *Meta-regression analysis in economics and business.* NY: Routledge. 2012.
9. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods* 2014; 5(1): 60–78.
10. Ioannidis JP, Stanley TD, Doucouliagos H. The Power of Bias in Economics Research. *The Economic Journal* 2017; 127(605): F236–F265.
11. Bom PRD, Rachinger H. A kinked meta-regression model for publication bias correction. *Research Synthesis Methods* 2019; 10(4): 497–514.
12. Furukawa C. Publication Bias under Aggregation Frictions: Theory, Evidence, and a New Correction Method. *MIT* 2019; working paper. www.jeameetings.org/2019s/Gabst/1161.pdf.
13. Hedges L. Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* 1984; 9: 61–85.
14. Iyengar S, Greenhouse JB. Selection Models and the File Drawer Problem. *Statistical Science* 1988; 3(1): 109–117.
15. Hedges LV. Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science* 1992; 72(2): 246–255.
16. Vevea J, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* 1995; 60(3): 419–435.
17. Andrews I, Kasy M. Identification of and correction for publication bias. *American Economic Review* 2019; 109(8): 2766–2794.
18. Aert vRC, Assen vM. Correcting for publication bias in a meta-analysis with the p-uniform* method. *Tilburg University & Utrecht University* 2021; working paper. doi: https://doi.org/10.31222/osf.io/zqjr9
19. Abadie A, Athey S, Imbens GW, Wooldridge JM. When Should You Adjust Standard Errors for Clustering?. *The Quarterly Journal of Economics* 2022; 138(1): 1–35.
20. Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* 2015; 50(2): 317–372.
21. Roodman D, Nielsen MØ, MacKinnon JG, Webb MD. Fast and Wild: Bootstrap Inference in Stata Using Boottest. *The Stata Journal* 2019; 19(1): 4–60.
22. White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 1980; 48(4): 817–838.
23. Bom PRD, Ligthart JE. What Have We Learned From Three Decades of Research on the Productivity of Public Capital?. *Journal of Economic Surveys* 2014; 28(5): 889–916.
24. Askarov Z, Doucouliagos A, Doucouliagos H, Stanley TD. The Significance of Data-Sharing Policy. *Journal of the European Economic Association* 2023; forthcoming. doi: doi.org/10.1093/jeea/jvac053
25. Putz P, Bruns SB. The (Non-)Significance Of Reporting Errors In Economics: Evidence From Three Top

Journals. *Journal of Economic Surveys* 2021; 35(1): 348-373.

26. Lane H, Tranel B. The Lombard Sign and the Role of Hearing in Speech. *Journal of Speech and Hearing Research* 1971; 14(4): 677–709.
27. McCloskey DN, Ziliak ST. What quantitative methods should we teach to graduate students? A comment on Swann's Is precise econometrics an illusion?. *The Journal of Economic Education* 2019; 50(4): 356–361.
28. Kunc HP, Morrison K, Schmidt R. A meta-analysis on the evolution of the Lombard effect reveals that amplitude adjustments are a widespread vertebrate mechanism. *Proceedings of the National Academy of Sciences* 2022; 119(30): e2117809119.
29. Hedges LV. A random effects model for effect sizes. *Psychological Bulletin* 1983; 93(2): 388–395.
30. Hedges LV, Olkin I. *Statistical methods for meta-analysis*. Orlando, FL: Academic Press. 1985.
31. Taylor LR. Aggregation, variance and the mean. *Nature* 1961; 189(4766): 732–735.
32. Cohen JE, Xu M. Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. *Proceedings of the National Academy of Sciences* 2015; 112(25): 7749-7754.
33. Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-regression. *Research Synthesis Methods* 2017; 8(1): 19–42.
34. Havranek T, Stanley TD, Doucouliagos H, et al. Reporting Guidelines for Meta-Analysis in Economics. *Journal of Economic Surveys* 2020; 34(3): 469–475.
35. Ugur M, Awaworyi Churchill S, Luong H. What do we know about R&D spillovers and productivity? Meta-analysis evidence on heterogeneity and statistical power. *Research Policy* 2020; 49(1): 103866.
36. Xue X, Reed WR, Menclova A. Social capital and health: A meta-analysis. *Journal of Health Economics* 2020; 72(C): 102317.
37. Neisser C. The Elasticity of Taxable Income: A Meta-Regression Analysis. *Economic Journal* 2021; 131(640): 3365–3391.
38. Zigraiova D, Havranek T, Irsova Z, Novak J. How puzzling is the forward premium puzzle? A meta-analysis. *European Economic Review* 2021; 134(C): 103714.
39. Nakagawa S, Lagisz M, Jennions MD, et al. Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution* 2022; 13(1): 4–21.
40. Brown AL, Imai T, Vieider F, Camerer C. Meta-Analysis of Empirical Estimates of Loss-Aversion. *Journal of Economic Literature* 2023; forthcoming. doi: 10.1257/jel.20221698
41. Heimberger P. Do Higher Public Debt Levels Reduce Economic Growth?. *Journal of Economic Surveys* 2023; forthcoming. doi: 10.1111/joes.12536
42. Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science* 2019; 2(2): 115-144.
43. Brodeur A, Cook N, Heyes A. Methods Matter: P-Hacking and Causal Inference in Economics. *American Economic Review* 2020; 110(11): 3634–3660.
44. DellaVigna S, Linos E. RCTs to Scale: Comprehensive Evidence From Two Nudge Units. *Econometrica* 2022; 90(1): 81–116.
45. Gechert S, Havranek T, Irsova Z, Kolcunova D. Measuring Capital-Labor Substitution: The Importance of Method Choices and Publication Bias. *Review of Economic Dynamics* 2022; 45(C): 55–82.
46. Imai T, Rutter TA, Camerer CF. Meta-Analysis of Present-Bias Estimation Using Convex Time Budgets. *The Economic Journal* 2021; 131(636): 1788–1814.
47. Gechert S, Heimberger P. Do corporate tax cuts boost economic growth?. *European Economic Review* 2022; 147(C): 104157.
48. Havranek T, Irsova Z, Laslopova L, Zeynalova O. Publication and Attenuation Biases in Measuring Skill Substitution. *The Review of Economics and Statistics* 2023; forthcoming. doi: 10.1162/rest_a_01227
49. Matousek J, Havranek T, Irsova Z. Individual discount rates: A meta-analysis of experimental evidence. *Experimental Economics* 2022; 25(1): 318–358.
50. Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-Drawer. *Journal of Experimental Psychology: General* 2014; 143(2): 534–547.

51. Simonsohn U, Nelson LD, Simmons JP. p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science* 2014; 9(6): 666-681. PMID: 26186117.

52. Assen vM, Aert vRC, Wicherts JM. Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods* 2015; 20(3): 293–309.

53. Simonsohn U, Simmons JP, Nelson LD. Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General* 2015; 144(6): 1146–1152.

54. Aert vRC, Assen vM. Bayesian evaluation of effect size after replicating an original study. *Plos ONE* 2017; 12(4): e0175302.

55. Aert vRC, Assen vM. Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods* 2018; 50: 1515–1539.

56. Stanley TD. Beyond Publication Bias. *Journal of Economic Surveys* 2005; 19(3): 309–345.

57. Kranz S, Putz P. Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment. *American Economic Review* 2022; 112(9): 3124–3136.

58. Sanchez-Meca J, Marín-Martínez F. Weighting by Inverse Variance or by Sample Size in Meta-Analysis: A Simulation Study. *Educational and Psychological Measurement* 1998; 58(2): 211-220.

59. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews. *Journal of Clinical Epidemiology* 2005; 58(9): 882–893.

60. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of Two Methods to Detect Publication Bias in Meta-analysis. *JAMA* 2006; 295(6): 676-680.

61. Hong S, Reed WR. Using Monte Carlo experiments to select meta-analytic estimators. *Research Synthesis Methods* 2021; 12(2): 192–215.

62. Schmidt FL, Oh IS, Hayes TL. Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology* 2009; 62(1): 97–128.

63. Stanley TD, Doucouliagos H, Ioannidis JPA. Beyond Random Effects: When Small-Study Findings Are More Heterogeneous. *Advances in Methods and Practices in Psychological Science* 2022; 5(4): 25152459221120427.

64. Bartos F, Gronau QF, Timmers B, Otte WM, Ly A, Wagenmakers EJ. Bayesian model-averaged meta-analysis in medicine. *Statistics in Medicine* 2021; 40(30): 6743–6761.

65. Bartos F, Maier M, Quintana DS, Wagenmakers EJ. Adjusting for Publication Bias in JASP and R: Selection Models, PET-PEESE, and Robust Bayesian Meta-Analysis. *Advances in Methods and Practices in Psychological Science* 2022; 5(3): 1–19.

66. Maier M, Bartos F, T. D. Stanley TD, Shanks D, Harris AJ, Wagenmakers EJ. No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences* 2022; 119(31): e2200300119.

67. Bartos F, Maier M, Wagenmakers EJ, Doucouliagos H, Stanley TD. Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods* 2023; 14(1): 99–116.

68. Maier M, Bartos F, Wagenmakers EJ. Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods* 2023; forthcoming. doi: doi.org/10.1037/met0000405

69. Stanley TD, Doucouliagos H. Harnessing the power of excess statistical significance: Weighted and iterative least squares. *Psychological Methods* 2023; forthcoming. doi: 10.1037/met0000502

70. Kvarven A, Stromland E, Johannesson M. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behavior* 2020; 4: 423–434.

71. Mathur MB, VanderWeele TJ. Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society Series C* 2020; 69(5): 10911119.

72. Copas JB, Li HG. Inference for Non-random Samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1997; 59(1): 55–95.

73. Copas JB. What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 1999; 162(1): 95–109.

74. Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in*

Medical Research 2001; 10(4): 251–265.

75. Copas JB. A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2013; 62(1): 47–66.

76. McShane BB, Böckenholt U, Hansen KT. Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science* 2016; 11(5): 730–749.

77. Mathur MB. Sensitivity analysis for p-hacking in meta-analyses. *Quantitative Sciences Unit and Department of Pediatrics, Stanford University* 2022; working paper. doi: 10.31219/osf.io/ezjsx

78. Hausman J. Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. *Journal of Economic Perspectives* 2001; 15(4): 57–67.

79. Card D, Krueger AB. Time-series minimum-wage studies: A meta-analysis. *American Economic Review* 1995; 85(2): 238–243.

80. Olken BA. Promises and Perils of Pre-analysis Plans. *Journal of Economic Perspectives* 2015; 29(3): 61–80.

81. Havranek T. Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association* 2015; 13(6): 1180–1204.

82. Egger M, Schneider M, Smith GD. Spurious precision? Meta-analysis of observational studies. *BMJ* 1998; 316(7125): 140–144.

83. Hansen TF. On bias and precision in meta-analysis: the error in the error. *Journal of Evolutionary Biology* 2016; 29(10): 1919–1921.

84. Nakagawa S, Noble DWA, Lagisz M, Spake R, Viechtbauer W, Senior AM. A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations. *Ecology Letters* 2023; 26(2): 232-244. doi: https://doi.org/10.1111/ele.14144

85. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001; 54(10): 1046–1055.

86. Andrews I. Conditional Linear Combination Tests for Weakly Identified Models. *Econometrica* 2016; 84(6): 2155–2182.

87. Andrews I. Valid Two-Step Identification-Robust Confidence Sets for GMM. *The Review of Economics and Statistics* 2018; 100(2): 337–348.

88. Sun L. Implementing valid two-step identification-robust confidence sets for linear instrumental-variables models. *Stata Journal* 2018; 18(4): 803–825.

89. Andrews I, Stock JH, Sun L. Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics* 2019; 11(1): 727–753.

90. Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. Questionable research practices in ecology and evolution. *PloS ONE* 2018; 13(7): e0200303.

91. Bom PRD, Rachinger H. A generalized-weights solution to sample overlap in meta-analysis. *Research Synthesis Methods* 2020; 11(6): 812–832.

92. Bruns SB, Ioannidis JP. p-Curve and p-Hacking in Observational Research. *PloS ONE* 2016; 11(2): e0149144.

93. Freedman DA. On regression adjustments to experimental data. *Advances in Applied Mathematics* 2008; 40(2): 180–193.

94. Hausman JA. Specification Tests in Econometrics. *Econometrica* 1978; 46(6): 1251–1271.

95. Brodeur A, Carrell S, Figlio D, Lusher L. Unpacking p-hacking and publication bias. *American Economic Review* 2023; forthcoming. faculty.econ.ucdavis.edu/faculty/scarrell/unpacking.pdf.

96. Bruns SB. Meta-Regression Models and Observational Research. *Oxford Bulletin of Economics and Statistics* 2017; 79(5): 637–653.

97. Simonsohn U, Simmons J, Nelson LD. Above averaging in literature reviews. *Nature Reviews Psychology* 2022; 1: 551–552.

98. Lewis M, Mathur MB, VanderWeele TJ, Frank MC. The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science* 2022; 9(2): 211499.

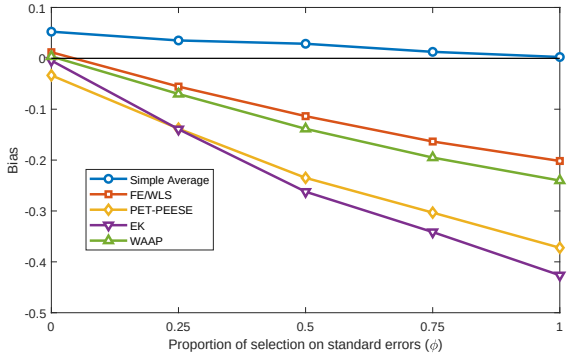
99. Stanley TD, Doucouliagos H, Ioannidis JPA. Retrospective median power, false positive meta-analysis and large-scale replication. *Research Synthesis Methods* 2022; 13(1): 88-108.

Appendices

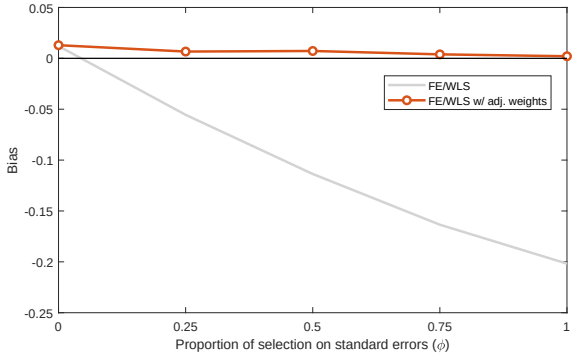
A Cheating Selection for a Large Underlying Effect (for Online Publication)

Figure 10: Bias: cheating selection, large effect ($\alpha_1 = 2$)

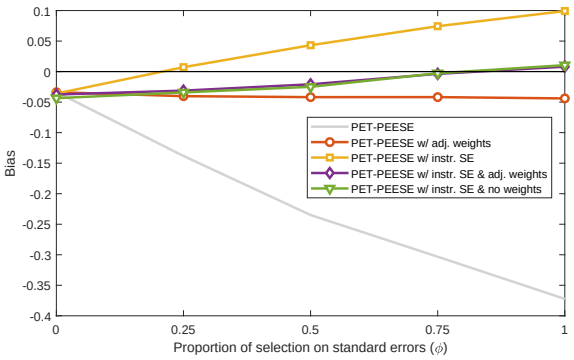
(a) Unadjusted estimators



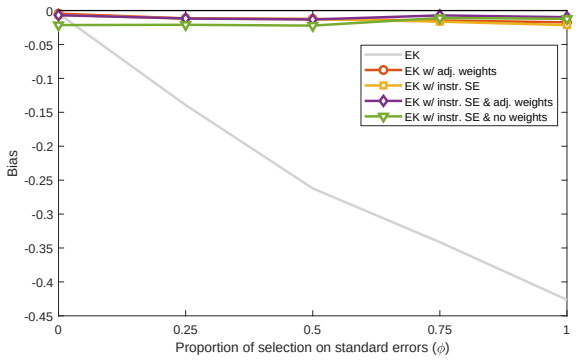
(b) Adjusting FE/WLS



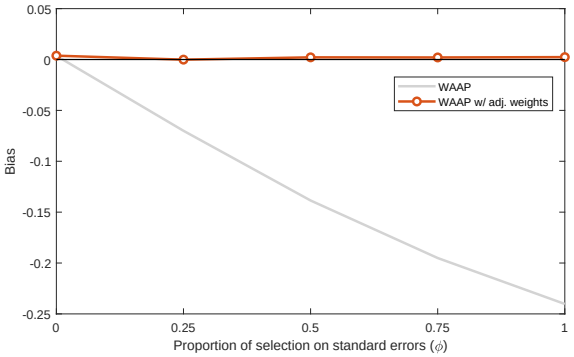
(c) Adjusting PET-PEESE



(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators

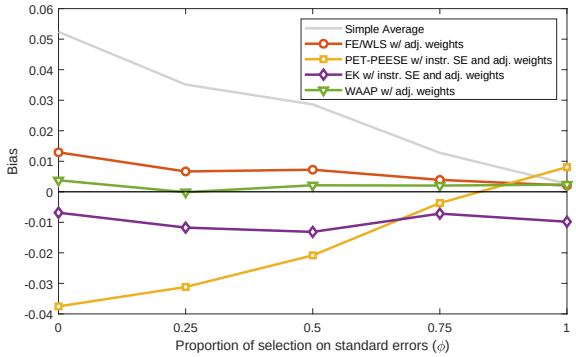
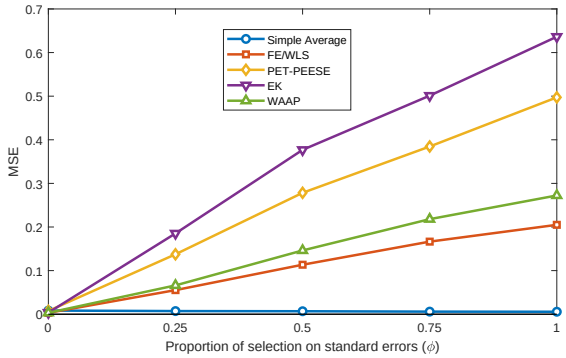
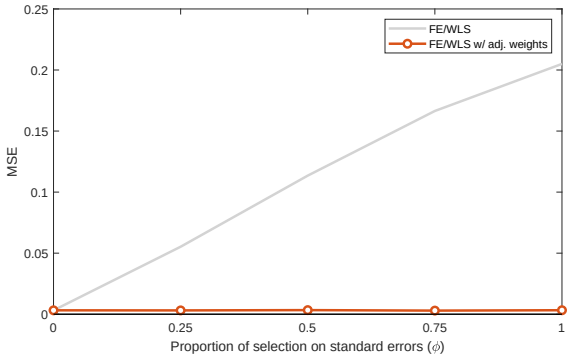


Figure 11: MSE: cheating selection, large effect ($\alpha_1 = 2$)

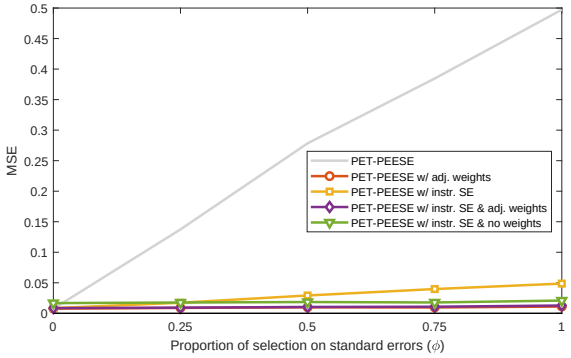
(a) Unadjusted estimators



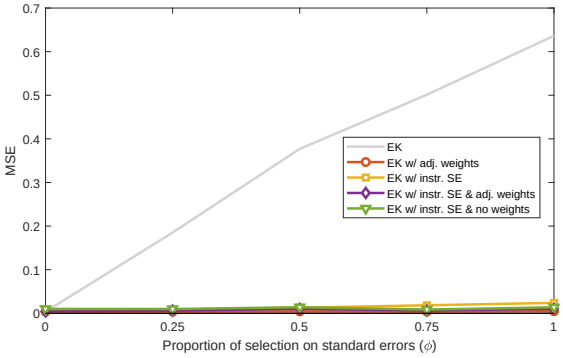
(b) Adjusting FE/WLS



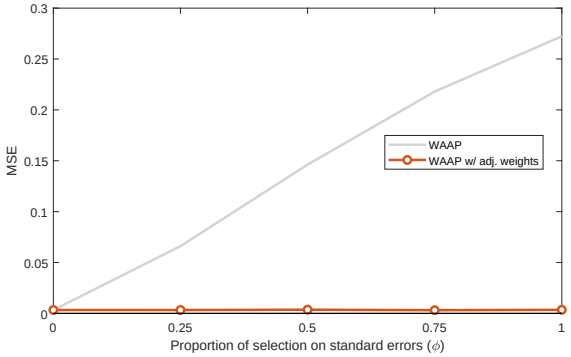
(c) Adjusting PET-PEESE



(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators

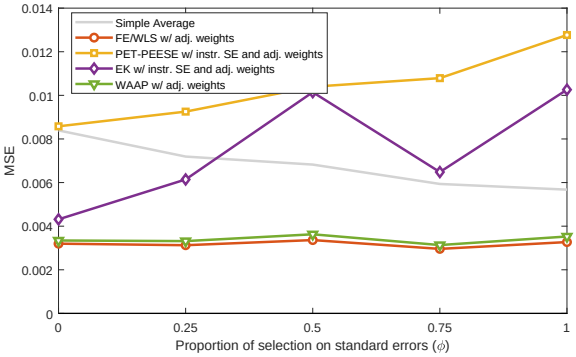
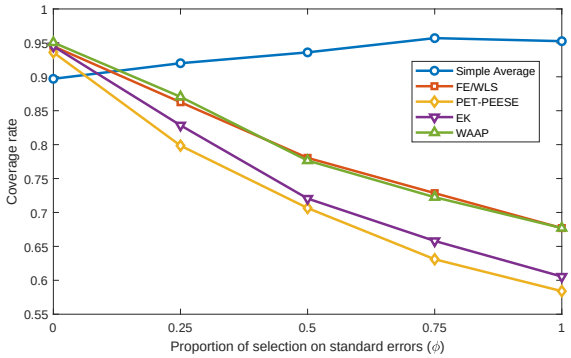
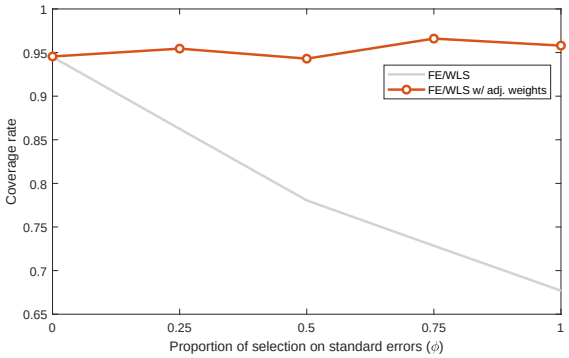


Figure 12: Coverage: cheating selection, large effect ($\alpha_1 = 2$)

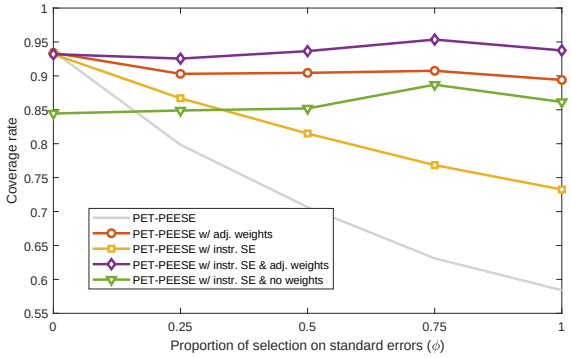
(a) Unadjusted estimators



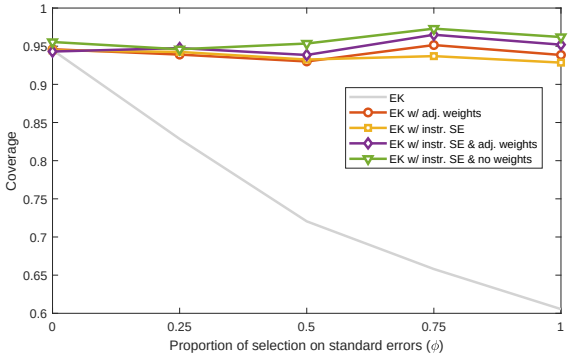
(b) Adjusting FE/WLS



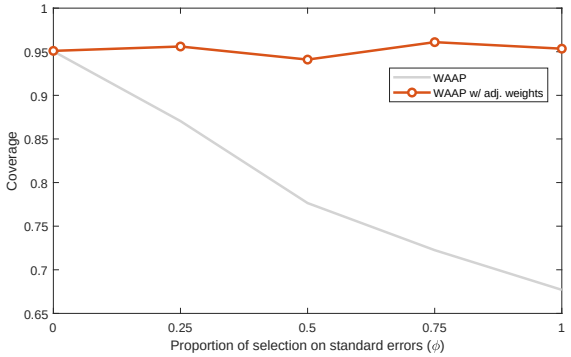
(c) Adjusting PET-PEESE



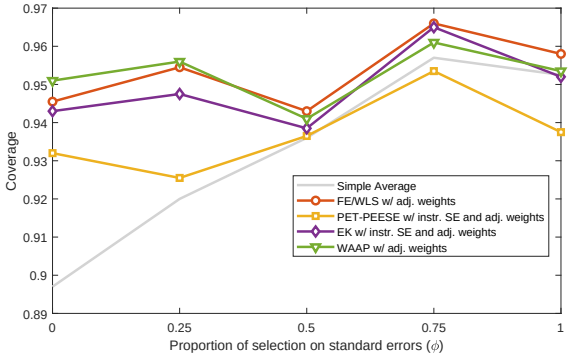
(d) Adjusting EK



(e) Adjusting WAAP



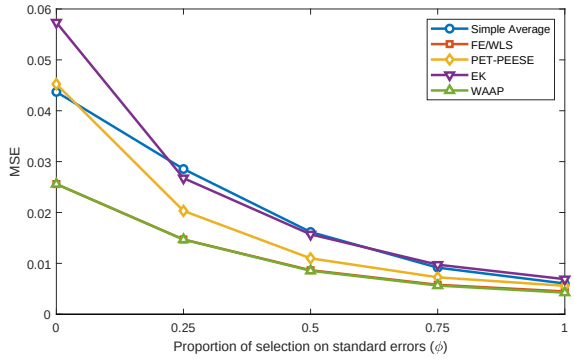
(f) Comparison of adjusted estimators



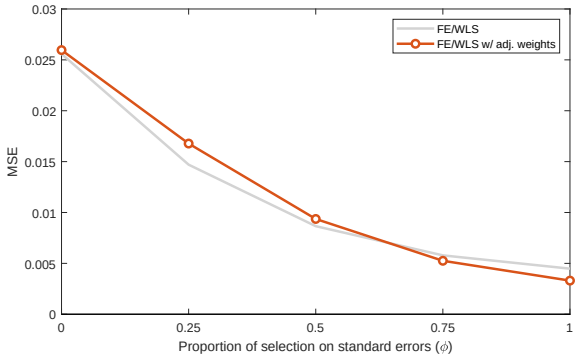
B Additional Simulation Results: Mean Squared Error (for Online Publication)

Figure 13: MSE: cheating selection, no effect ($\alpha_1 = 0$)

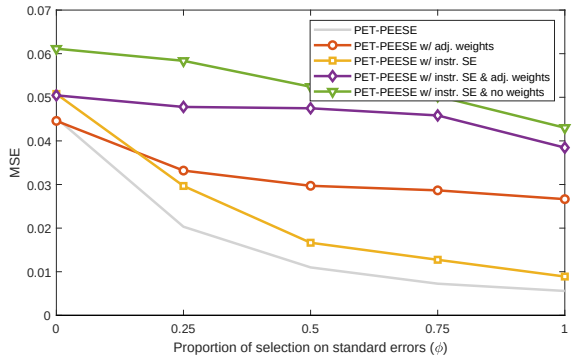
(a) Unadjusted estimators



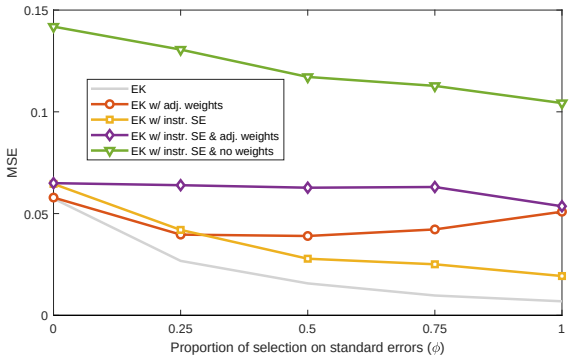
(b) Adjusting FE/WLS



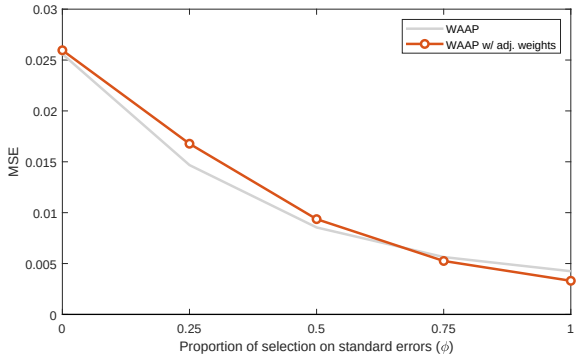
(c) Adjusting PET-PEESE



(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators

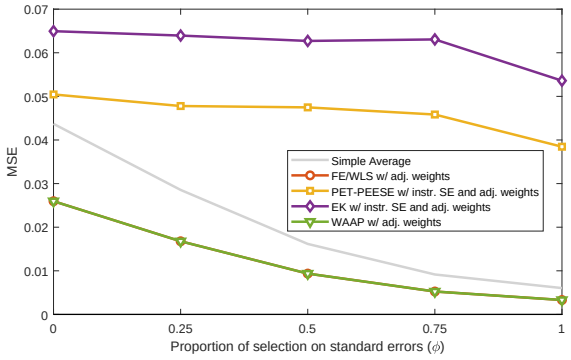
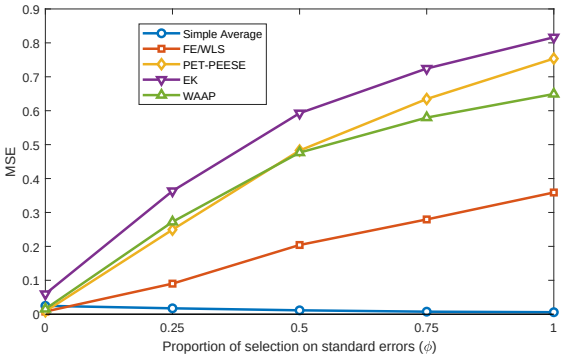
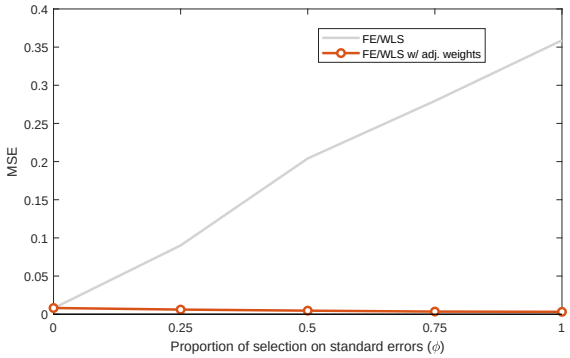


Figure 14: MSE: cheating selection, moderate effect ($\alpha_1 = 1$)

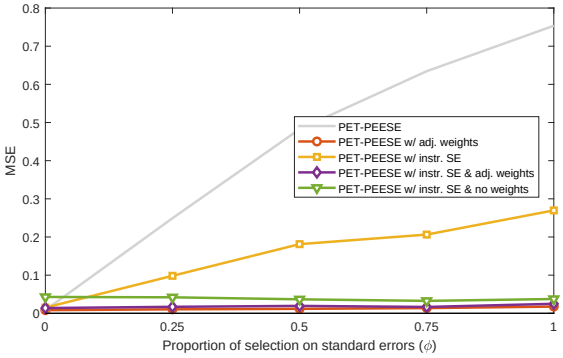
(a) Unadjusted estimators



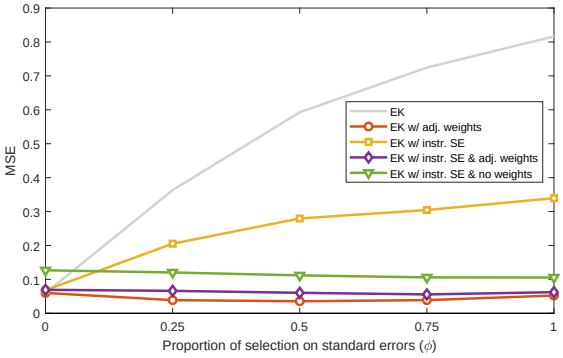
(b) Adjusting FE/WLS



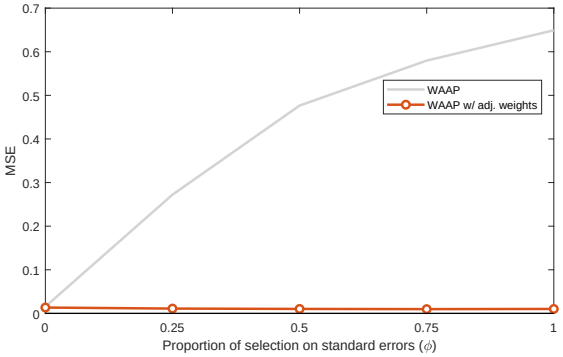
(c) Adjusting PET-PEESE



(d) Adjusting EK



(e) Adjusting WAAP



(f) Comparison of adjusted estimators

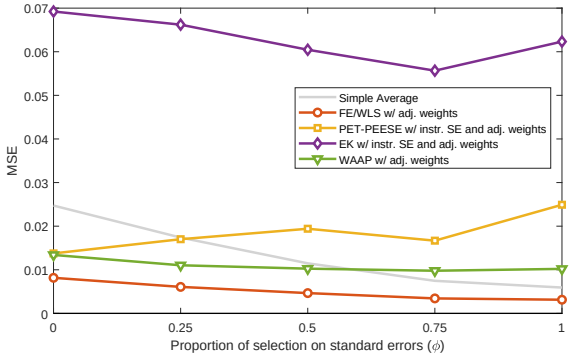
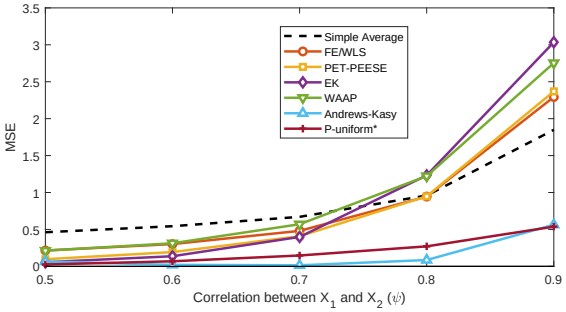
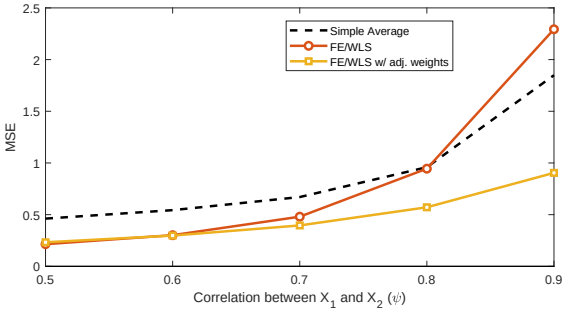


Figure 15: MSE: p-hacking selection, no effect ($\alpha_1 = 0$), various values of ψ

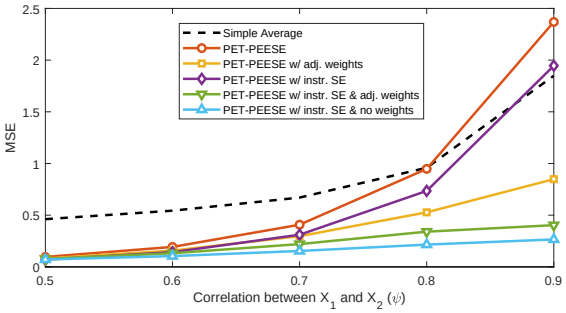
(a) Unadjusted estimators



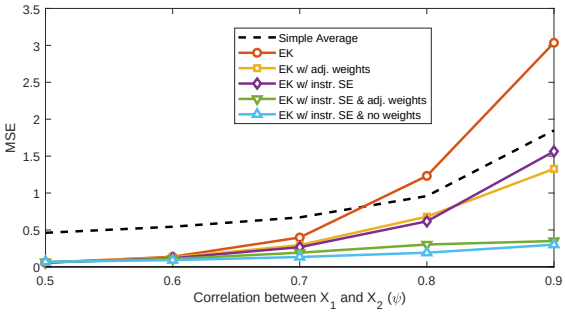
(b) Adjusting FE/WLS



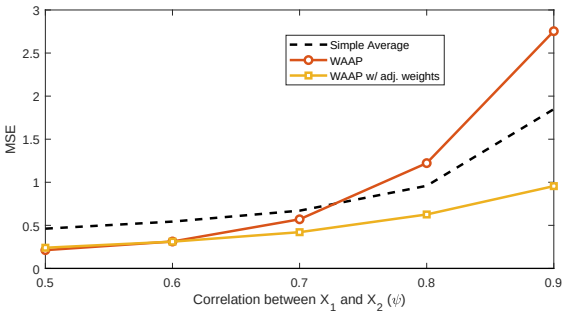
(c) Adjusting PET-PEESE



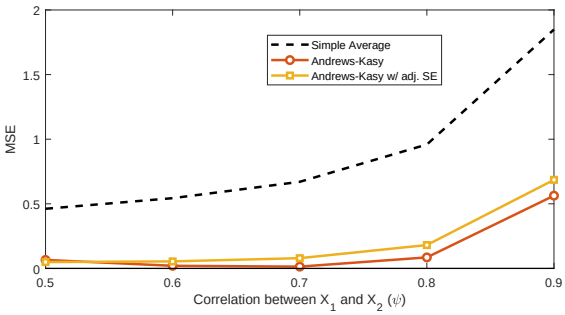
(d) Adjusting EK



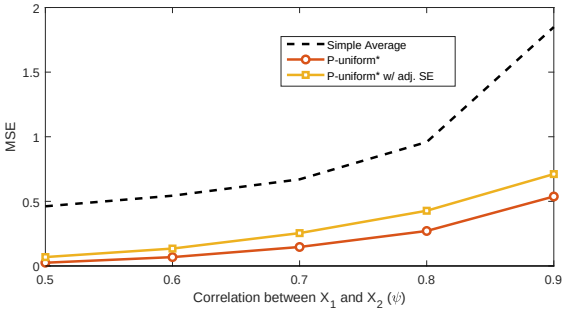
(e) Adjusting WAAP



(f) Adjusting Andrews-Kasy



(g) Adjusting p-uniform*



(h) Comparison of adjusted estimators

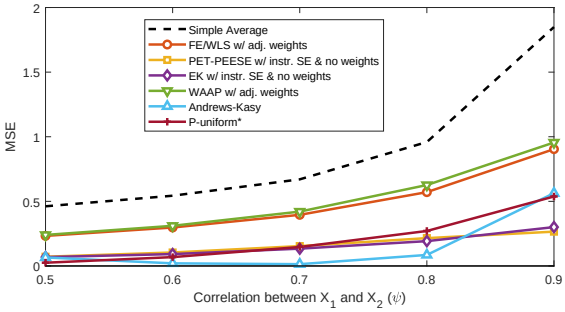
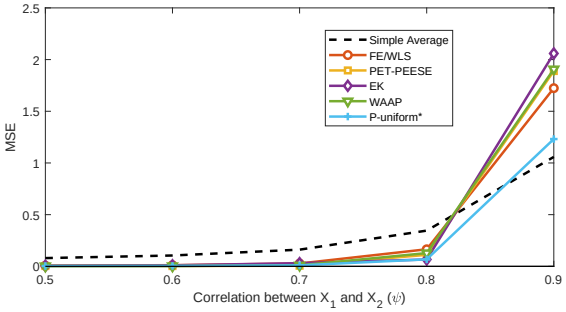
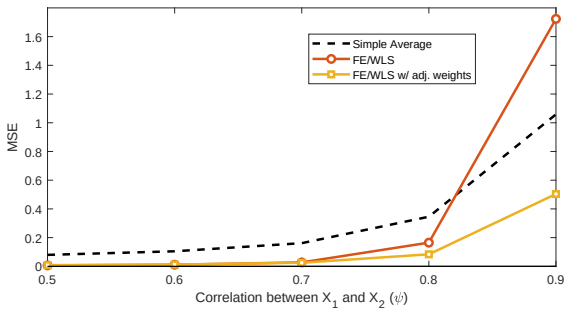


Figure 16: MSE: p-hacking selection, positive effect ($\alpha_1 = 1$), various values of ψ

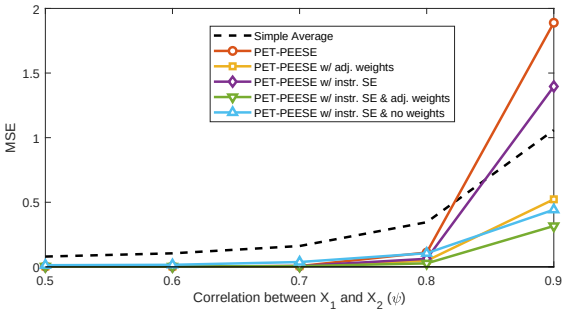
(a) Unadjusted estimators



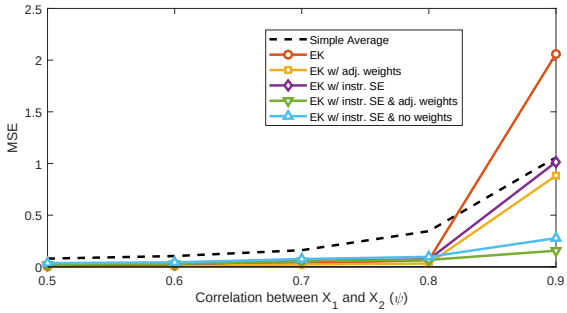
(b) Adjusting FE/WLS



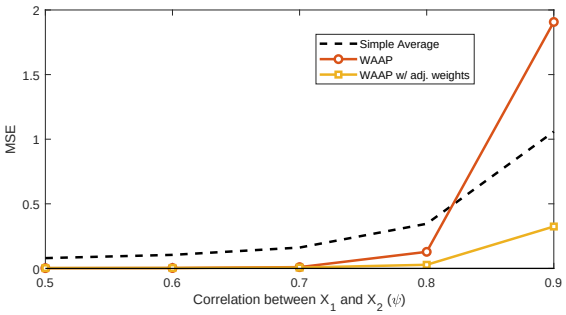
(c) Adjusting PET-PEESE



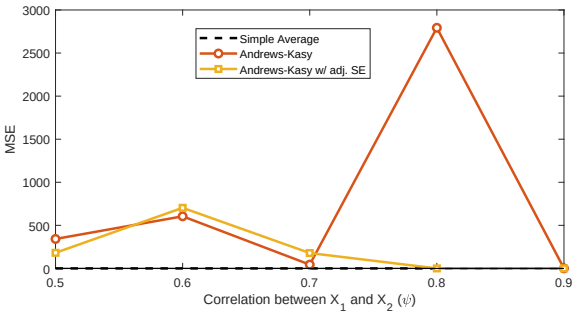
(d) Adjusting EK



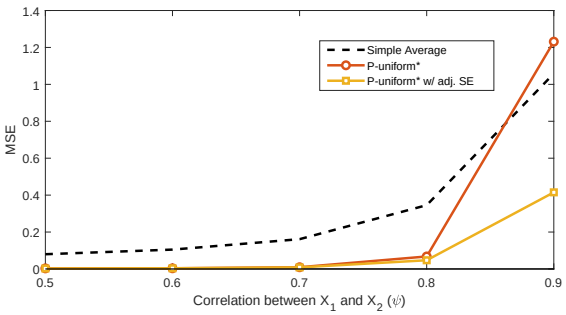
(e) Adjusting WAAP



(f) Adjusting Andrews-Kasy



(g) Adjusting p-uniform*



(h) Comparison of adjusted estimators