# Omitted variable bias in interacted models: A cautionary tale

Benjamin Feigenberg, Ben Ost, and Javaeria A. Qureshi [a]

We highlight that analyses using interaction terms to study treatment effect heterogeneity are susceptible to a form of omitted variable bias that is often overlooked in economics. Unlike most instances of omitted variable bias, the omitted variables in this case are available to the researcher but were not included in the model. We demonstrate that this exclusion matters based on a replication of 200 estimates across 17 papers published in the American Economic Review over the past 5 years. For approximately half of the results, failing to account for the omitted variables changes the estimate by more than 100%.

[a] Department of Economics, University of Illinois at Chicago, 601 South Morgan Chicago, IL 60607, United States. Feigenberg: bfeigenb@uic.edu, Ost: bost@uic.edu, Qureshi: javaeria@uic.edu

## I. Introduction

Though economists vigilantly worry about omitted variable bias (OVB), we highlight that a subtle type of omitted variable bias in interacted models is systematically overlooked. We suspect that the issue has gone unnoticed because it is tempting to infer that if a treatment variable, T, is exogenous conditional on a vector of controls, X, then the interaction between treatment and a covariate, H, is exogenous conditional on X and H. This erroneous logic may lead to situations where researchers carefully consider omitted variables with respect to T but fail to consider omitted variables with respect to T*H.

Our study offers two contributions. First, we highlight and provide intuition for why the typical approach to employing interaction terms can lead to omitted variable bias in settings where treatment is only conditionally exogenous. Second, we perform a systematic literature review and replication exercise based on articles recently published in the *American Economic Review*. The literature review demonstrates that it is common to use interaction terms in contexts where treatment is only conditionally exogenous and that researchers rarely account for the potential for omitted variable bias when estimating treatment effect heterogeneity in this setting. The replication exercise shows that controlling for the omitted variables often overturns research conclusions.

To clarify the source of the bias we study, consider a context where treatment T is exogenous, but only conditional on vector X. To examine how the effect of T varies with a dimension of heterogeneity H, researchers often estimate a model that includes main effects for T and H along with their interaction and control for X. In this setting, the interaction X*H can be correlated with both the outcome and T*H, so failure to include it in the model can lead to

omitted variable bias. Fortunately, the solution is straightforward: controlling for X*H in the regression eliminates this potential source of bias.

For binary H, including X*H makes the interacted approach equivalent to estimating the effect of T across models run separately for H=0 versus H=1. As such, an alternative framing of the issue we highlight is that estimating heterogeneity based on typical interacted models can yield different conclusions from running separate models for H=0 versus H=1. We propose this as a litmus test for interacted models: to claim that the effect of T is different for H=1 versus H=0, it is important to show that the estimated effect of T is different when estimated for the H=1 sample and the H=0 sample.

To help build intuition, consider a study that examines the effect of having a female teacher (T) on student outcomes in a context where students are randomly assigned to teachers within schools. Suppose schools with more female teachers happen to also have better performing students overall and smaller gender gaps in outcomes (perhaps because these schools place greater emphasis on gender equity). When studying the effect of having a female teacher, it is important to account for the differences in average performance across schools through the inclusion of school fixed effects (X). Importantly, cross-school differences in gender gaps in outcomes do not bias estimates of the average effect of being assigned to a female teacher. Suppose instead that the question of interest is how female teachers affect girls' outcomes differently than boys' outcomes. In this case, the school fixed effects only account for the average difference across schools and do nothing to account for differential gender gaps across schools, which threaten estimate validity. To address this potential source of bias, one would need to include student gender*school fixed effects (i.e., X*H interactions, where H represents student gender).

In a mechanical sense, any regression model that uses interactions to study heterogeneity could have omitted variables of the type we highlight, but we find a sharp contrast in current practice depending on the underlying identification strategy. When the baseline identification strategy is difference-in-differences (DID) or regression discontinuity (RD), researchers tend to correctly specify the heterogeneity analysis.[1] In contrast, we find that in other specifications that rely on conditional exogeneity – henceforth referred to as "standard regressions" – 86% neglect to control for X*H.[2] In our literature review, "standard regressions" make up more than 40% of all reduced form papers and are more than twice as common as DID papers.

Though applied researchers often overlook the source of omitted variable bias we highlight, analytically it is no different from more commonly recognized sources of omitted variable bias. Consequently, we see our paper as contributing to applied fields as opposed to econometric theory. In this regard, we view our study as similar in spirit to Bertrand, Duflo, and Mullainaithan (2004). In that paper, the authors argue that a "well understood" econometric problem (serial correlation) has been "largely ignored by researchers using [difference-in-differences] estimation" and has the potential to significantly alter research conclusions. Similarly, there is an existing literature studying biases arising from non-additive models (e.g. Gibbons, Suárez Serrato, and Urbancic, 2019; Balli and Sørensen, 2013; Giesselmann and Schmidt-Catran, 2018) but the OVB we highlight continues to be overlooked in empirical work.

[1] In traditional DID analysis, the X vector includes the "ever treated" and "post" controls and the interaction term "ever treated*post" is the "T" variable. Correct heterogeneity analysis according to H would add H*T, H*"ever treated" and H*post. Incorrect heterogeneity analysis would add only H*T. None of the studies we review use DIDID or RK designs.

[2] We can think of no theoretical justification for this difference. We speculate that one potential factor might be that many researchers (including ourselves) often think of RD/DID designs in a different, more visual way. For DID, another possible explanation is that heterogeneity analysis uses a specification equivalent to a triple difference design and omitting X*H controls would be equivalent to failing to control for the complete set of two-way interactions. Researchers may be less likely to omit these two-way interactions given that the correct specification for triple difference designs is well known.

Our study also relates to the many recent papers in the literature on treatment effect heterogeneity more broadly, much of it relating to the implications in difference-in-differences models (Borusyak and Jaravel, 2017; Callaway and Sant'Anna, 2020; Goldsmith-Pinkham, Hull, and Kolesar, 2021; Goodman-Bacon, 2021; Muralidharan, Romero and Wüthrich, 2021; Sloczynski, 2020; Sun and Abraham, 2021).[3] A key theme from this literature is that unmodelled treatment effect heterogeneity has the potential to bias estimated average treatment effects. More generally, any form of unmodelled non-linearity in the controls has the potential to introduce bias. In our replications, we do not attempt to address all of these potential issues, nor do we probe the validity of other assumptions implicit in the authors' original models. Instead, as in other papers that illustrate issues in empirical practice, we take the author's preferred model as given and just assess how the estimates of that model are affected by the inclusion of the X*H variables that are our focus. Accounting for the full scope of heterogeneity may be further consequential, but the purpose of our article is to make a simple point that has the potential to impact empirical practice and research conclusions.

To the best of our knowledge, ours is the first paper to document the prevalence of this misspecification in the literature and to systematically assess its empirical relevance. We use 3 complementary metrics to assess how the estimates we replicate are affected by controlling for X*H. First, we use seemingly unrelated regression to test whether controlling for X*H leads to a statistically significant change in the estimated effect of T*H. We find that the change in the T*H coefficient is statistically significant at the 5% level in 29% of cases – far more than would

[3] A separate concern, raised in Athey and Imbens (2017) in the context of experimental analyses, is that controlling additively for binary covariates that partition the population without including interactions between these covariates and an indicator for treatment will introduce finite sample bias in the estimation of average treatment effects.

be expected by chance suggesting that the coefficient changes we identify are unlikely to be explained by statistical noise alone.

While our first approach confirms the presence of bias, the statistical test results could be driven by small but precisely estimated changes in coefficient magnitudes that do not alter qualitative conclusions. Alternatively, even in cases where we fail to reject that coefficient estimates are unchanged, it is possible that sizable changes in magnitudes and/or significance levels occur which meaningfully affect the interpretation of findings. For instance, if the inclusion of X*H controls changes an estimate from 0.2 (0.08) to 0.05 (0.1), one would be hesitant to conclude that the findings are indicative of a robust, 0.2 treatment effect (even if the two estimates are not statistically distinguishable from each other). As such, our second approach describes the magnitude of coefficient changes from controlling for X*H.

Strikingly, in roughly half of the cases, failing to account for the omitted variables changes the estimate by more than 100%. The changes in coefficients, which we term the "estimated bias," are not clustered in a small number of papers: 81% of papers we replicate have at least one estimate where the bias is larger than our preferred estimate, and 56% of papers have at least one estimate where the bias is three times larger than our preferred estimate.[4]

Our third approach examines the sensitivity of research conclusions by characterizing changes in sign and significance levels. Though this approach is more coarse than considering the magnitude, statistical significance and sign are particularly relevant for thinking about the ramifications for how research informs policy. We find that research conclusions are often sensitive to controlling for X*H. Among estimates that were statistically significant in the

[4] For convenience, we occasionally use the term "bias" rather than "estimated bias" to refer to the difference in estimates derived from a model that omits X*H versus a model that controls for these interactions. Naturally, there is uncertainty regarding the true bias, both because of sampling variation in the preferred model and because there may be other biases that are not accounted for in either the original or our preferred model.

original paper, accounting for the omitted variables changes the sign or significance of 81.2% of the estimates. This lack of robustness to controlling for X*H partly reflects that standard errors tend to rise when controlling for X*H, but we provide several pieces of evidence suggesting that the increase in standard errors does not drive all of the sensitivity.

Taken together, our findings highlight how a simple source of bias that occurs regularly in economics research has the potential to influence research conclusions. Fortunately, the omitted variables that generate this bias are available to the researcher.[5]

## II. Omitted variable bias in interacted models

To describe the omitted variable bias we study, consider the setting where treatment T is exogenous conditional on a set of controls X. The researcher is interested in how the effect of T varies with a covariate H and estimates what we term the naïve interacted regression:

$$\mathrm{Y} = \beta T + \delta \mathrm{H} + \lambda \mathrm{TH} + \gamma \mathrm{X} + \epsilon \qquad (1)$$

Though Equation (1) estimates treatment effect heterogeneity according to H, it omits X*H interactions (as we show is typical in the literature). The key question is whether X*H belongs in the model (i.e., whether the effect of X on Y varies according to H).[6] Whether X*H belongs in

[5] Adding X*H controls may come at a cost with respect to statistical power. Researchers generally err on the side of controlling for potential confounders even when doing so may reduce power and we believe that this approach is just as appropriate with omitted interactions as it is with omitted main effects. None of the articles we reviewed argued that X*H controls should be omitted based on power considerations. We provide additional discussion in the "Power considerations and conclusion" section.

[6] If X*H were uncorrelated with T*H (after removing the linear effects of X, H and T), this would also provide justification for excluding X*H from the model. It is straightforward for researchers to directly assess whether this condition holds in their specific study context. For our purposes, it is sufficient to provide counter examples to show that it does not generally hold. Define $\widetilde{XH}$ and $\widetilde{TH}$ to be the residuals from regressing X*H and T*H on X, H, and T. It is possible for $\widetilde{XH}$ to be uncorrelated with $\widetilde{TH}$, but only if X, H and T are all uncorrelated with each other or if various complex products of correlations cancel. For example, in a simulation, we find that for normal random variables that are linearly related, if corr(X,T)=0.5, the correlation between $\widetilde{XH}$ and $\widetilde{TH}$ is only less than 0.1 in absolute value when corr(X,H) and corr(H,T) are approximately equal in magnitude and opposite signed. Interestingly, there are also some special cases where $\widetilde{XH}$ and $\widetilde{TH}$ *cannot* be uncorrelated. In a derivation available upon request, we show that if H is strictly exogenous, but X and T are correlated (as is the case if T is exogenous only conditional on X), $\widetilde{XH}$ is necessarily correlated with $\widetilde{TH}$.

the model is necessarily context specific and likely needs to be assessed empirically as we suspect that there are few cases where the researcher can confidently exclude it based on theoretical considerations.

When H is binary, it is informative to compare the naïve interacted model (equation 1) to the split-sample approach, whereby, instead of using interaction terms, the researcher could estimate the following two regressions and compute the difference ($\beta_1 - \beta_2$).

$$Y = \beta_1 T + \gamma_1 X + \epsilon_1 \ \ if\ H = 1 \qquad (2)$$

$$Y = \beta_2 T + \gamma_2 X + \epsilon_2 \ \ if\ H = 0 \qquad (3)$$

We view the split-sample approach to be a useful benchmark as it literally shows the difference in the effect of T for H=1 versus H=0 samples. Adding X*H to the naïve interacted model makes it equivalent to the split-sample approach since it relaxes the equation (1) constraint that $\gamma_1 = \gamma_2$. When H is continuous, the split-sample approach is not possible, but analogous reasoning implies that X*H interactions should be included.[7] As such, even in cases where H is continuous, we use the term "naïve model" to refer to equation (1) and the terms "split-sample equivalent" or "preferred" to refer to a model that adds the vector X*H to equation (1).

It is rarely possible to estimate fully flexible, fully interacted models and so functional form assumptions are generally necessary. In theory, one could assess the consequences from relaxing several other parametric assumptions imposed by the "naïve model" in equation (1) – for instance, allowing the individual covariates within the vector X to be interacted with each other - so why do we focus on whether $\gamma_1 = \gamma_2$? Our focus on this particular assumption is motivated by its importance for understanding how the interacted model compares to split sample models. When interpreting the results of interacted models with binary H, it is common

[7] For continuous H, even the split-sample equivalent model is susceptible to omitted variable bias if there are non-linear interactions between H and the covariates.

to present $\beta$ as the effect of T for H=0 and $\beta + \lambda$ as the effect of T for H=1. This interpretation is analogous to how one would interpret the results of separate estimates for the H=0 and H=1 samples, but the estimates are identical only when X*H is included. Considering how the naive model is affected by the inclusion of X*H is thus precisely the check necessary to evaluate whether the interacted model yields results that are substantially different from those produced by models estimated separately in each subsample.

When the split-sample model and the naïve interacted model yield different results, which estimates should be preferred? There is no definitive answer to this question as researchers must consider bias reduction versus potential efficiency loss. That said, the split-sample approach is typically simpler and the fact that it imposes fewer parametric assumptions make it a natural starting point. This is particularly clear when $\beta_1$or $\beta_2$ are of direct interest. For example, suppose a researcher is interested in the effect of a treatment for boys and has a sample including boys and girls. One possible approach is to estimate the effect of the treatment for the sample of boys and simply exclude girls from the regression. A second possible approach is to pool the boys and girls and estimate the effect of T and T*girls, where the coefficient on the main effect is the estimate for boys. If the two approaches yield different findings, the researcher would need to provide a clear justification for why the more complex interacted approach is preferred. When the question of interest is the gap between boys and girls, the reasoning is analogous and the split-sample model should be taken as the benchmark, with justification required to instead adopt the more parametric, interacted model. This argument is key to interpreting divergent results from the split-sample and naïve interacted approaches. Revisiting the example presented earlier, suppose that the split sample-equivalent estimate is 0.05 (0.1), the naive interacted approach yields an estimate of 0.2 (0.08), and the two estimates are not statistically distinguishable. A

researcher faced with this set of estimates may argue for using the naïve interacted model, but the burden of proof lies on that researcher to show that the required additional assumption that underpins the interacted model is reasonable in their context. Our literature review suggests that the status quo in economics is to present the estimate from the naïve interacted model with no discussion of the assumption behind this estimate and no consideration of the split-sample equivalent estimate.

## III. Illustrative example

For concreteness, we illustrate our point in the context of Gong, Lu and Song (2018), which was recently published in *Journal of Labor Economics*. This example is used because it is intuitive and has a clear justification for why treatment is as good as randomly assigned, but only conditional on a set of controls. It is important to emphasize that we did not select this example randomly and the select results we present are not representative of those in the paper more broadly. As such, we are not aiming to generalize from this example in any way and we provide a systematic assessment of the empirical importance of the issue in the following section.

In Table 3 of Gong, Lu and Song (2018), the authors investigate whether teacher gender has a differential effect on boys' versus girls' test scores and self-assessed performance. They estimate these effects in a setting where teachers are randomly assigned to students within school-grade-subjects so that the conditional independence assumption is plausibly satisfied within these blocks. The main estimating equation is

$$Y_i = \beta T_i + \delta H_i + \lambda T_i H_i + \gamma X_i + \epsilon_i \quad (4)$$

where $Y_i$ is test score (either Chinese, Math or English), $T_i$ represents teacher gender, $H_i$ is student gender and $X_i$ is a vector of controls that includes student covariates, subject fixed effects and school-grade fixed effects.

Gong, Lu and Song (2018) carefully assess the possibility that omitted variables are correlated with assignment to a female teacher ($T_i$), but once they establish that $T_i$ is plausibly random controlling for $X_i$, they do not further assess the exogeneity of the interaction term, T*H. The key question for our context is whether X*H belongs in the model, i.e. do the effects of the controls vary by student gender?[8] Though there are potential reasons that the effect of several of the X controls could vary by student gender, it seems particularly likely that the coefficients on the subject fixed effects will vary by student gender since girls typically outperform boys in language skills but usually perform worse in math. In this case, a model that does not allow the subject fixed effects to vary by student gender would yield biased estimates.

To illustrate how this bias manifests in practice, in Panel A of Table 1 we replicate the Gong, Lu and Song (2018) analysis of test scores and self-assessed subject performance (Table 3 from their paper). Columns 1 and 2 show the original results that exclude X*H interactions. For both test scores and self-assessed performance, the results suggest that boys perform worse when assigned to a female teacher and girls perform substantially better when assigned to a female teacher. For self-assessment in particular, the negative effect for boys is quite large at -0.125 standard deviations and the interaction estimate of 0.289 suggests that girls benefit by 0.164 standard deviations from assignment to a female teacher. When we estimate separate models for boys and girls, however, the magnitudes are quite different. For boys, female teachers only decrease self-assessed performance by an insignificant 0.018 and for girls, female teachers

[8] Gong, Lu, and Song (2018) fits the special case mentioned near the end of footnote 6 where H*T is necessarily correlated with X*H because T and X are correlated and H is approximately uncorrelated with both T and X.

increase self-assessed performance by only 0.056. Thus, the difference in the estimated effect of a female teacher is only 0.056 – (-0.018) = 0.074 in contrast to the original estimate of 0.289. Columns 3 and 4 of Panel A, Table 1 show the split-sample equivalent estimates that control for X*H interactions. These models mechanically match the separate estimates by subgroup. Though the qualitative story is broadly consistent with the original findings, the magnitudes are quite different when X*H is included in the model versus not, and these differences are statistically significant.

In Panel B of Table 1 we replicate a variety of non-cognitive outcomes from Gong, Lu, and Song (Table 4 in their paper). The first 4 columns show the original estimates and the last 4 columns show the split-sample equivalent versions. We reject the null hypothesis that coefficients are unchanged by the inclusion of X*H controls for all but the "pessimistic" outcome. Moreover, the non-cognitive estimates appear to be quite sensitive with some estimates even changing sign. Naturally, the sensitivity of the estimates is context-specific, so it is important not to generalize from this assessment. That said, the Gong, Lu, and Song (2018) example establishes that it is possible for the omitted variable bias to be empirically important and motivates the replication exercise to which we now turn.

## IV. Use of interaction terms in recent AER articles

We conduct a literature review and replication exercise to answer the following three questions:

1. *How often do researchers use interaction terms to study treatment effect heterogeneity in a context where treatment is only conditionally exogenous? In other words, how common is the context where the misspecification we highlight might apply?*

2. *In papers that fit our context, how often are the X*H interactions omitted?*
3. *In papers where the interacted model omits X*H interactions, how often is the omitted variable bias empirically important?*

We answer these questions by performing two complementary searches of articles in the *American Economic Review*. Our primary search is a keyword search on 5 years (2015-2019) of AER publications with the goal of identifying articles that fit our context and have publicly available data that permits replication.[9] This keyword search identifies a set of 21 replicable articles. We complement the keyword search by performing a manual review of articles published during the first year of our sample period (2015).[10]

For the keyword search, we start by searching for articles published in AER from 2015-2019 that include the phrase "interaction term" along with some related phrases (details on the exact search are in Appendix B). Of the 145 articles that match our search terms, 62 have publicly available data and do files. Our keyword search yields many false positives because researchers often use our search terms in other contexts (e.g., theory). To eliminate false positives, we read the 62 articles that match our search terms and have replication files available. Of the 62 articles, 21 use interaction terms to conduct heterogeneity analysis and exploit variation in a treatment measure that is only conditionally exogenous.[11]

The keyword search is unlikely to identify the universe of relevant articles so it cannot be used to evaluate the prevalence of our context (question 1). Relatedly, if articles that correctly include X*H are less likely to use the phrase "interaction term," judging question (2) solely

[9] Though papers with publicly available data are not likely to be a random subset, there is no reason to expect that the bias in these studies would be unusual.
[10] We focus on one year because it is time-consuming to manually examine every article and our goal is only to provide a general sense of prevalence.
[11] Appendix B provides a detailed breakdown of the false positives.

based on the keyword search would overstate the fraction of articles that omit X*H.[12] The limitations of the keyword search motivate our manual search, where we reviewed all 112 articles published in 2015 to identify the field and approach associated with each article. The manual review is useful for answering questions (1) and (2) because it does not rely on keyword matches, and it provides a more nuanced characterization of the prevalence of our context by describing the universe of papers published in 2015. In practice, our conclusions regarding question (2) are similar whether we use the manual search or the keyword search.

## V. Question 1: How common is our context?

We assess the frequency of our context using the manual review of 2015, which begins with the universe of 112 full-length articles that were not comments or replies. Of the 112 articles published in 2015, 37 are pure theory papers and 3 are purely descriptive. Our concerns are also not directly relevant for the 31 papers that use structural or time-series approaches. Of the 41 papers that use reduced-form style regression models, 17 use standard regression analyses of the form described in equation (1) where there is a treatment variable T, that is argued to be exogenous conditional on a set of covariates X.[13] An additional 10 articles use DID or RD designs and the remaining 14 are lab or field experiments where treatment is typically unconditionally exogenous.

Of the 10 articles that use DID or RD, 5 study treatment effect heterogeneity using interactions and one of the RD articles shows specifications using separate subsamples but does not conduct statistical testing across subsamples. We count the 5 articles that statistically test for

[12] This is plausible because articles that present separate analyses by subsample may use interaction terms to statistically test for differences across subsamples without ever using a phrase like "interaction term." Such analyses correctly account for X*H and would be missed by our keyword search.

[13] Categorizing "standard regression" designs is somewhat subjective as some models are in the spirit of DID, but use a continuous treatment or deviate slightly from DID in other ways. We follow the original authors' presentation to classify papers. Though separately analyzing "standard regression" and DID/RD designs reveals an interesting difference in current practice, the distinction is not vital to the main conclusions of our study.

heterogeneity as fitting our context. Of the 17 articles that use standard regression, 14 study heterogeneity using interactions and none of the papers show split sample analyses. Putting this all together, we identify 14 articles that use interaction terms to study heterogeneity in a standard regression context and an additional 5 articles that use interaction terms to study heterogeneity in a DID or RD context.

The 14 articles that use standard regressions make up 13% (14/112) of all articles published, 19% (14/72) of articles excluding pure theory and descriptive papers, and 34% (14/41) of reduced-form style papers. Adding in the DID and RD papers, 46% (19/41) of the reduced form papers are from contexts where treatment is only conditionally random and treatment heterogeneity is tested using interaction terms. As a point of comparison, the number of articles that use interactions in a standard regression where treatment is only conditionally exogenous is 14, which is the same as the total number of randomized experiments and larger than the total number of articles using DID or RD designs.

## VI. Question 2: In papers that fit our context, how often are X*H interactions omitted?

The manual search identifies 14 articles that use standard regression and 5 papers that use DID or RD to investigate treatment effect heterogeneity. Of the 14 standard regression articles, 2 articles control for X*H interactions and 12 do not. In 4 of the 5 papers that use DID or RD designs to study treatment effect heterogeneity, the authors control for X*H interactions. Given that analytically there is no difference between the standard regression and DID/RD contexts, it is striking that 80% of the RD/DID papers include the necessary interactions but only 14% of the standard heterogeneity analysis papers include these interactions.

## VII. Question 3: How large is the bias?

To evaluate the magnitude of the omitted variable bias, we require a set of papers that neglected to control for X*H interactions and have publicly available data and replication files. Of the 21 articles identified from the keyword search, one is a special case where the vector of covariates X subsumes the relevant H variable.[14] An additional 2 articles control for X*H interactions. These articles thus have no interacted omitted variable bias to contend with. For the remaining 18 articles, we replicate the interaction term estimates and assess the extent of bias from omitting the X*H interactions. In one article that uses interaction terms in an IV analysis, adding the X*H controls makes it so the instrument is no longer relevant in the first stage. We exclude this paper from our description of coefficient bias since without a first stage, it makes little sense to examine the IV estimates. With this exclusion, we have a total of 17 articles that we use to characterize the magnitude of the omitted variable bias.

We focus on interaction terms that are intended to characterize treatment effect heterogeneity as opposed to interactions that control for potential confounders. We generally replicate every estimate in a paper that fits this criteria, but we make some exclusions to avoid double counting or including very tangential analyses. We exclude estimates found only in an appendix and we exclude specifications that are very similar to already replicated specifications. For example, if a table showing coefficient stability includes multiple estimates of essentially the same parameter but with varying controls, we only include the author's preferred specification for replication purposes. Relatedly, if an author shows robustness to various ways of measuring the same outcome (e.g. binary or continuous), we include just the author's preferred metric. In

[14] In general, if X subsumes H then X*H is equivalent to X. As an example, if a study is interested in whether class size has a different effect in rural vs urban areas, school fixed effects are the same as school-by-urbanicity fixed effects. In a split-sample model, each school appears in either the rural or urban sample.

two cases, the original paper used a non-linear model (e.g. probit) for some specifications, and we replicate these specifications using a linear model and assess the degree of bias from omitting X*H in the linear model.[15]

Other than the aforementioned exceptions, we include all interaction term estimates from the 17 papers in our sample, resulting in a total of 200 estimates. Table A1 lists the 17 included papers and identifies the number of estimates included from each paper. As can be seen from Table A1, the number of estimates varies substantially by paper (from 2 to 50). Estimate sensitivity is unlikely to be independent for estimates that are derived from the same paper, and it would be concerning if the paper with 50 estimates happened to be a context where OVB was severe and therefore skews the results. As such, we present both results where each estimate counts equally and results where each paper counts equally.

We employ three complementary approaches to characterize the degree of bias associated with the 200 estimates in our sample. First, we statistically test whether the naïve estimated effect of T*H differs from the split-sample equivalent estimate. Second, we describe the distribution of the magnitude of the estimated bias for the 200 estimates. Finally, we assess whether the qualitative conclusions from the original paper would have been different had they used a split-sample equivalent model. Because each approach has differing strengths and weaknesses, the findings from the 3 approaches are best taken as a whole.

[15] We estimate linear models because the vast majority of estimates we replicate are derived from linear models and the marginal effects from the linear models are generally very similar to the marginal effects from the non-linear models. For each estimate that was originally based on a non-linear model, we verify that the linear marginal effect is within 10% of the marginal effect from the non-linear model.

*VII.A First approach: Statistically testing the naïve versus split-sample equivalent estimates*

We use seemingly unrelated regression to test whether the estimated effect of T*H in each naïve model is statistically different from the estimated effect in the corresponding split-sample equivalent model. We are only able to estimate the seemingly unrelated regression for 182 of the 200 estimates, because the joint covariance-variance matrix of the remaining 18 specifications is too sparse.[16] For the 182 estimates that we test, 28.6% of the split-sample estimates are statistically distinguishable (at the 5% level) from the naive estimates – far more than would be expected by chance. These cases are distributed fairly evenly across the 17 papers as 81% of the papers have at least one case where the split-sample equivalent estimate statistically differs from the naïve estimate. If we restrict attention to the estimates that were statistically significant in the original paper, we find that 40% are statistically distinguishable from the naïve estimates.[17] Importantly, our goal is to assess whether omitting X*H *can* be important, not to demonstrate that it is always important. The rate of rejection here far exceeds this proof of concept benchmark.

Though statistically testing the change in the coefficient is a useful starting point, we would not want to judge the importance of omitting X*H based on this criterion alone because it provides no information on the magnitude of the omitted variable bias. Depending on the level of power, it may be possible to statistically distinguish between estimates that are not meaningfully different, or there may be a large change in the estimate that is not statistically significant. Judging bias solely based on the statistical test would lead to the spurious conclusion that omitted variable bias tends to be unimportant in contexts with less power.

[16] In these 18 cases, the naïve and split-sample equivalent models can both be estimated with standard errors, but the data is too sparse to estimate the covariance between the two models.
[17] The higher rejection rate reflects the fact that estimates that are more precise in the original paper are more likely to be statistically different from the split-sample equivalent estimate because of power.

The analysis above establishes that there are many cases where the omitted variable bias meets the 95% statistical threshold, but it is possible for cases that do not meet this threshold to still be concerning. Put differently, failure to reject the null of zero bias does not imply zero bias and researchers are likely to be wary of estimates where there is a high possibility of bias, even if the likelihood of that bias is less than 95%. As such, for the cases where the researcher fails to reject the hypothesis that the split-sample equivalent estimate is identical to the naïve estimate, it does not follow that the naïve estimate should be treated as the preferred estimate or taken as unbiased.

No single metric perfectly captures whether result sensitivity is concerning, and assessing this requires a degree of judgement based on a combination of coefficient movement, statistical significance, standard error movement, ex-ante theory and other context-specific considerations. For example, if the split-sample and naïve estimates are quite different from each other, this suggests that the original results are not robust to controlling for X*H, but if the standard error on the split-sample equivalent estimate is sufficiently large, the naïve estimate might still be preferrable and the lack of robustness discounted. That said, a large proportional increase in the standard error does not have to lead to an uninformative estimate. For example, the largest change in standard errors we observe in the replications is a situation where the naïve estimate is 0.0026 (0.0012) and the split-sample equivalent estimate is -0.036 (0.0071).

Ultimately, for a researcher trying to bolster the credibility of their findings, showing that one estimate is statistically indistinguishable from another estimate based on an alternative model is less valuable than showing that the estimated coefficients are similar across the two models. As such, in the following section, we consider how the estimated effect of T*H differs across the naïve and split-sample equivalent models.

*VII.B Second approach: Describing the magnitude of the estimated bias*

A key challenge in describing the magnitude of the estimated bias is that the coefficient and therefore, the bias, is measured in different units in different contexts making it hard to summarize. The best way to judge the severity of bias is on an individual basis with a full understanding of the context and the units of T, H and Y.[18] However, this sort of case-by-case analysis does not lend itself to summary, so instead we present two measures of bias that are more comparable across all 200 estimates: bias in percentage terms and bias in standardized units.

First, we calculate the percent change from the split-sample estimate to the naïve estimate.[19] In panel A of Figure 1, we present a histogram of the bias in percentage terms across the 200 estimates, capping the bias at +/- 200% to prevent the scale from being distorted by particularly large biases. Applied researchers are often more concerned with biases that move estimates away from zero or alter the sign of estimates, so we highlight these cases. The red shaded region shows the 104 instances (52% of the total) where the naïve estimate is larger than the split-sample equivalent estimate and of the same sign. The green shaded region shows the 46 instances (23% of the total) where the bias is more extreme than -100% and the sign therefore differs between the naïve and split-sample equivalent estimates. The blue shaded region shows the 50 instances (25% of the total) where the naïve estimate is smaller in absolute value than the split-sample equivalent.

[18] For example, if researchers are estimating a parameter that characterizes returns to scale, an estimate of 0.9 might have very different implications than an estimate of 1.1, but in other contexts, 0.9 and 1.1 estimates might have roughly the same implications.

[19] We construct percent changes relative to the split-sample benchmark because we prefer to define the bias in terms of the ratio $\frac{estimated\ bias}{preferred\ estimate}$ rather than defining it in terms of the ratio $\frac{estimated\ bias}{preferred\ estimate + estimated\ bias}$.

The key takeaway from panel A of Figure 1 is that relatively few estimates have small biases in percentage terms. Only 25 (12.5%) of the naïve estimates are within +/-20% of the split-sample estimate and only 58 (29%) of the naïve estimates are within +/-50% of the split-sample estimate. For 99 (49.5%) of the estimates, the bias is more than +/-100%. The bars corresponding to biases larger than +/-200% represent more than 30% of the sample. Though not shown in panel A, 21.5% of the estimates have biases larger than 300% of the split-sample equivalent estimates.[20]

One concern with reporting bias in percentage terms is that large percentage changes may be inconsequential for statistically insignificant estimates. In particular, if both the naïve and split-sample equivalent estimates are statistically insignificant, a large percentage change may not alter the qualitative interpretation of the result. To assess whether the large biases shown in panel A are driven by insignificant estimates, in panel B of Figure 1 we show the same histogram split according to whether the estimate from the original paper was statistically significant at the 5% level. This figure illustrates that the distribution of bias is similar for these two types of estimates, though the bias is somewhat more extreme for estimates that were originally significant. Of the 90 estimates that were originally statistically significant, 50 (55.6%) have biases larger than +/-100%.

Table 2 summarizes key points from the distributions shown in Figure 1 at the estimate-level and at the paper-level, allowing us to assess how the takeaways vary depending on whether each estimate is given equal weight or each paper is given equal weight.[21] Column (1) shows the

[20] If we focus on only cases where the naïve and split-sample equivalent are statistically distinguishable, the biases are more extreme as expected. For example, 81% have biases more than 100% and 58% have biases larger than 200%.

[21] The paper-level measures are constructed by first calculating paper-level averages across included estimates and then treating each paper as a single observation. To aggregate outcomes such as "bias exceeds 100%", we calculate the proportion of cases that exceed the threshold within each paper and then average these proportions across all papers.

distribution of bias for all estimates whereas column (2) shows the distribution of bias for the 90 naïve estimates that were statistically significant in the original paper. Columns (3) and (4) show paper-level versions of columns 1 and 2. In column 4, we include only the 90 estimates that were significant in the original papers and then collapse to the paper level. Column 4 has one fewer paper than column 3 because one paper had no statistically significant estimates. The paper-level results are broadly similar to the estimate-level results, but the estimated bias appears slightly less severe when assessed at the paper level. This suggests that estimates from papers that include more interactions (where interactions may be more central to the analysis) have slightly worse bias than estimates from papers that include fewer interactions. Panel B describes the degree of bias associated with the most sensitive estimate from each paper. All but one paper (94.1%) include at least one interaction where the bias is larger than the split-sample equivalent estimate. Focusing on just statistically significant estimates within each paper, 13 of the 16 papers (81.2%) have at least one instance of bias of this magnitude. Strikingly, the majority of papers have at least one statistically significant estimate for which the bias is more than 300% of the split-sample equivalent estimate.

As an alternative to describing the bias in percentage terms, we also examine the magnitude of the bias in standard deviation units.[22] Panel A of Figure 2 presents a scatterplot comparing measured bias to the corresponding split-sample equivalent coefficient estimates, both in standard deviation units. If omitted variable bias were generally unimportant, panel A of Figure 2 would show data points clustering near 0, but the scatterplot shows many instances of biases that are large in standard deviation units, as well as large relative to the split-sample

[22] Though more readily interpretable than unstandardized coefficients, standardized coefficients are still not likely to be comparable across very different contexts. For example, a 0.01 standard deviation effect size is a small change in many settings, but a very large change if the outcome of interest is the level of GDP and the standard deviation is constructed from a sample that includes both rich and poor countries.

equivalent estimates. The shaded bands in panel A of Figure 2 highlight the 29% of estimates for which the bias is smaller than 50% of the split-sample equivalent estimate. To facilitate examination of estimates near the origin, panel B of Figure 2 shows a magnified version of panel A, restricted to split-sample estimates that are smaller than 0.02 SD in magnitude. The basic pattern of findings is qualitatively similar.

Including additional controls has the potential to increase standard errors, and this raises two considerations. First, even if we have correctly rejected the null of no bias, it is possible that a researcher would prefer the biased naïve model if the split-sample model is sufficiently noisy to be uninformative. Second, if the split-sample equivalent model is noisy, it is possible that the difference between the split-sample equivalent and naïve model coefficients are driven by noisiness of that model rather than omitted variable bias. Even if the true bias were zero, one expects some estimated bias by chance and the magnitude of this spurious estimated bias is larger when standard errors are large.

To describe how standard errors differ across the split-sample equivalent and naïve models, Figure 3 presents a histogram of the ratio $\frac{naive\ SE}{ss-equiv\ SE}$. The peak of the distribution is slightly below 1, suggesting that the standard error is somewhat larger for the split-sample equivalent model in the modal case. Forty-five percent of cases have a ratio of 0.8 or larger and 16% of cases have a standard error ratio larger than 1. Nineteen percent of cases have a ratio smaller than 0.5, corresponding to a doubling or more of the standard error. As such, there is a non-trivial proportion of cases where standard errors rise sufficiently so that bias/variance tradeoffs may be relevant. In such cases, judging whether X*H should be omitted requires careful consideration of context-specific factors such as ex-ante theory and the magnitude of the

standard errors and point estimates. That said, there are also many cases where the change in standard error is sufficiently modest so that bias considerations are likely to dominate.

A key question for interpreting our results is the degree to which the large magnitudes of estimated bias reflect the larger standard errors in the split-sample equivalent model. It is worth noting that we can definitively rule out the possibility that the difference in coefficients across split-sample and naïve models solely reflect noise. The statistical tests in the previous section reject no bias for 29% of cases. Furthermore, the shape of the bias histogram shown in Figure 1 is inconsistent with a pure noise explanation because it is not peaked at zero.

Beyond simply establishing that noise cannot explain all of the estimated bias, it is useful to assess whether the cases with large estimated biases only occur when the increase in standard errors from including H*X is large.[23] Figure 4 shows the severity of the bias versus the ratio of the standard errors, where bias severity is measured as the fraction of cases where the naïve coefficient differs from the split-sample estimate by more than X%. The key takeaway from Figure 4 is that regardless of the standard error ratio, large estimated biases are common. For example, even for cases where the T*H standard error is mostly unchanged from the inclusion of H*X (standard error ratio=1), nearly 40% have estimated biases larger than 100%. Among the cases where the split-sample equivalent standard error is smaller (standard error ratio>1), 56% of models have estimated biases larger than the split-sample equivalent estimate. This is inconsistent with the notion that controlling for X*H blows up the standard errors and this alone is driving the large estimated biases.

[23] Of course, changes in standard errors are not independent of changes in coefficients so the overall shape of the relationship characterized by Figure 4 should not be over-interpreted.

*VII.C Third approach: Comparing statistical significance and sign*

For our final approach, we consider how the qualitative conclusions of each study would have differed had the authors originally controlled for X*H interactions. If the research conclusions change when estimating split-sample equivalent models, this has implications for future scholarly work and policy recommendations. Naturally, changes in statistical significance could be driven by random fluctuations so this approach does not test whether the naïve model is biased. Nevertheless, it is useful to know whether the researchers would have made the same policy recommendations had they estimated a split-sample equivalent model instead of the naïve model.

We categorize each coefficient coarsely into 3 categories (positive and significant at the 5% level, negative and significant at the 5% level, and insignificant at the 5% level) and compare the split-sample equivalent estimates to the naïve estimates that exclude X*H interactions. Panel A of Table 3 shows the results from this cross-tabulation. Forty-three of the original estimates were significant and negative, but only 14 of these estimates remain negative and significant after controlling for X*H. Forty-seven estimates were originally significant and positive and only three of these estimates remain significant and positive after controlling for X*H. Of the 110 original estimates that were insignificant, 10 are statistically significant after controlling for X*H.[24]

The large drop in statistical significance shown in Panel A could be driven by changes in coefficients, standard errors, or both. In particular, since standard errors tend to be larger in the split-sample equivalent model, estimates may become insignificant primarily because standard

[24] To assess whether the changes in significance are driven by substantively unimportant changes that happen to cross significance thresholds, in Appendix Figure A1, we plot the p-values from the split-sample model against the p-values form the naïve model. This figure shows no evidence that the result in Table 3 are driven by small changes in p-values that happen to cross the significance threshold.

errors typically rise. We show that standard errors do not drive the changes in statistical significance in Panel B, where we include the same point estimates for the coefficients as in Panel A, but we use the naïve (original) standard errors to determine statistical significance for the split-sample equivalent models. Thus, we hold the standard errors fixed so that the changes in Panel B (relative to the naïve models) are driven by coefficient movement only. Among the originally significant positive estimates and originally significant negative estimates, approximately 81% of the estimates change either their sign or significance in Panel A compared to 70% in Panel B. Panel B thus demonstrates that even holding standard errors fixed, the conclusions based on split-sample equivalent models are qualitatively different from the conclusions based on naïve models for the majority of estimates.

Though changes in the standard error cannot directly drive the changes in statistical significance in Panel B, it remains possible that they indirectly affect the observed sensitivity by increasing the noise of the split-sample equivalent estimate. In other words, the split-sample equivalent point estimate may differ from the naive estimate, not because either is biased, but because the split-sample equivalent estimate is noisy. The evidence described earlier (shown in Figure 4) suggests that standard error changes are not central to coefficient movement in general. To assess whether inflated standard errors explain the patterns shown in Table 3 specifically, we replicate Panel B of Table 3 for just the subset of cases where the split-sample equivalent model has smaller standard errors than the naïve model. For this subsample, in 92% of cases that were originally statistically significant, the split-sample equivalent estimate becomes insignificant due to coefficient movement alone. We view this last result as providing support for the conclusion

that coefficient changes are not driven by falling precision since this analysis is restricted to cases where the split-sample equivalent model is more precise.[25]

## VIII. Power considerations and conclusion

Though not explicitly argued by any of the articles we reviewed, one reason for omitting X*H interactions is to potentially increase power. Some fields use diagnostics such as variance-inflation factors to help guide the bias-efficiency tradeoff, but economists generally rely more on theoretical considerations to determine which controls are important to include. If including X is well-motivated theoretically when studying the effect of T, then there is often theoretical justification for controlling for X*H when studying the effect of T*H. Though power differences across models should certainly be considered, our replications suggest that there are many cases with little power loss and large changes in the estimated coefficient. The reduction in standard errors from excluding X*H controls is less than 20% in forty-five percent of cases and of these, 43% have estimated biases larger than 100% of the split-sample equivalent estimate.

Because power and bias considerations are necessarily context-specific, we do not aim to provide universal guidance regarding whether X*H should be controlled for. That said, the process for determining whether to include X*H should be similar to the process that determines the X vector in the first place. Current practice frequently excludes all of the X*H interactions with no discussion, while including an extensive set of uninteracted covariates in X that have similar implications for power. We suspect that in many cases, it might be more prudent to control for some X*H interactions rather than additional X covariates, but thus far researchers

[25] It is worth emphasizing that focusing on cases where the standard error becomes smaller does not address the broader point that change in statistical significance could come from random fluctuations in the coefficient. The key point is that research conclusions would have differed had the researchers used the split-sample equivalent model, even for cases where the standard error is smaller in the split-sample equivalent model.

have mostly not seriously considered the X*H controls because of a perception that they are of secondary importance.

The fact that controlling for X*H makes it equivalent to estimating separate models for H=1 vs H=0 is helpful for thinking through power considerations. For example, some might worry that in the Gong, Lu, and Song (2018) example, interacting every fixed effect with student gender leads to too many parameters and an unwieldly, noisy model. While there can certainly be power loss, estimating separate models for boys and girls allows fixed effects to vary across samples and therefore has the same number of parameters. Conceptually, it seems appealing to start with a model estimated separately for the two subgroups of interest, and then to ask whether there is theoretical justification to force some of the parameters to be the same in the interest of power.

Addressing the form of omitted variable bias we highlight is critical given its empirical implications. Eighty-one percent of the papers we replicate have at least one estimate where the estimated bias is larger than the split-sample equivalent estimate. Combining our estimates from the manual review and keyword search suggests that these cases constitute approximately 9% of all articles published and 26% of reduced form articles.[26] As such, substantively important biases are sufficiently common so that the omission of X*H interactions should not be taken lightly. To fully understand how concerning the omitted variable bias is requires a more nuanced understanding of the contexts than our summary measures can provide. That said, taken as a whole, the three approaches suggest that omitting X*H is often consequential.

Ultimately, the omitted variables we highlight are like any other and theoretical considerations that are context-specific need to guide decisions regarding their inclusion. Though

[26] This calculation assumes that the manual and keyword searches are representative of the same population.

excluding X*H may be reasonable in certain contexts, it requires justification. Given the extensive discussion of other potential biases present in most empirical work and the absence of discussion related to the inclusion of X*H terms, we suspect that the common practice of omitting these interactions may be a result of oversight rather than intentional choice.

**Bibliography**


Anderson, S. (2018). Legal origins and female HIV. *American Economic Review*, *108*(6), 1407-39.

Anderson, S., Francois, P., & Kotwal, A. (2015). Clientelism in Indian villages. *American Economic Review*, *105*(6), 1780-1816.

Athey, S., and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments* (Vol. 1, pp. 73–140). Elsevier.

Bazzi, S., Gaduh, A., Rothenberg, A. D., & Wong, M. (2016). Skill transferability, migration, and development: Evidence from population resettlement in Indonesia. *American Economic Review*, *106*(9), 2658-98.

Balli, H., and Sørensen, B. (2013). Interaction effects in econometrics. *Empirical Economics,* 45(1), 583-603.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in differences estimates? *The Quarterly Journal of Economics*, 119 (1), 249-275.

Bobonis, G. J., Cámara Fuertes, L. R., & Schwabe, R. (2016). Monitoring corruptible politicians. *American Economic Review*, *106*(8), 2371-2405.

Borusyak, K. and Jaravel, X. (2017). Revisiting Event Study Designs. Working Paper.

Callaway, B., and Sant'Anna, P. (2021). Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment. *Journal of Econometrics*, forthcoming.

Favara, G., & Imbs, J. (2015). Credit supply and the price of housing. *American Economic Review*, *105*(3), 958-92.

Frydman, C., & Hilt, E. (2017). Investment banks as corporate monitors in the early twentieth century United States. *American Economic Review*, *107*(7), 1938-70.

Gertler, P. J., Shelef, O., Wolfram, C. D., & Fuchs, A. (2016). The demand for energy-using assets among the world's rising middle classes. *American Economic Review*, *106*(6), 1366-1401.

Gibbons, C., Suárez Serrato, J.C., and Urbancic, M. (2018). Broken or fixed effects? *Journal of Econometric Methods*, 8(1).

Giesselmann, M., and Schmidt-Catran, A. (2018). Getting the within estimator of cross-level interactions in multilevel models with pooled cross-sections: Why country dummies (sometimes) do not do the job. *Sociological Methodology*, 0081175018809150.

Goldsmith-Pinkham, P., Hull, P., Kolesar, M. (2021). On estimating multiple treatment effects with regression. Working paper.

Gong, J., Lu, Y., and Song, H. (2018). The Effect of Teacher Gender on Students' Academic and Noncognitive Outcomes. Journal of Labor Economics, 36(3), 743-778.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics,* 225(2).

Greenland, Sander, Malcolm Maclure, James J. Schlesselman, Charles Poole, and Hal Morgenstern. 1991. "Standardized Regression Coefficients: A Further Critique and Review of Some Alternatives." *Epidemiology, 2*(5), 387–92.

Hsu, J. W., Matsa, D. A., & Melzer, B. T. (2018). Unemployment insurance as a housing market stabilizer. *American Economic Review*, *108*(1), 49-81.

Huang, Z., Li, L., Ma, G., & Xu, L. C. (2017). Hayek, local information, and commanding heights: Decentralizing state-owned enterprises in China. *American Economic Review*, *107*(8), 2455-78.

Jensen, R., & Miller, N. H. (2018). Market integration, demand, and the growth of firms: Evidence from a natural experiment in India. *American Economic Review*, *108*(12), 3583-3625.

Kaur, S. (2019). Nominal wage rigidity in village labor markets. *American Economic Review*, *109*(10), 3585-3616.

Levinson, A. (2016). How much energy do building energy codes save? Evidence from California houses. *American Economic Review*, *106*(10), 2867-94.

Markevich, A., & Zhuravskaya, E. (2018). The economic effects of the abolition of serfdom: Evidence from the Russian Empire. *American Economic Review*, *108*(4-5), 1074-1117.

Muralidharan, K., Romero, M., and Wüthrich, K. (2021). Factorial designs, model selection, and (incorrect) inference in randomized experiments. Working paper.

Qin, B., Strömberg, D., & Wu, Y. (2018). Media bias in China. *American Economic Review*, *108*(9), 2442-76.

Rajan, R., & Ramcharan, R. (2015). The anatomy of a credit crisis: The boom and bust in farm land prices in the United States in the 1920s. *American Economic Review*, *105*(4), 1439-77.

Słoczyński, T. (2020). Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights. Review of Economics and Statistics, forthcoming.

Sun, L., and Abraham, S. (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, forthcoming.

Voena, A. (2015). Yours, mine, and ours: Do divorce laws affect the intertemporal behavior of married couples? *American Economic Review*, *105*(8), 2295-2332.

Xu, G. (2018). The costs of patronage: Evidence from the British empire. *American Economic Review*, *108*(11), 3170-98.

**Table 1: Gong et al. (2018) replication**

***Panel A: Cognitive outcomes***

| | Estimates from Gong et al. | | Controlling for X*H interactions | |
|---|---|---|---|---|
| | Test Score | Self-Assessment | Test Score | Self-Assessment |
| Fem Stud. X Fem Subj. Tch | 0.183*** | 0.289*** | 0.0769** | 0.0741* |
| | (0.0518) | (0.0444) | (0.0345) | (0.0419) |
| Female Subject Teacher | -0.0357 | -0.125*** | 0.0162 | -0.0178 |
| | (0.032) | (0.0292) | (0.0282) | (0.0279) |
| N | 18202 | 18601 | 18202 | 18601 |

***Panel B: Non-cognitive outcomes***

| | Estimates from Gong et al. | | | | Estimates controlling for X*H interactions | | | |
|---|---|---|---|---|---|---|---|---|
| | Blue | Depressed | Unhappy | Pessimistic | Blue | Depressed | Unhappy | Pessimistic |
| Fem Stud. X Fem Head Tch | -0.173*** | -0.198*** | -0.140** | -0.0245 | -0.00207 | -0.0539 | 0.0468 | 0.0364 |
| | (0.0604) | (0.0568) | (0.0562) | (0.0446) | (0.101) | (0.0801) | (0.102) | (0.104) |
| Female Head Teacher | 0.0233 | 0.09 | 0.0525 | 0.00659 | -0.0638 | 0.0182 | -0.0382 | -0.0244 |
| | (0.0506) | (0.0616) | (0.0559) | (0.0663) | (0.062) | (0.0736) | (0.0622) | (0.0685) |
| N | 6895 | 6895 | 6895 | 6895 | 6895 | 6895 | 6895 | 6895 |

Notes: The first set of columns in each panel show the original specification from Gong et al. and the second set show the split-sample equivalent specification. The coefficient of interest is Female Teacher * Female Student, which is supposed to capture how the effect of teacher gender varies by student gender. All specifications also control for the main effect of student gender. The estimates from Gong et al. control for school-by-grade fixed effects, subject fixed effects and student observable characteristics. The estimates that control for X*H interact each of the original controls with student gender to make the specification equivalent to estimating separate models for boys and girls.

***Table 2: Describing the magnitude of the estimated bias***

**Panel A**

| | | | Paper level | |
|---|---|---|---|---|
| | All Estimates | Naive is significant | All Estimates | Naïve is significant |
| | Proportion | Proportion | Proportion | Proportion |
| Naive estimate is more than 20 percent off | 0.875 | 0.867 | 0.868 | 0.825 |
| Naive estimate is more than 30 percent off | 0.815 | 0.8 | 0.804 | 0.692 |
| Naive estimate is more than 50 percent off | 0.71 | 0.711 | 0.707 | 0.598 |
| Naive estimate is more than 70 percent off | 0.62 | 0.644 | 0.625 | 0.539 |
| Naive estimate is more than 100 percent off | 0.495 | 0.556 | 0.502 | 0.462 |
| Naive estimate is more than 200 percent off | 0.315 | 0.389 | 0.247 | 0.278 |
| Naive estimate is more than 300 percent off | 0.215 | 0.278 | 0.183 | 0.247 |
| Number of observations | 200 | 90 | 17 | 16 |

**Panel B**

| | Paper level | |
|---|---|---|
| | All Estimates | Naïve is significant |
| At least 1 naive estimate is more than 20 percent off | 1 | 1 |
| At least 1 naive estimate is more than 30 percent off | 1 | 0.875 |
| At least 1 naive estimate is more than 50 percent off | 0.941 | 0.812 |
| At least 1 naive estimate is more than 70 percent off | 0.941 | 0.812 |
| At least 1 naive estimate is more than 100 percent off | 0.941 | 0.812 |
| At least 1 naive estimate is more than 200 percent off | 0.706 | 0.562 |
| At least 1 naive estimate is more than 300 percent off | 0.706 | 0.562 |
| Number of observations | 17 | 16 |

*Notes:* Table 2 shows the distribution of bias at both the estimate and paper level. Each row shows the proportion of cases where the naïve estimate differs from the split-sample equivalent estimate by more than X%. Column 1 includes all 200 estimates. Column 2 focuses only on cases where the naïve estimate was statistically significant at the 5% level. Columns 3 and 4 of Panel A are analogous, but estimates are weighted so that each paper contributes equally. In Panel B, we report the proportion of papers that have any instances of bias larger than X%.

**Table 3: Naïve versus Split-sample equivalent estimates**

**Panel A**

| | *Split-sample equivalent estimates* | | | |
|---|---|---|---|---|
| *Naive estimates* | Neg. Signif. | Insignif. | Pos. Signif. | Total |
| Neg. Signif. | 14 | 28 | 1 | 43 |
| Insignif. | 6 | 100 | 4 | 110 |
| Pos. Signif. | 1 | 43 | 3 | 47 |
| Total | 21 | 171 | 8 | 200 |

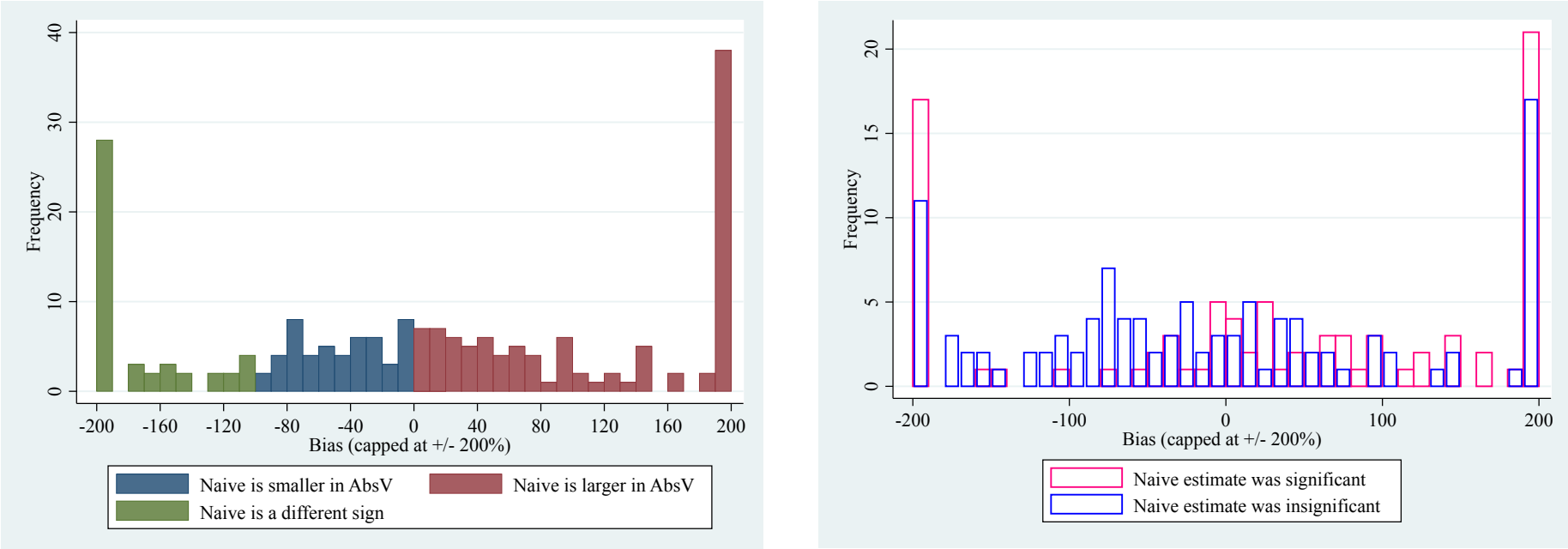
**Panel B**

| | *Split-sample equivalent estimates with naive standard errors* | | | |
|---|---|---|---|---|
| *Naive estimates* | Neg. Signif. | Insignif. | Pos. Signif. | Total |
| Neg. Signif. | 17 | 22 | 4 | 43 |
| Insignif. | 14 | 88 | 8 | 110 |
| Pos. Signif. | 1 | 36 | 10 | 47 |
| Total | 32 | 146 | 22 | 200 |

*Notes:* Table 3 presents a cross-tabulation of the significance (at the 5% level) and sign of the naïve estimates versus the split-sample equivalent estimates. For both the naïve and the split-sample equivalent estimates, we follow the original paper's approach to constructing standard errors (for example, the level of clustering). In Panel B, both the naïve and the split-sample equivalent estimates are divided by the naïve standard errors in order to construct t-statistics and assess statistical significance.

**Figure 1: Histogram of Magnitude of Estimated Bias in Naïve Interacted Model Estimates**

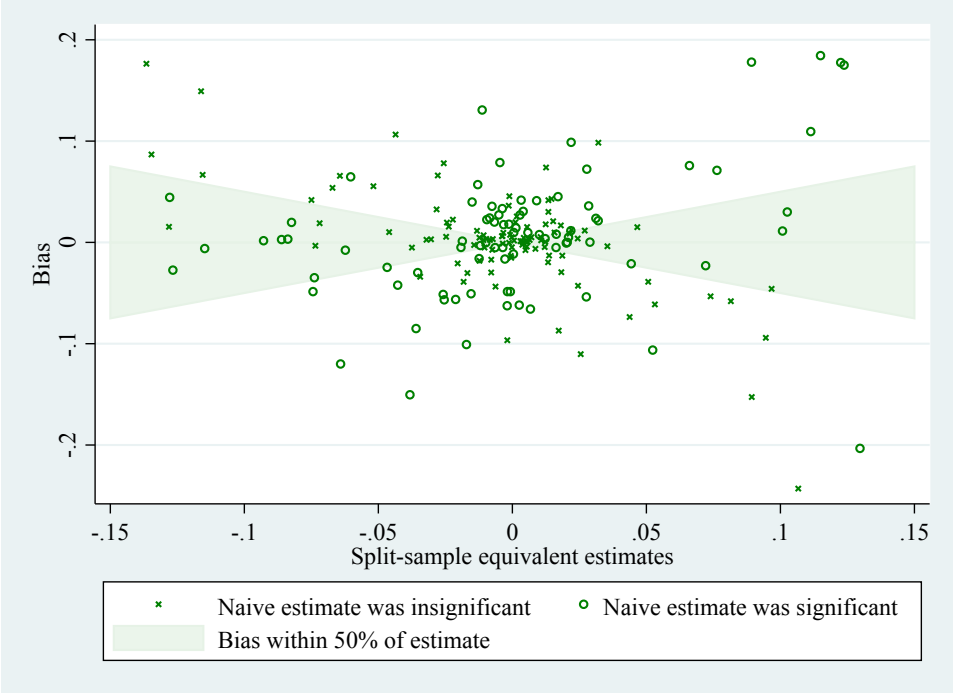
**Panel A (all estimates combined)** **Panel B (split by statistical significance of naïve estimate)**



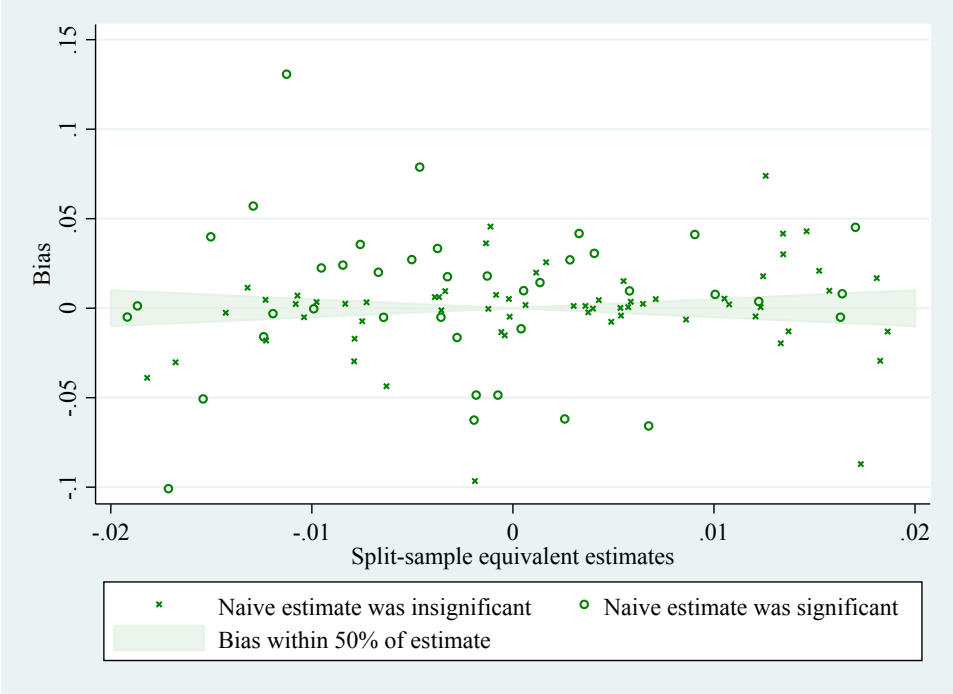

*Notes:* Figure 1 shows the distribution of the magnitude of the estimated bias in the naïve interacted model estimates relative to the split-sample equivalent estimates. Estimated bias refers to the difference between the split-sample equivalent estimate and the naïve estimate divided by the split-sample equivalent estimate. The bias is capped at +/- 200% to prevent the scale from being distorted. A 100% bias corresponds to the case where the naïve estimate is twice the magnitude (in absolute value) of the split-sample equivalent estimate. A -200% bias corresponds to the case where the split-sample equivalent estimate and the naïve estimates have the same magnitudes and opposite signs. Panel A shows the distribution of the bias for all estimates combined. Panel B shows the distribution of the bias for estimates separately based on whether or not the naïve estimate is statistically significant at the 5% level.

**Figure 2: Magnitude of Estimated Bias versus Split-sample Equivalent Estimate**

**Panel A (bias in SD units)**

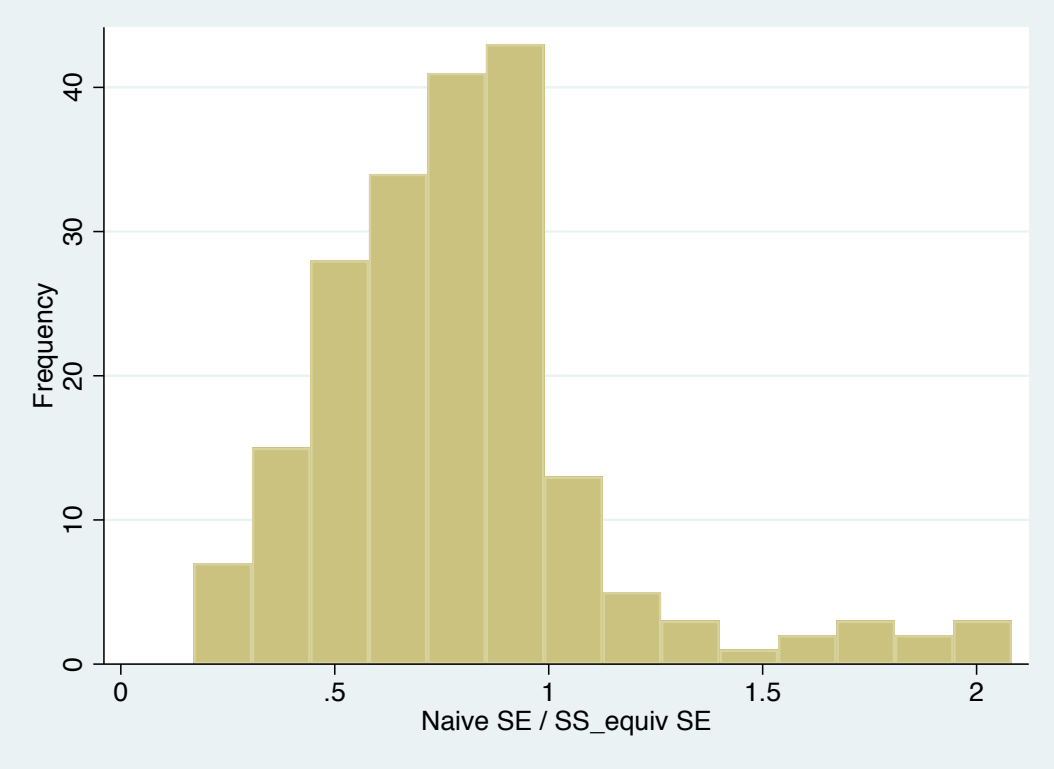



**Panel B (magnified version of panel A)**



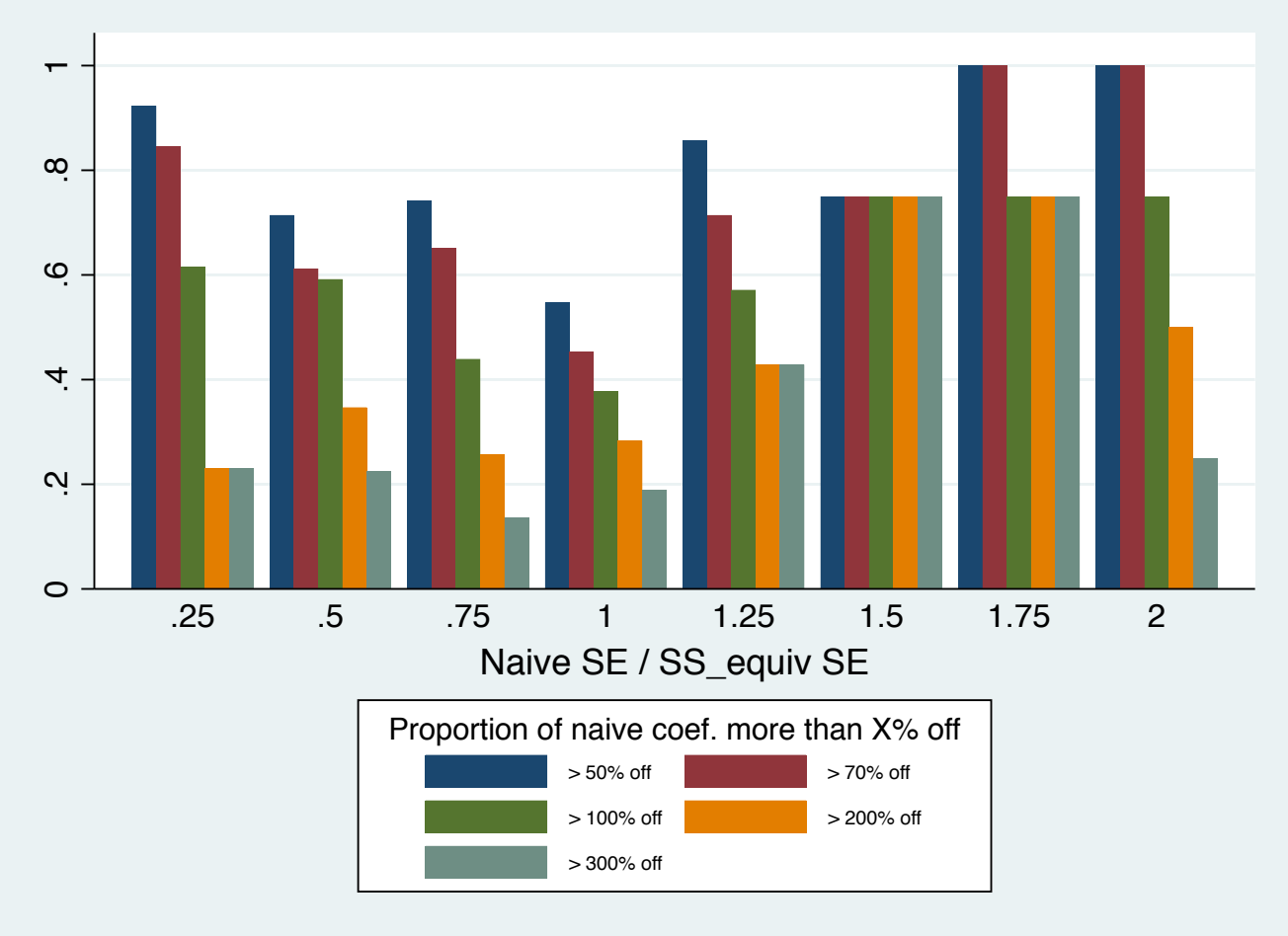

*Notes:* Figure 2 shows a scatter plot of the magnitude of the estimated bias in standard deviation units (along the vertical axis) versus the split-sample equivalent estimate in standard deviation units (along the horizontal axis). Panel B is identical to panel A except that the figure includes only split-sample equivalent estimates between -0.02 and 0.02. To avoid scale distortion, we exclude the 10 most extreme outliers (in terms of bias) from panel A.

**Figure 3: Histogram of Standard Error Ratios**




*Notes:* Figure 3 shows a histogram of the ratio of the naïve standard error to the split-sample equivalent standard error for the T*H coefficient. Ratios larger than 1 correspond to cases where the split-sample equivalent standard error is smaller than the naïve standard error.

**Figure 4: Describing the Magnitude of the Estimate Bias, split by the Standard Error Ratio**

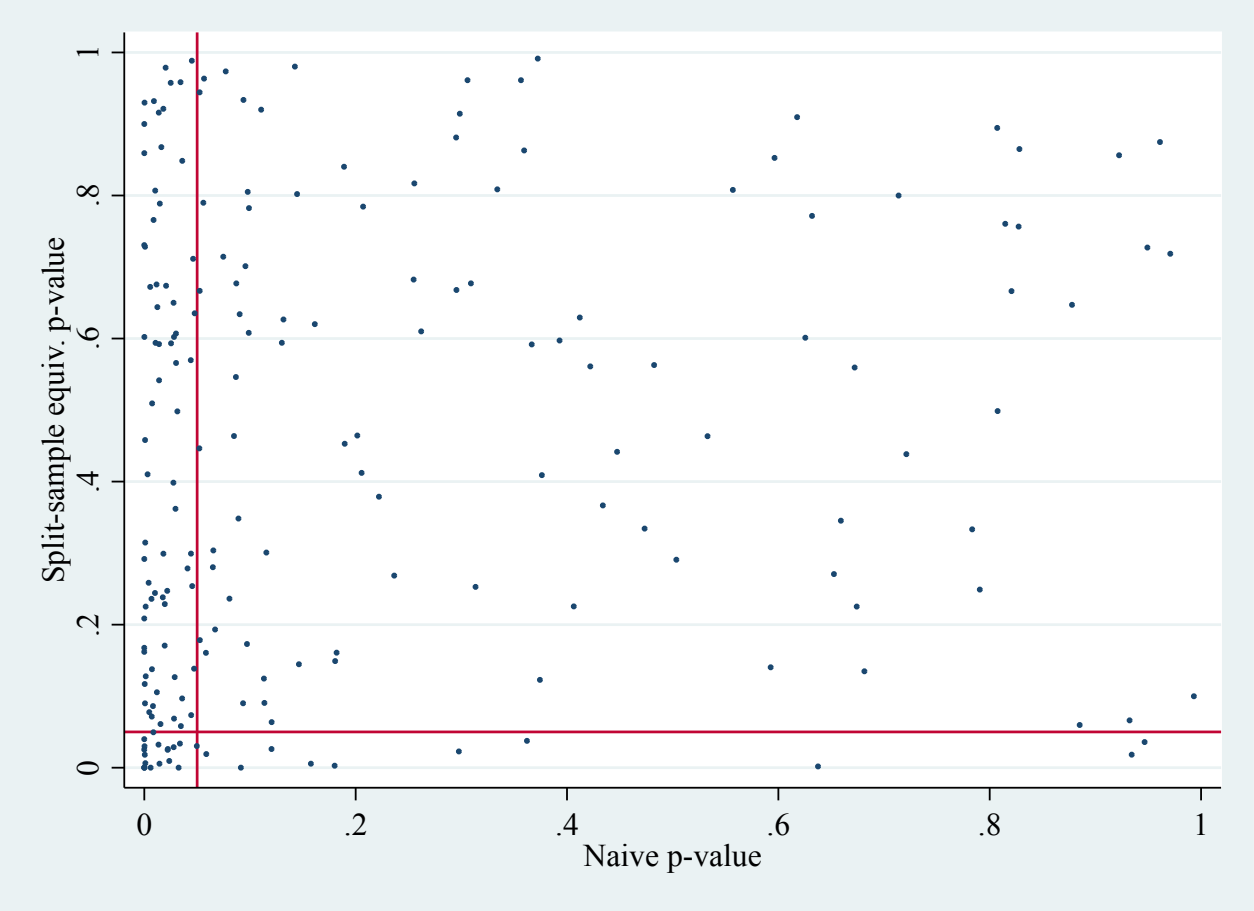



*Notes:* Figure 4 shows the proportion of estimates where the estimated bias is larger than X% off (similar to Table 2). The x-axis is the ratio of the naïve standard error to the split-sample equivalent standard error, rounded to the nearest 0.25.

**Appendix Table A1: Papers included in replication**

| *Paper reference:* | Number of specifications | Number of specifications that were statistically significant in the original paper |
|---|---|---|
| Anderson (2018) | 3 | 2 |
| Anderson Francois and Kotwal (2015) | 50 | 25 |
| Bazzi et al. (2016) | 4 | 2 |
| Bobonis, Camara Fuertes and Schwabe (2016) | 13 | 2 |
| Favara and Imbs (2015) | 4 | 4 |
| Frydman and Hilt (2017) | 8 | 0 |
| Gertler et al. (2016) | 3 | 2 |
| Hsu, Matsa, and Melzer (2018) | 2 | 2 |
| Huang et al. (2017) | 22 | 4 |
| Jensen and Miller (2018) | 11 | 7 |
| Kaur (2019) | 18 | 8 |
| Levinson (2016) | 22 | 15 |
| Markevich and Zhuravskaya (2018) | 5 | 5 |
| Qin, Strömberg, and Wu (2018) | 7 | 4 |
| Rajan and Ramcharan (2015) | 13 | 3 |
| Voena (2015) | 2 | 2 |
| Xu (2018) | 13 | 3 |
| Total | 200 | 90 |

**Appendix Figure A1**




Notes: The figure plots the p-values testing whether the coefficient on the interaction term T*H is zero. The x-axis gives the p-value for the naïve model that excludes X*H interactions and the y-axis gives the p-value from the split-sample equivalent model. The 2 red lines correspond to p-values of 0.05.

**Appendix B: Search procedure**

On April 22nd of 2020, we performed a search of all articles published from 2015-2019 in the *American Economic Review* using EBSCOhost to search "Academic Search Complete".

The search aimed to identify articles that use interaction terms to study treatment effect heterogeneity in a setting where T is only conditionally exogenous.

To do this, we used EBSCOhost and performed a search with the following components.

- American economic review from 2015-2019, excluding Papers and Proceedings.
- A match to any of the following
    - "interaction term*"
    - "interacted with"
    - "include* interactions"
    - "include* an interaction"
    - "include* the interaction"
    - "add* an interaction"
    - "add* interactions"
    - (equation* or table* or row* or column*) + W8 + interact*
        - The W8 operator specifies that the term "interact*" must be within 8 words of one of the terms in the parentheses. This search is intended to catch phrases such as "Column 3 includes an interaction between H and T".
- Articles that do NOT match the following
    - "AEA RCT registry"
    - Abstract includes "field experiment"
    - Abstract includes "randomization"
    - Title is "Front matter"

The included terms are meant to identify articles that may be using interaction terms. We exclude articles that are likely randomized experiments because treatment is likely to be unconditionally exogenous in those settings and therefore does not match our context.

The exact search is:

( ( ( ( SO american economic review NOT ( VI (97 OR 98 OR 99 OR 100 OR 101 OR 102 OR 103 OR 104 OR 105 OR 106 OR 107 ) AND IP 5 ) ) ) NOT SO insights AND DT 2015-2019 AND TX ( "interaction term*" or "interacted with" or "include* interactions" or "include* an interaction" or "include* the interaction" or "add* an interaction" or "add* interactions" or ((equation* or table* or row* or column*) W8 interact*) ) ) NOT TX "AEA RCT registry" NOT AB "field experiment" NOT AB randomization ) NOT TI Front matter

Using the above search as a starting point, we manually identify 62 articles that have publicly available data and replication files. Of the 62 articles, 9 are based on randomized experiments

and do not require controls. An additional 10 articles use interaction terms in a context such as a difference-in-differences or regression discontinuity model rather than studying treatment effect heterogeneity. An additional 9 articles use a phrase that happens to match our search, but do not actually use interaction terms in the analysis. 6 articles use interaction terms in a structural model. 4 articles use interaction terms but only for incidental controls. 1 article explores heterogeneity over time, but because they also include group-specific time trends in the preferred specification, there is no sensible split-sample analog of their specification. 2 articles only use interaction terms in an appendix and we exclude these because it is not central to their analysis. We are left with 21 papers that match our context.