NET WORK LISER

# Discussion Paper Series

IZA DP No. 18741

June 2026

# Misleading Estimates from Nonlinear Models with a Binary Outcome

**Brian Curran**
University of Chicago

**Bruce D. Meyer**
University of Chicago,
NBER and AEI

**Derek Wu**
University of Virginia
and IZA@LISER




ISSN: 2365-9793

IZA DP No. 18741 June 2026

# Misleading Estimates from Nonlinear Models with a Binary Outcome*

## Abstract

When estimating nonlinear models for binary outcomes, such as probit and logit models, researchers often rely on average partial effects (APEs) to summarize the effect of a regressor. Because the marginal effect of a variable in these models depends on the values of all other variables, the value of an APE hinges on the portion of the sample used for the calculations. When averaged over parts of the sample drawn from a subpopulation not used to define the object of interest, the APE may be misleading. This paper highlights common situations, such as differences-in-means with a secondary group and difference-in-differences designs, where APEs calculated for the full sample deviate from marginal effects for the appropriate part of the sample. We propose a simple and costless solution in specific cases and demonstrate through simulations that recalculating APEs over the appropriate subsample yields unbiased results. Reexamining published results from multiple papers, we find statistically significant discrepancies between the reported estimates and the appropriately calculated APEs.

## JEL classification

C25, C21, C23

## Keywords

average partial effects, nonlinear models, binary choice models, difference-in-differences, randomized controlled trials

## Corresponding author

Derek Wu
derek.wu@virginia.edu

* *Acknowledgments*: We thank Josh Angrist, Gabby Rocco, and Jeff Wooldridge for helpful comments. We also appreciate the support of the Alfred P. Sloan Foundation, the Russell Sage Foundation, the Charles Koch Foundation, the Menard Family Foundation, and the American Enterprise Institute.

# 1. Introduction

Researchers often examine discrete outcomes that have a limited set of values, such as whether an individual has children or which health insurance plan they choose. Applied research utilizes various models to analyze these outcomes, including probit and logit models for binary outcomes (i.e., which take one of two values), tobit models for censored data, and ordered probit or multinomial logit models for choices among more than two categories. These nonlinear specifications often provide more sensible predictions when the outcomes are categorical or bounded. For instance, probit or logit models enforce the plausible restriction that marginal effects of variables and their combinations diminish as predictions approach zero or one. Consequently, unlike a linear probability model (LPM), these models have the advantage of always giving fitted values between zero and one.

While these nonlinear models have important advantages, interpreting their outputs can be challenging. Thus, researchers often rely on summary statistics to provide a clearer understanding of how an outcome changes with a given regressor. With models for binary outcomes, researchers commonly calculate an average partial effect APE($x$) to summarize the effect of a regressor $x$, holding other variables constant. The standard practice for calculating an APE is to take the average over the entire sample; however, in nonlinear models, the marginal effect of a variable depends on the values of all other variables in the model. This dependency in nonlinear models contrasts with LPMs, where the marginal effect of each variable is constant regardless of the other variables' values. As a result, the value of the APE in a nonlinear model hinges on what part of the sample is used for averaging. If averaged over portions of the sample that are not relevant for defining the object of interest, then the calculated APE may deviate from the most relevant marginal effect. This phenomenon implies that APEs from nonlinear models such as probit and logit models can systematically differ from LPM coefficients, a result at odds with what might be inferred from standard econometric textbooks.[1]

[1] Angrist and Pischke (2009), for example, state "…in our empirical experience, the average derivatives (also called "marginal effects") constructed from parametric nonlinear models for limited dependent variables (e.g., Probit or Tobit) are usually indistinguishable from the corresponding regression coefficients, regardless of the distribution of regressors." Similarly, Greene (2018) states that "…[the LPM] appears to reliably reproduce the average partial effects obtained from the formal models such as probit and logit." Finally, Wooldridge (2020) – while not making any strong statements on an equivalence – emphasizes an example in Section 17-1d that shows only trivial differences between the probit APEs, logit APEs, and LPM coefficients. Consequently, these discussions leave the impression that APEs from binary choice models should closely mirror LPM coefficients.

In this paper, we describe a set of common situations where APEs from nonlinear binary outcome models are likely to be misleading. This issue arises when an APE is calculated over an entire sample, even though only a subset of the sample is relevant for defining the APE. We first lay out a stylized model of the problem and describe two classes of applied problems that are approximated by this model. The first is differences-in-means with a secondary group, for which researchers might run a regression of a binary outcome on mutually exclusive and exhaustive group indicators but focus on the estimate for a single indicator. This is often done with randomized controlled trials (RCTs) with multiple arms. The second is difference-in-differences, for which the APE should be calculated over treated individuals during the post-treatment period. We then simulate the degree to which probit APEs (as commonly calculated) differ from relevant marginal effects, showing that the approach used in many papers and suggested by textbooks can yield misleading estimates. For simplicity, we focus on probit as our prototypical nonlinear model, although the results generalize to other nonlinear models. We suggest a simple solution to the problem in the cases we study: calculating APEs using only the appropriate subsample.

We then show that published results often differ from this ideal solution, after recalculating empirical estimates from nine papers released since 2010. These papers span a wide variety of topics, such as the labor supply effects of the Earned Income Tax Credit (EITC), the effects of transfers on intimate partner violence, and the gains from liberalizing tariffs. We find that the published results are often statistically significantly different from the appropriately specified marginal effect. The magnitudes of the differences are also large (exceeding 100% of the baseline estimate) in several cases. While these biases are not always empirically important, the correction is straightforward and costless, and it ensures that the estimated parameter accurately reflects the object of interest. We provide guidance to practitioners by showing that differences are larger when the appropriate subsample is a smaller share of the overall sample or has an average slope of the probit function (i.e., normal density) that markedly differs from that of the non-relevant subsample.

The rest of the paper proceeds as follows. Section 2 lays out a simple stylized model of the issue and discusses its relevance to two classes of applied problems. Section 3 discusses the results of Monte Carlo simulations showing the biases from calculating the probit APE over the entire sample in these cases. Section 4 lays out the empirical applications, showing the biases associated with the uncorrected problem that we highlight. Section 5 concludes.

# 2. Stylized Model

Consider a binary dependent variable $y \in \{0,1\}$ and three mutually exclusive and exhaustive groups represented by dummy variables $x_0$, $x_1$, and $x_2$ (corresponding to assignment in Groups 0, 1, and 2). We refer to Groups 0 and 1 as the primary groups and Group 2 as the secondary group. This setup is illustrative for a differences-in-means design with a secondary group. We are interested in estimating the following model, which can be expressed in simplest LPM terms as:

$$\Pr(y = 1 \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{1}$$

where Group 0 is the omitted group. We are interested in estimating the effect of being in Group 1 relative to Group 0, which is equivalent to $\mathrm{E}[y = 1 \mid x_1 = 1] - \mathrm{E}[y = 1 \mid x_1 = 0]$ and is the coefficient on Group 1 ($\beta_1$). For now, we ignore any differences between the estimated and true values of the $\beta$s, which can be justified in large samples.

Because this is a linear specification, the average discrete difference in $y$ due to $x_1$ is always $\beta_1$, regardless of whether equation (1) is estimated over the entire sample (including $x_2$) or over just Groups 1 and 0 (excluding $x_2$). While it is possible in this setting to obtain $\beta_1$ by simply calculating the mean difference in $y$ between Groups 0 and 1, we may want to in practice control for a vector $\mathbf{Z}$ of covariates – which would require running a regression to obtain $\beta_1$. In this case, we may want to include the observations from Group 2 to help estimate the coefficients on $\mathbf{Z}$ and reduce the variance of the estimates of the $\beta$s. The example that motivated this setting was the problem of estimating the difference in a well-being indicator between two groups, and we included a third group to more precisely estimate how the indicator changed with age and other covariates (Meyer, Wu, and Curran 2020).

Alternatively, we can estimate effects using a probit specification:

$$\Pr(y = 1 \mid x) = \Phi(\beta_0' + \beta_1' x_1 + \beta_2' x_2), \tag{2}$$

where $\Phi$ is the standard normal cumulative distribution function (CDF) and $\beta_0'$, $\beta_1'$, and $\beta_2'$ are the coefficients from the probit regression. Unlike estimates from an LPM, the marginal effects from a probit model are no longer constant because the cumulative normal probit function $\Phi$ is nonlinear. In particular, the average discrete difference (i.e., average partial effect for a non-continuous regressor) due to $x_1$ is given by:

$$\text{APE}(x_1) = n^{-1} \sum_i \left[\Phi(\beta_0' + \beta_1' + \beta_2' x_{2i}) - \Phi(\beta_0' + \beta_2' x_{2i})\right], \tag{3}$$

where observations are indexed by $i$ and $n$ denotes the overall sample size.[2] Note that the average discrete difference due to $x_1$ in equation (3) incorporates $\beta_2' x_2$ in the normal CDF $\Phi$, so taking the average over the entire sample (including Group 2) may lead to misleading estimates depending on the size of Group 2 and the change in $\Phi$ as a function of $\beta_2' x_2$. As a result, this reveals a key feature of probit APEs: they will only be unbiased when calculated over the parts of the sample that are relevant for defining them.

In the case of this stylized setting, it is relatively straightforward to identify the relevant subsample that provides the appropriate contrast: observations in Groups 0 and 1. Calculating the APE due to $x_1$ over only these primary groups yields:

$$\text{APE}^{Restricted}(x_1) = (n_0 + n_1)^{-1} \sum_{i \in \{0,1\}} \left[\Phi(\beta_0' + \beta_1') - \Phi(\beta_0')\right], \tag{4}$$

where $n_0$ and $n_1$ are the sample sizes of Groups 0 and 1, respectively. The APE calculated in equation (4) provides an estimate of the desired quantity, unlike the estimate calculated over the full sample in equation (3).[3] In more complex examples where the identifying variation may come from continuous regressors, it will be more difficult to determine an appropriate subsample to use.

Another prototypical setting with a similar setup is difference-in-differences (DiD). Given a simple two-group, two-period setting, we again have three dummy variables $x_1$, $x_2$, and $x_3$, where $x_1$ is an indicator for assignment to group $G \in \{0,1\}$, $x_2$ is an indicator for being in time period $T \in \{0,1\}$, and $x_3$ is the interaction of the group and time dummies. We now estimate the following probit model:

$$\Pr(y = 1 \mid x) = \Phi(\beta_0' + \beta_1' x_1 + \beta_2' x_2 + \beta_3' x_3), \tag{5}$$

where the APE of $x_3$ represents the DiD estimate (the treatment effect of interest). Assuming that the probit model is the true model, the marginal effect of $x_3$ is equal to:

$$\Pr(y = 1 \mid x_3 = 1) - \Pr(y = 1 \mid x_3 = 0) = \Phi(\beta_0' + \beta_1' + \beta_2' + \beta_3') - \Phi(\beta_0' + \beta_1' + \beta_2').$$

[2] If $x_1$ were continuous, then the average partial effect would be expressed as $\text{APE}(x_1) = n^{-1} \Sigma_i \varphi(\beta_0' + \beta_1' x_{1i} + \beta_2' x_{2i}) \beta_1'$, where $\varphi$ is the standard normal density.

[3] An alternative (but equivalent) approach to calculating an unbiased APE of $x_1$ in our stylized model would be to calculate it over the entire sample but purge the influence of Group 2 by setting $x_2 = 0$ for all observations. This approach could be interpreted as the effect of switching from Group 0 to Group 1, calculated over the full sample.

One can recover the relevant marginal effect by calculating the APE over the observations in Group 1 ($x_1 = 1$) during time period 1 ($x_2 = 1$):

$$\text{APE}^{Restricted}(x_3) = (n_{G=1,T=1})^{-1} \sum_{G=1,T=1} [\Phi(\beta_0' + \beta_1' + \beta_2' + \beta_3') - \Phi(\beta_0' + \beta_1' + \beta_2')], \quad (6)$$

where $n_{G=1,T=1}$ denotes the number of observations in the relevant subsample. An alternative interpretation of equation (6) is that it reflects the average treatment effect for treated observations (ATT) during the post-treatment period, which is the object of interest in a DiD design. In contrast, calculating the APE over the full sample would yield:

$$\text{APE}^{Full}(x_3) = n^{-1} \sum_i [\Phi(\beta_0' + \beta_1' x_1 + \beta_2' x_2 + \beta_3') - \Phi(\beta_0' + \beta_1' x_1 + \beta_2' x_2)], \quad (7)$$

which may lead to misleading estimates since it is averaged over irrelevant observations that are untreated ($x_1 = 0$) or observed prior to treatment ($x_2 = 0$).

Intuitively, the curvature of the normal CDF implies that the marginal effects of treatment vary depending on the baseline probabilities. These can differ across time periods, aligning with the idea that policy impacts are often time-specific. Additionally, the untreated group – used to identify the time-fixed effect – can differ fundamentally from the treated group and therefore may not respond to treatment in the same way. These differences in baseline probabilities and responses underscore why restricting the APE calculation to treated observations ($G = 1$) during the post-treatment period ($T = 1$) ensures alignment with the object of interest.

Notably, when the true functional form of a DiD specification follows a probit model, even the LPM can yield biased estimates of the relevant treatment effect. This point is embedded in Puhani (2012) and discussed again in Wooldridge (2023), but it can be thought of as a specific example of the general problem discussed in this paper. This bias arises from the LPM's assumption that the marginal effect of $x_3$ is constant, whereas the probit model accounts for the true nonlinear relationship between the covariates and the outcome. To see this more formally, the LPM coefficient for $x_3$ represents the average difference in probabilities between treated and untreated observations before and after treatment: $[\Phi(\beta_0' + \beta_1' + \beta_2' + \beta_3') - \Phi(\beta_0' + \beta_2')] - [\Phi(\beta_0' + \beta_1') - \Phi(\beta_0')]$. This is in general not equal to the relevant APE in equation (6) unless the group- and time-fixed effects $\beta_1'$ or $\beta_2'$ are zero. Ultimately, these concerns about functional form are distinct from the primary issue addressed in this paper, which centers on calculating the APE over the subsample that defines the object of interest conditional on the relevant data generating process.

# 3. Simulations

In this section, we use Monte Carlo simulations to shed light on the magnitude of potential biases when the probit APE is calculated over the full sample rather than a smaller sample from the subpopulation that provides the object of interest. We examine two types of simulations that correspond to our two settings. The first simulation estimates differences-in-means between two groups that are drawn from a subset of the entire population. The second simulation estimates a difference-in-differences model with two groups and two periods. The estimates of the simulations are averaged over 1,000 replications, each of which has a sample size of 10,000 observations.

### 3.1. Simulation 1: Differences-in-Means with a Secondary Group

Our first simulation takes the form of differences-in-means with a secondary group. Consider a sample divided into three mutually exclusive and exhaustive groups designated by $x \in \{0,1,2\}$. In the simulation, the goal is to learn the difference in an outcome between Groups 1 and 0. As a share of the entire sample, the relative sizes (proportions) of Groups 0, 1, and 2 are $p_0$, $p_1$, and $p_2$, where $p_0 + p_1 + p_2 = 1$. The conditional means of $y$ for each group are $\bar{y}_{x=0}$, $\bar{y}_{x=1}$, and $\bar{y}_{x=2}$. Because Group 2 is the secondary (irrelevant) group, the properly specified APE should be calculated over Groups 0 and 1 only (omitting Group 2). This setup is analogous to a randomized controlled trial with multiple disjoint treatment arms (Groups 1 and 2) and a control arm (Group 0). We are interested in the causal effect of treatment arm 1 (relative to no treatment) on some binary outcome and include dummy variables for assignment into both treatments as separate regressors in a probit specification.

We estimate the APE in three ways: 1) probit specification where the APE is calculated over the full sample (which we call the "full-sample probit APE"), 2) probit specification where the APE is calculated over the relevant subsample (which we call the "restricted-sample probit APE"), and 3) standard LPM estimated over the entire sample.[4] We show the results of Monte Carlo simulations that vary the relative size ($p_2$) and conditional mean ($\bar{y}_{x=2}$) of the secondary group. We always set $\bar{y}_{x=0} = 0.05$ and $\bar{y}_{x=1} = 0.25$, meaning the true value of the APE is 0.2. To

[4] For (1) and (2), note that the probit regression is still being run over the entire sample, but the samples over which the APE is calculated are different.

make things more concrete, consider a setting where individuals are assigned to take either Vaccine 1 ($x = 1$), Vaccine 2 ($x = 2$), or no vaccine ($x = 0$) and we are interested in their effects on disease immunity. Our baseline parameters would suggest that the true share of individuals with immunity is 5% in the control group and 25% in the group assigned to Vaccine 1, implying that the true effect of assignment to Vaccine 1 is an increase in disease immunity of 20 percentage points.

*Varying the Size of the Secondary Group*

We start by varying the relative size of Group 2 ($p_2$) while holding constant the conditional means of $y$ for each group. In addition to the baseline parameters for $\bar{y}_{x=0}$ and $\bar{y}_{x=1}$, we set $\bar{y}_{x=2} = 0.45$. We vary $p_2$ from 0.01 to 0.99 (in increments of 0.01) and always fix $p_0 = p_1$. In other words, when Group 2 is set to be 10% of the sample, Groups 0 and 1 are each 45% of the sample; when Group 2 is set to be 90% of the sample, Groups 0 and 1 are each 5% of the sample. Figure 1a shows how the average discrete difference in $y$ due to $x_1$ going from 0 to 1 (i.e., being in Group 1 rather than 0) varies with $p_2$. We see that the simulated average discrete differences from the restricted-sample probit APE (using only Groups 0 and 1) are nearly constant and almost exactly equal to the true effects of 0.2. They are also practically identical to the LPM estimates. However, the average discrete difference using the full-sample probit APE becomes increasingly biased as $p_2$ increases. For example, the average discrete difference is overstated by 38% when Group 2 is half of the overall sample and by 68% when Group 2 is 90% of the overall sample. The relationship between the bias and the proportional size of Group 2 is approximately linear.

*Varying the Mean of the Secondary Group*

We next vary the conditional mean of $y$ for Group 2 ($\bar{y}_{x=2}$) while holding constant the conditional means of $y$ for Groups 0 and 1 and the relative sizes of each group, which we fix at $p_0 = p_1 = p_2 = 1/3$. We vary $\bar{y}_{x=2}$ from 0.01 to 0.99 (in increments of 0.01). Figure 1b shows how the average discrete difference in $y$ from being in Group 1 rather than 0 varies with $\bar{y}_{x=2}$. Once again, the average discrete differences using the restricted-sample probit APE – which are identical to the LPM coefficients – are almost exactly equal to the true effect of 0.2, regardless of the magnitude of $\bar{y}_{x=2}$. However, the average discrete difference using the full-sample probit APE can become biased in different directions depending on the value of $\bar{y}_{x=2}$. Indeed, the full-sample probit APE produces unbiased estimates when $\bar{y}_{x=2}$ is approximately 0.05 or 0.75, while it yields an

upward bias when $\bar{y}_{x=2} \in (0.05,0.75)$ and a downward bias when $\bar{y}_{x=2} < 0.05$ or $\bar{y}_{x=2} > 0.75$. The upward bias from the full-sample probit APE is largest when $\bar{y}_{x=2}$ is approximately 0.3, suggesting the conditional mean of the secondary group need not be dissimilar from that of a primary group for bias to exist. This exercise therefore reveals that the magnitude of the bias depends less on the differences in average outcomes and more on the differences in average derivatives of the probit function or normal densities between groups. Moreover, the non-monotonic nature of the bias with respect to $\bar{y}_{x=2}$ can be attributed to the symmetry of the standard normal density.

*Varying Both the Size and Mean of the Secondary Group*

We finally vary *both* the size of Group 2 ($p_2$) and the conditional mean of $y$ for Group 2 ($\bar{y}_{x=2}$) while holding constant the conditional means of $y$ for Groups 0 and 1. Specifically, we vary $p_2$ from 0.1 to 0.9 (in increments of 0.2) and always fix $p_0 = p_1$, and we vary $\bar{y}_{x=2}$ from 0.01 to 0.99 (in increments of 0.01). Using only the full-sample probit APE, Figure 1c shows how the average discrete difference in $y$ due to $x_1$ going from 0 to 1 varies over $\bar{y}_{x=2}$ for various sizes of Group 2 ($p_2$). For any size of Group 2, it is always the case that the full-sample probit APE produces an upward bias when $\bar{y}_{x=2} \in (0.05, 0.75)$ and a downward bias when $\bar{y}_{x=2}$ lies outside of [0.05, 0.75]. The values of $\bar{y}_{x=2}$ at which the biases are largest are also identical across values of $p_2$. However, the interactive effects of $p_2$ and $\bar{y}_{x=2}$ on the bias of APE($x_1$) are large enough such that we observe almost a doubling of the estimated effect when $\bar{y}_{x=2} = 0.3$ and the size of Group 2 is 90% of the overall sample. Indeed, it is possible to devise more extreme examples where the bias is arbitrarily large. Overall, these simulations make clear that calculating the APE using the full-sample probit APE can lead to considerable discrepancies, although the sign will usually be correctly estimated.

### 3.2. Simulation 2: Difference-in-Differences

Our next simulation takes the form of difference-in-differences in a panel data setting with two groups and two time periods. Here, we would like to learn the difference in an outcome for the treated group in the time period when they are treated compared to their outcome if they were not treated in that period, while accounting for additive time and group effects. We denote the untreated and treated groups by $G = 0$ and $G = 1$, respectively, and the pre-treatment period and

period of treatment by $T = 0$ and $T = 1$, respectively. We assume an even split in observations across time periods as in a balanced panel.

Unlike in the previous simulation, we assume that the data generating process (DGP) follows a probit model specified as:

$$\Pr(y = 1 \mid G, T) = \Phi(0.1 + G + 0.3T + 0.5G*T),$$

We are interested in the average partial effect of $G*T$, whose true value is:

$$\Phi(0.1 + 1 + 0.3*1 + 0.5*1) - \Phi(0.1 + 1 + 0.3*1 + 0.5*0) \approx 0.052.$$

We simulate the APE in four ways. We start with a probit specification that calculates the APE over only the relevant subsample ($G = 1$, $T = 1$). Continuing with the probit specification, we also calculate the APE over two broader samples that include some observations not relevant for defining the object of interest: treated observations in both periods ($G = 1$, Any $T$) and all observations in the full sample (Any $G$, Any $T$). Finally, we calculate the LPM coefficient estimated over the entire sample.

Figure 2 shows the results of simulations that vary the relative size of the untreated group ($G = 0$) as a proportion of the overall sample. First, the probit APE calculated over treated observations during the post-treatment period (seen in the orange diamonds) consistently matches the true marginal effect for most sizes of the untreated group. However, when the untreated group is particularly small or large, the restricted-sample probit APE estimates become noisier and less reliable. When calculated over a broader sample that includes treated observations prior to being treated (brown squares), the probit APE becomes biased and understates the true marginal effect by approximately 50%. Yet, the estimates continue to remain stable across different sizes of the untreated group, as no untreated observations are included in the APE calculation.

Using the full sample for the APE calculations – i.e., additionally including untreated observations in both periods (green triangles) – amplifies the bias further in proportion to the size of the untreated group. Using the full-sample probit, the APE is understated by 75% when the treated and untreated groups are of the same size and by 95% when the untreated group constitutes 90% of the overall sample. While these biases can be large in magnitude, the sign of the treatment effect is still correctly estimated. However, in contrast with results of the prior simulation on differences-in-means, the LPM coefficients (blue circles) deviate substantially from the true marginal effect and can even have the wrong sign – taking a value of approximately -0.008 in this

setting. This result illustrates the risks of relying on linear models when the underlying data generating process is nonlinear.

# 4. Empirical Applications

In this section, we recalculate estimates from nine papers published or released since 2010 to show that the issue we document in this paper is common and empirically relevant. These nine papers fall into two categories corresponding to our model/simulations: one category is differences-in-means/multi-armed RCTs and the other is difference-in-differences with two groups and two time periods. Section 4.1 describes how we chose the relevant papers, and Sections 4.2 and 4.3 discuss the results of our recalculations.

## 4.1. Choice of Applications

As discussed in Sections 2 and 3, there are at least two major classes of empirical problems where misleading APEs are likely to arise. The first is RCTs with multiple treatment arms or differences-in-means with more than two groups. In this class of problems, the APE should only be calculated over the two arms one is interested in directly comparing and not any other group. The second is difference-in-differences. While misleading APEs can arise in any kind of difference-in-differences design, our stylized model and corresponding simulation only considered the case of a design with two time periods and two groups. Hence, for this section, we only looked for two-period, two-group difference-in-differences applications. In such a design, we want to calculate the APE over only the treated group in the post-treatment period.[5]

We implemented a systematic search to find published papers that fit within our framework. We started by searching on JSTOR for papers published since 2010 in the *American Economic Review* (*AER*), *American Economic Journals* (*AEJs*), *Quarterly Journal of Economics* (*QJE*), and *European Economic Review* (*EER*) that mentioned either "probit" or "logit" as well as "randomized control trial" or "difference-in-differences" in the text. We chose these journals given

[5] In certain applications, it may make more sense to calculate the APE over the treated group in both time periods. For example, if the time periods in a difference-in-differences setup have very similar settings (e.g., the economic conditions after treatment are roughly similar to those before), then it may be useful to include both time periods when calculating the APE. For this reason, we also include results calculating APEs in this way, detailed in the Appendix. Qualitatively, we find that the differences relative to full-sample probit APEs and LPM coefficients are similar regardless of how we calculate the restricted-sample probit APEs.

that they are high-profile journals in economics that also have strong data availability rules, which aids replication. From this initial search, we narrowed down the papers using two additional criteria. First, we removed any RCT/differences-in-means papers that had a single treatment arm along with any difference-in-differences papers that had more than two time periods or used a continuous treatment. Second, we removed any papers which did not have datasets and a replication package readily available via the authors' website or journal.

Using our systematic search, we found eight papers that fit all of our criteria. To these eight papers, we added the working paper that catalyzed this project (Meyer et al. 2020).[6] Of these nine papers, four are RCTs/differences-in-means while the other five are difference-in-differences. These papers tackle a wide variety of questions across many fields of economics, including the labor supply effects of the EITC, the effects of cash transfers on intimate partner violence, and the gains from liberalizing tariffs. Some of these papers use probit models in their main specifications while others use the LPM in their main tables and include probit results in the Appendix.[7] Finally, some papers do not explicitly show results using a probit specification, but instead present LPM results and state that their results are similar using a probit specification.[8]

For each of the nine papers, we calculate the probit APEs using the full sample, the probit APEs using the relevant (restricted) portion of the sample, and the LPM coefficients. We try to do this for as many unique outcomes as we can for each paper (the minimum number of outcomes for which we recalculate estimates is three while the maximum is ten). Our focus is on how the full-sample probit APEs or LPM coefficients (which reflect the published results) differ from the correctly specified restricted-sample probit APEs. To that end, we estimate fractional differences relative to the restricted-sample probit APE as well as the statistical significance of these differences. It is worth noting that LPM coefficients are not clearly inferior to probit APEs, since the choice between LPM and probit hinges on the true functional form, which will vary by application. Yet, if researchers opted to use a probit model, then there *is* a correct approach to calculating the APE that averages over the relevant subsample rather than the entire sample. Thus, while our comparisons between restricted- and full-sample probit APEs aim to correct the

[6] The problem highlighted in this paper was recognized in Meyer and Rosenbaum (2001), who focused on calculating average partial effects over single mothers only (i.e., the treatment group) rather than over all observations in their sample (which additionally consisted of single women without children).
[7] While we searched for papers that used logits, all of the papers that fit within our criteria use probits.
[8] These papers include Nudges at the Dentist (2014), Nudges at the Library (2013), and DACA and Education (2020).

calculation for accuracy, the comparisons between restricted-sample probit APEs and LPM estimates are meant to illustrate how outcomes might differ when employing a widely used alternative approach in the literature.

### 4.2. Main Results

Figure 3 summarizes our results, reporting the total number of estimates and the number of outcomes in each paper for which the full-sample probit APE is statistically significantly different from the restricted-sample probit APE.[9] This figure also displays the counts of outcomes in each paper for which the LPM coefficient differs significantly from the restricted-sample probit APE, although these differences should be viewed not necessarily as errors but as the result of a different functional form assumption. Blue bars represent the difference between the full-sample probit APE and restricted-sample probit APE, while orange bars represent the difference between the LPM coefficient and restricted-sample probit APE. Dark shading represents a difference that is statistically significant at the 10% level, while light shading represents a difference that is not statistically significant at the 10% level.

For the RCT/differences-in-means papers, we find that 13 of the 32 outcomes have a statistically significant difference (at the 10% level) between the full- and restricted-sample probit APEs. In contrast, only 6 out of the 32 outcomes have a statistically significant difference between the LPM coefficient and the restricted-sample probit APE. These results align qualitatively with our simulations in Figures 1a and 1b, in which we found pronounced differences between the full- and restricted-sample probit APEs but essentially no differences between the LPM coefficient and the restricted-sample probit APE. For the difference-in-differences papers, 9 of the 25 outcomes have a statistically significant difference between the full- and restricted-sample probit APEs, while 15 of the 25 outcomes have a statistically significant difference between the LPM coefficient and the restricted-sample probit APE.

To learn more about how outcome magnitudes in these papers differ from the properly calculated APEs, Figure 4 shows the percent difference between the full-sample estimates (either the LPM coefficient or full-sample probit APE) and the restricted-sample probit APE for the *main* outcome on which each set of authors focuses their discussion. Appendix Figure A1 is a similar

[9] Appendix Figure A4 shows the same figure but calculates the restricted-sample probit APE for difference-in-differences papers over the treated group in both time periods.

plot but instead examines the outcome with the *largest* difference between the full-sample probit APE and restricted-sample probit APE.[10] As in Figure 3, blue bars correspond to the difference between the full-sample probit APE and restricted-sample probit APE, while orange bars correspond to the difference between the LPM coefficient and restricted-sample probit APE.

We find that the absolute value of the difference between the full- and restricted-sample probit APEs ranges from 0 to 115 percent (using the restricted-sample probit APE as a baseline). Of the nine papers we consider, five have differences over 10 percent, three have differences over 20 percent, and one has a difference over 100 percent. While the differences are larger (by construction) when we look at the outcome with the largest difference in Appendix Figure A1, we still see fairly large differences even when analyzing just the main outcome.

Of the four RCT/differences-in-means papers, the differences are generally larger when using the full-sample probit APE than when using the LPM. For instance, in Geography of Disadvantage (2020), the full-sample probit APE difference is 37 percent compared to 3 percent with the LPM. In contrast, in the difference-in-differences papers, the LPM tends to have larger differences than the full-sample probit APE. The two papers that show this result most clearly are Anti-Sweatshop Activism (2010) and DACA and Education (2020). In particular, the full-sample probit APE in the latter paper differs from the restricted-sample probit APE by only 17 percent, while the LPM coefficient differs by 83 percent.[11] Both of these patterns align with our prior simulations, in which the LPM coefficients were very similar to restricted-sample probit APEs under a differences-in-means design, while they had pronounced discrepancies in a difference-in-differences design.

Overall, for the RCT/differences-in-means papers, we find that differences between the full-sample and restricted-sample probit APEs vary in magnitude and statistical significance across studies. In Nudges at the Dentist (2014), these differences are particularly pronounced in both magnitude and significance, while in Geography of Disadvantage (2020), the differences are substantial in magnitude but not statistically significant. The other two papers have small differences in general. What leads to some papers having larger differences between the full- and restricted-sample probit APEs than others? Nudges at the Dentist (2014), for example, has large

[10] Appendix Figures A5 and A6 are the corresponding figures where we calculate the restricted-sample probit APE for difference-in-differences papers over the treated group in both time periods.

[11] Notably, the main specification in *DACA and Education* (2020) relied on an LPM, indicating that the authors may have considered a linear functional form to be more appropriate, making the LPM coefficient the relevant estimate.

differences that appear to be due to two factors. First, their experiment has a large number of treatment arms (and hence the secondary group is large). Second, the treatment arms each have different effect sizes, which leads to large differences in the average density of the secondary group relative to that of the primary group. Transfers and Domestic Violence (2016) and Nudges at the Library (2013) have small differences for the opposite reasons: there are not many treatment arms and the average densities are similar across the primary and secondary groups. We thus argue that the choice of which sample to calculate the APE over is more important when there are a large number of treatment arms and the treatment arms have different effect sizes.

In the difference-in-differences papers, we again see that the papers with the largest differences between the full- and restricted-sample probit APEs are those where the secondary group (observations in the untreated group or the treated group in the pre-treatment period) makes up a large portion of the sample and where the average derivative of the probit function of the secondary observations is different from that of the primary observations. Hence, practitioners should be especially cautious in cases where a difference-in-differences design has a small treatment group or a short post-treatment period.

### 4.3. Exploring Reasons Behind Differences

We have so far posited two main factors that influence the extent of the differences between the full- and restricted-sample probit APEs: the size of the secondary group and the difference in average derivatives of the probit function between the primary and secondary groups. In this subsection, we explore these factors in a more systematic way to provide suggestive evidence that these factors do shape the difference between the full- and restricted-sample probit APEs in our sample of papers.

Figures 5a and 5b display two scatterplots that show the relationship between each of the two aforementioned factors and the magnitude of the difference between the full- and restricted-sample probit APEs. On the y-axis, we plot the absolute value of the percent difference between the full- and restricted-sample probit APEs (measured relative to the restricted-sample probit APE). Figure 5a plots this difference against the share of the sample in the secondary group, while Figure 5b plots this difference against the absolute difference in average probit function derivatives

between the primary group and secondary group. We again focus on the main outcome for each paper, but Appendix Figures A2 and A3 also show corresponding figures for the largest outcome.[12]

Figure 5a shows a strong positive relationship between the share of the sample in the secondary group and the difference between the full-sample and restricted-sample probit APEs. The line of best fit has a slope of 113.9 with standard error 43.9 (and hence is statistically significant at the 5% level). Figure 5b shows that there is likewise a strong positive relationship when we use the difference in average probit function derivatives between the primary group and the secondary group as the independent variable. For this relationship, the line of best fit has a slope of 813.6 with standard error 262.3 (and hence is statistically significant at the 5% level). These two scatterplots confirm that our initial observations from theory and simulations successfully predicted the pattern of results in this sample of papers. We see that having a larger secondary group and a bigger difference in average probit function derivatives between the primary and secondary groups are each associated with having a larger difference between the full- and restricted-sample probit APEs. These relationships are always positive in both of these figures and in the appendix figures. The fact that these relationships persist across papers with very different settings suggests that the observed patterns reflect a robust and meaningful difference.

# 5. Conclusions

In this paper, we show that average partial effects from nonlinear models can be misleading in certain situations when calculated over the full study sample. The underlying issue results from an APE being calculated over an entire sample when only a subsample is relevant for defining the object of interest. We first describe a stylized model of the problem and apply this model to two common empirical problems: differences-in-means with a secondary group (e.g., multi-armed RCTs) and difference-in-differences. We then use simulations to show that probit APEs, as commonly calculated, differ from true marginal effects.

We propose a simple and costless solution to this problem – calculating APEs over the relevant subsample – and show that this issue is often overlooked in empirical work. Recalculations of results from nine published or working papers reveal that the published results are often statistically significantly different from the appropriately calculated APEs. Our findings

[12] Appendix Figures A7-A10 are the corresponding figures where we calculate the restricted-sample probit APE for the difference-in-differences papers over the treated group in both time periods.

suggest that differences are larger when the relevant subsample either is a smaller share of the sample or has an average density that is markedly different from that of the non-relevant subsample.

Though our discussion in this paper has focused primarily on probit models, the issues we highlight extend to other nonlinear models such as logit.[13] Furthermore, even though we focus on difference-in-differences as one of the two designs analyzed, these issues are likely to be even more pronounced in extensions like triple-differences (where the relevant subsample may be an even smaller share of the overall sample). More generally, they apply to situations where covariates are continuous and where most of the identifying variation is confined to a specific subset of observations. In such cases, the challenge of determining the appropriate subsample for calculating APEs becomes even more complex. Ultimately, our results call for greater care in calculating APEs in empirical applications of nonlinear models. By focusing on the subset of the sample that provides the identifying variation, researchers can avoid potentially large biases and improve the accuracy of their estimates, leading to more reliable conclusions in applied work.

[13] Two papers using logit estimators to evaluate RCTs with multiple treatment arms are Luoto et al. (2014) and Rommel et al. (2015). Luoto et al. (2014) has 12 treatment arms, so we predict that the results would differ substantially if APEs were calculated over the relevant subsamples. However, the authors average their results over many different treatment arms, so the differences may not end up being substantial. But if they reported coefficients for each arm, we would expect large differences. On the other hand, we do not expect Rommel et al. (2015) to have very large differences because there are only three treatment arms and the effect sizes are relatively similar across the arms.

# References


**Altmann, Steffen and Christian Traxler.** 2014. "Nudges at the Dentist." *European Economic Review*, 72: 19-38.

**Angrist, Joshua D. and Jörn-Steffen Pischke.** 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press. Princeton, NJ.

**Apesteguia, Jose, Patricia Funk, and Nagore Iriberri.** 2013. "Promoting Rule Compliance in Daily-Life: Evidence from a Randomized Field Experiment in the Public Libraries of Barcelona." *European Economic Review*, 64: 266-284.

**Bailey, Martha J.** 2010. "'Momma's Got the Pill': How Anthony Comstock and Griswold v. Connecticut Shaped US Childbearing." *American Economic Review*, 100(1): 98-129.

**Bastian, Jacob.** 2020. "The Rise of Working Mothers and the 1975 Earned Income Tax Credit." *American Economic Journal: Economic Policy,* 12(3): 44-75.

**Greene, William H.** 2018. *Econometric Analysis, Eighth Edition*. Pearson.

**Harrison, Ann and Jason Scorse.** 2010. "Multinationals and Anti-Sweatshop Activism." *American Economic Review*, 100(1): 247-273.

**Hidrobo, Melissa, Amber Peterman, and Lori Heise.** 2016. "The Effect of Cash, Vouchers, and Food Transfers on Intimate Partner Violence: Evidence from a Randomized Experiment in Northern Ecuador." *American Economic Journal: Applied Economics*, 8(3): 284–303.

**Kuka, Elira, Na'ama Shenhav, and Kevin Shih.** 2020. "Do Human Capital Decisions Respond to the Returns to Education? Evidence from DACA." *American Economic Journal: Economic Policy*, 12(1): 293-324.

**Luoto, Jill, David Levine, Jeff Albert, and Stephen Luby.** 2014. "Nudging to Use: Achieving Safe Water Behaviors in Kenya and Bangladesh." *Journal of Development Economics*, 110: 13-21.

**Meyer, Bruce D. and Dan T. Rosenbaum.** 2001. "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers." *Quarterly Journal of Economics*, 116(3): 1063-1114.

**Meyer, Bruce D., Derek Wu, and Brian Curran.** 2020. "Does Geographically Adjusting Poverty Thresholds Improve Poverty Measurement and Program Targeting?" Working Paper.

**Puhani, Patrick A.** 2012. "The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear 'Difference-in-Differences' Models." *Economics Letters*, 115(1): 85-87.

**Rommel, Jens, Vera Buttmann, Georg Liebig, Stephanie Schönwetter, and Valeria Svart-Gröger.** 2015. "Motivation Crowding Theory and Pro-Environmental Behavior: Experimental Evidence." *Economics Letters*, 129: 42-44.

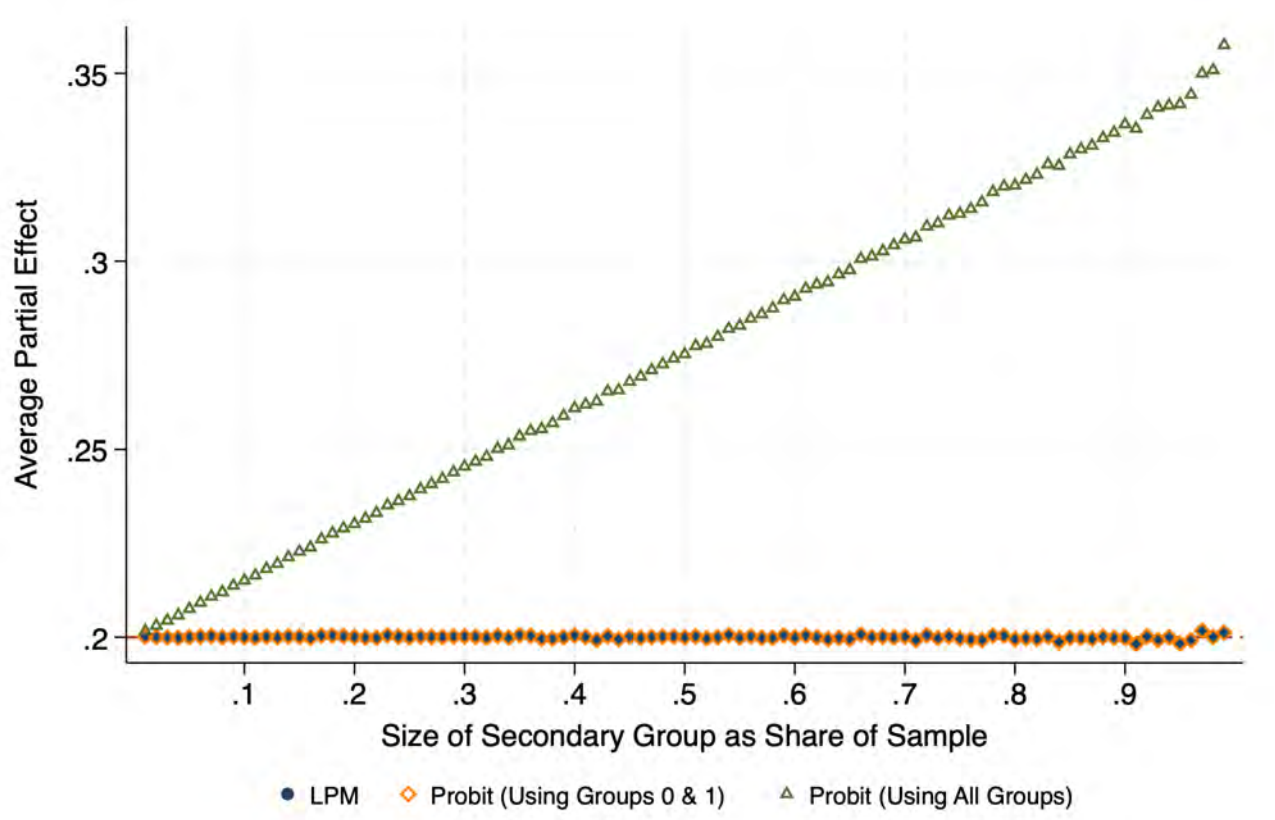
**Sequeira, Sandra.** 2016. "Corruption, Trade Costs, and Gains from Tariff Liberalization: Evidence from Southern Africa." *American Economic Review*, 106(10): 3029-3063.

**Wooldridge, Jeffrey M.** 2020. *Introductory Econometrics: A Modern Approach, 7th Edition*. Cengage Learning.

**Wooldridge, Jeffrey M.** 2023. "Simple Approaches to Nonlinear Difference-in-Differences with Panel Data." *Econometrics Journal*, 26(3): C31-C66.

# Figures and Tables

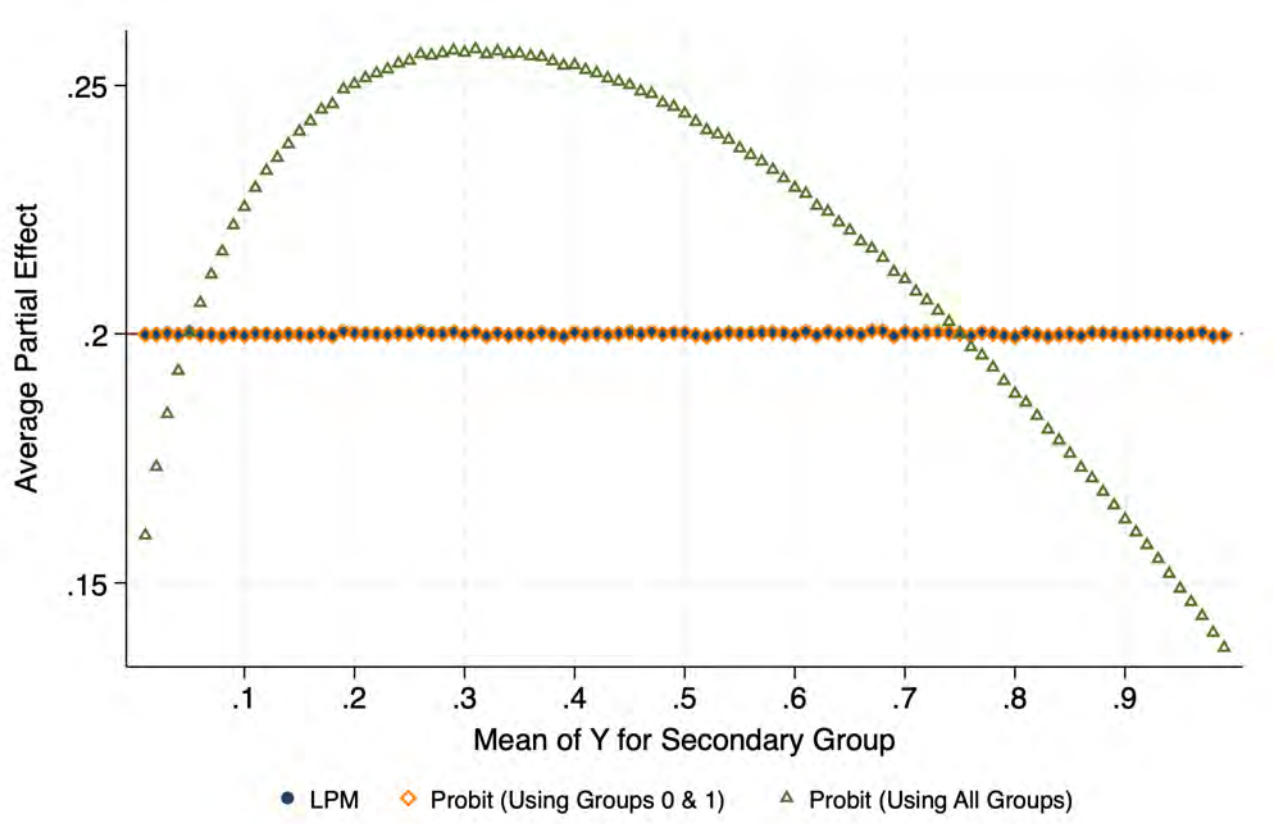
**Figure 1a. Bias in Average Partial Effect as a Function of Secondary Group Share of Sample (Differences-in-Means)**




Parameters: $\bar{y}_{x=0} = 0.05$, $\bar{y}_{x=1} = 0.25$, $\bar{y}_{x=2} = 0.45$

Notes: This figure shows estimates of the average discrete difference in $y$ due to being in Group 1 relative to Group 0, estimating it using three different specifications: (1) the basic LPM run over the entire sample, (2) a probit model where the APE is calculated over only Groups 0 and 1, and (3) a probit model where the APE is calculated over the entire sample. We use Monte Carlo simulations, averaging over 1,000 replications each with sample size 10,000, and set $\bar{y}_{x=0} = 0.05$, $\bar{y}_{x=1} = 0.25$, $\bar{y}_{x=2} = 0.45$. We vary $p_2$ from 0.01 to 0.99 (in increments of 0.01) and always fix $p_0 = p_1$.

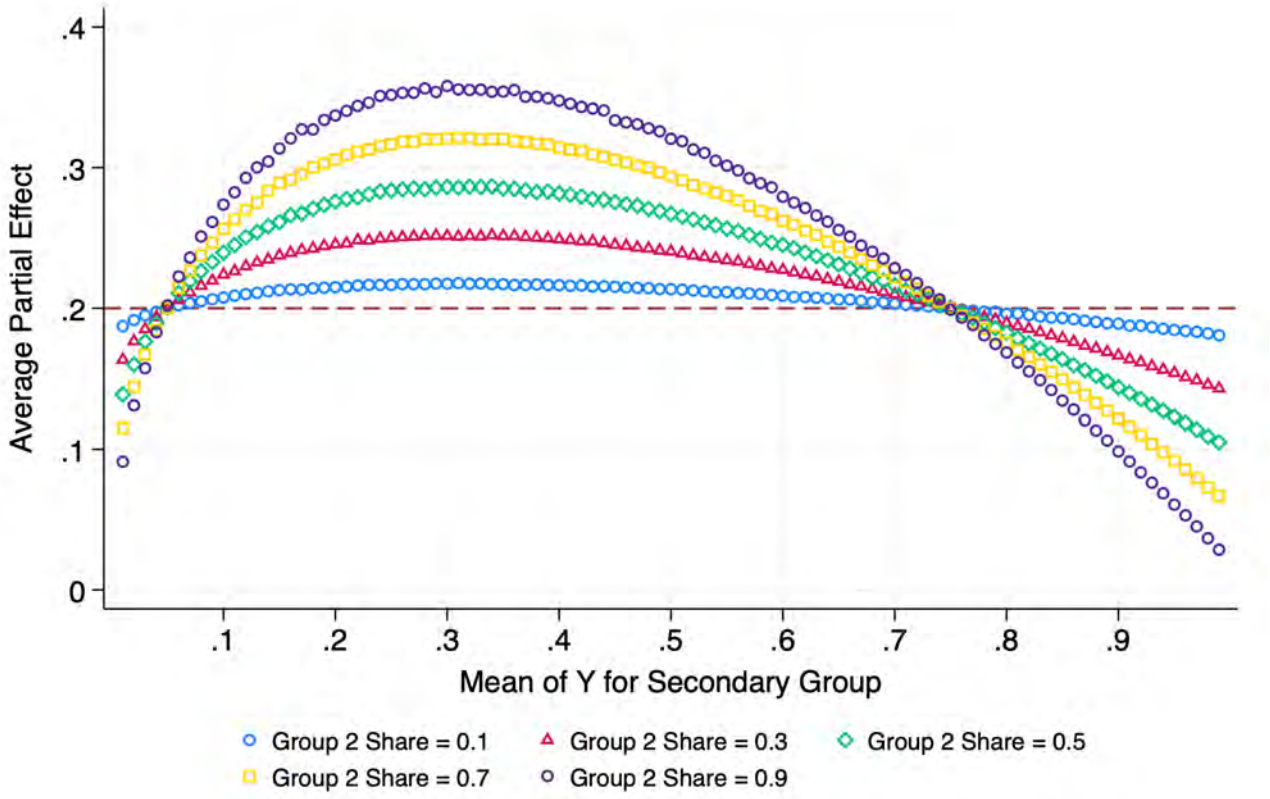
**Figure 1b. Bias in Average Partial Effect as a Function of Outcome Mean in Secondary Group (Differences-in-Means)**




Parameters: $\bar{y}_{x=0} = 0.05$, $\bar{y}_{x=1} = 0.25$

Notes: This figure shows estimates of the average discrete difference in $y$ due to being in Group 1 relative to Group 0, estimating it using three different specifications: (1) the basic LPM run over the entire sample, (2) a probit model where the APE is calculated over only Groups 0 and 1, and (3) a probit model where the APE is calculated over the entire sample. We use Monte Carlo simulations, averaging over 1,000 replications each with sample size 10,000, and set $\bar{y}_{x=0} = 0.05$ and $\bar{y}_{x=1} = 0.25$. We also equate the relative sizes of each group at $p_0 = p_1 = p_2 = 1/3$. We vary $\bar{y}_{x=2}$ from 0.01 to 0.99 (in increments of 0.01).

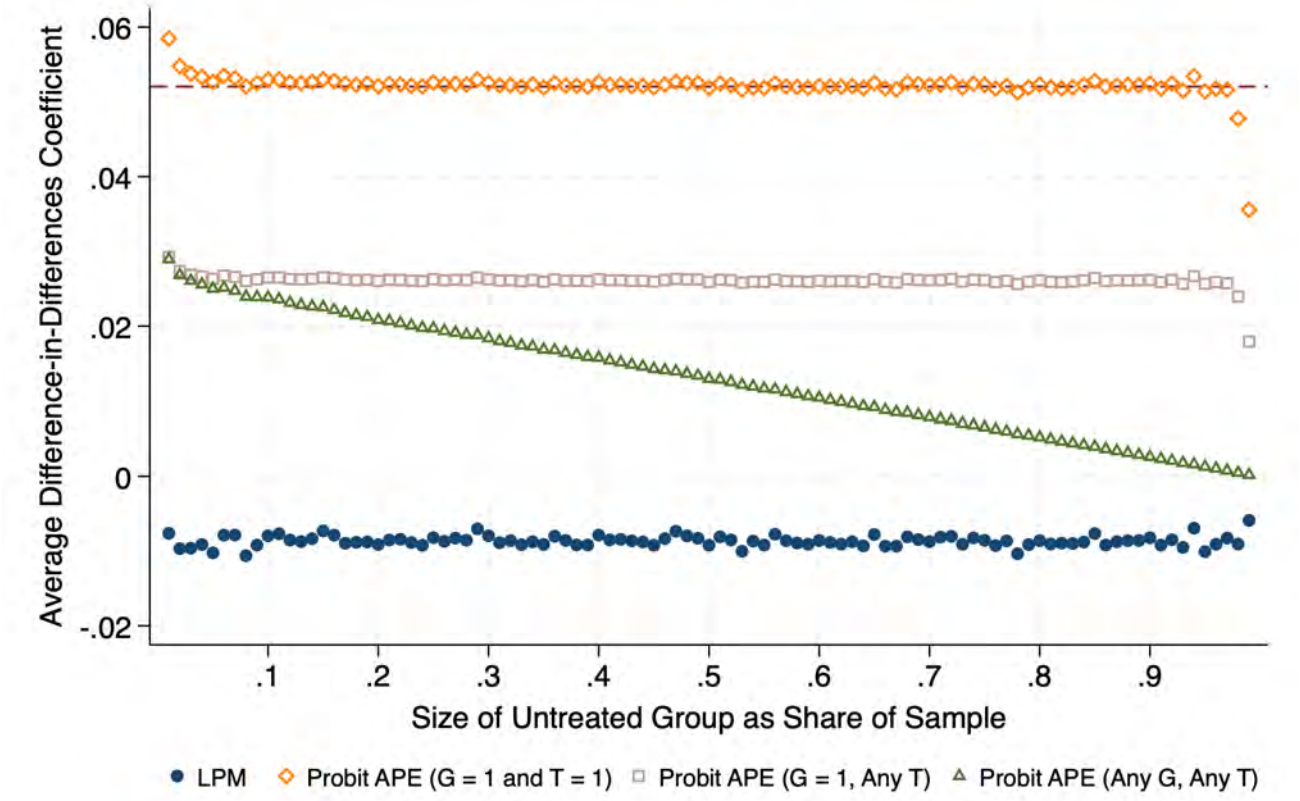
**Figure 1c. Bias in Average Partial Effect as a Function of Outcome Mean and Sample Share of Secondary Group (Differences-in-Means)**




Parameters: $\bar{y}_{x=0} = 0.05$, $\bar{y}_{x=1} = 0.25$

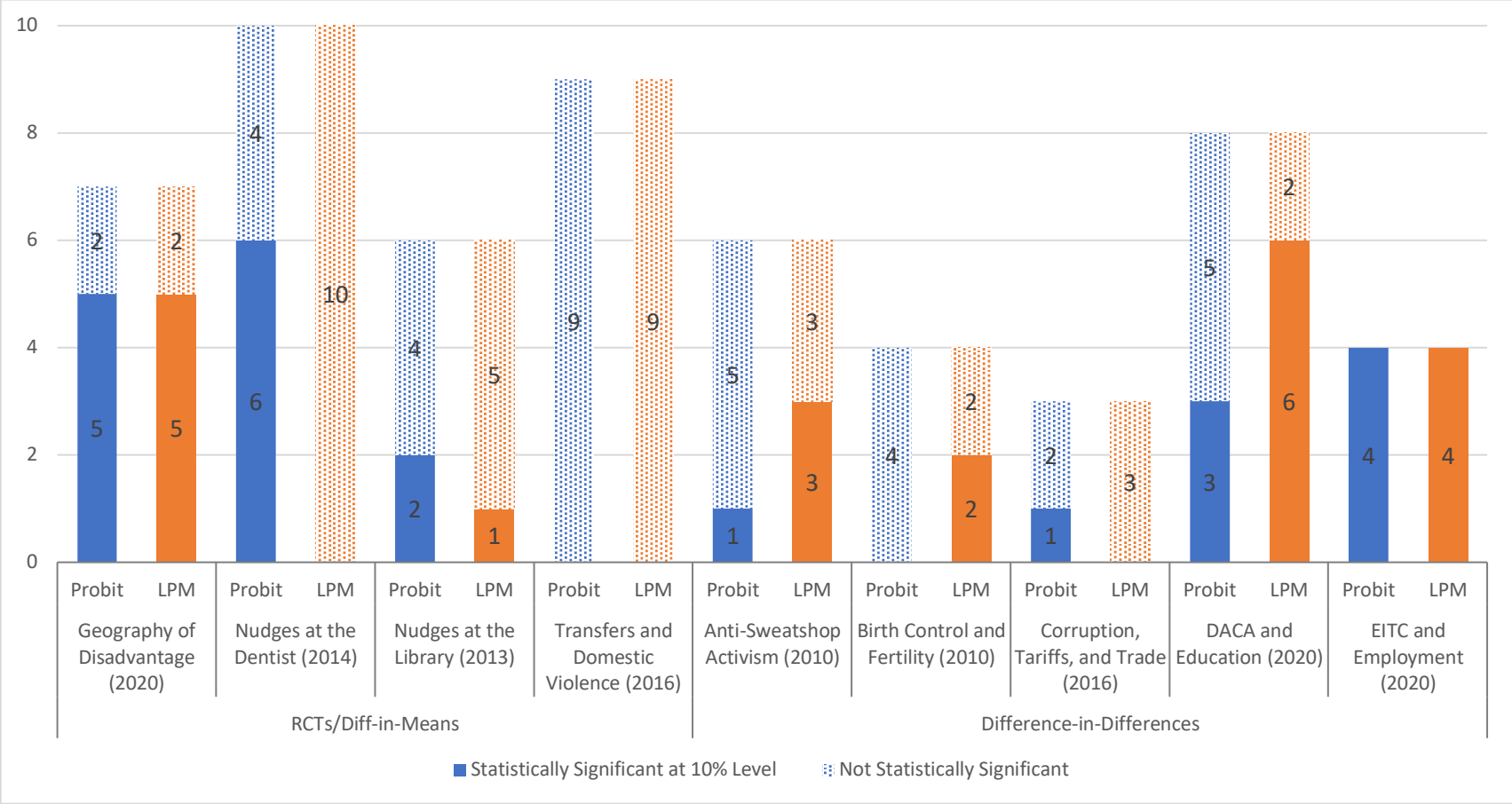
Notes: This figure shows estimates of the average discrete difference in $y$ due to being in Group 1 relative to Group 0, estimating it using a probit model where the APE is calculated over the entire sample. We use Monte Carlo simulations, averaging over 1,000 replications each with sample size 10,000, and set $\bar{y}_{x=0} = 0.05$ and $\bar{y}_{x=1} = 0.25$. We vary $p_2$ from 0.1 to 0.9 (in increments of 0.2) and always fix $p_0 = p_1$, and we vary $\bar{y}_{x=2}$ from 0.01 to 0.99 (in increments of 0.01).

**Figure 2. Bias in Average Partial Effect as a Function of the Untreated Group Share of Sample (Difference-in-Differences)**



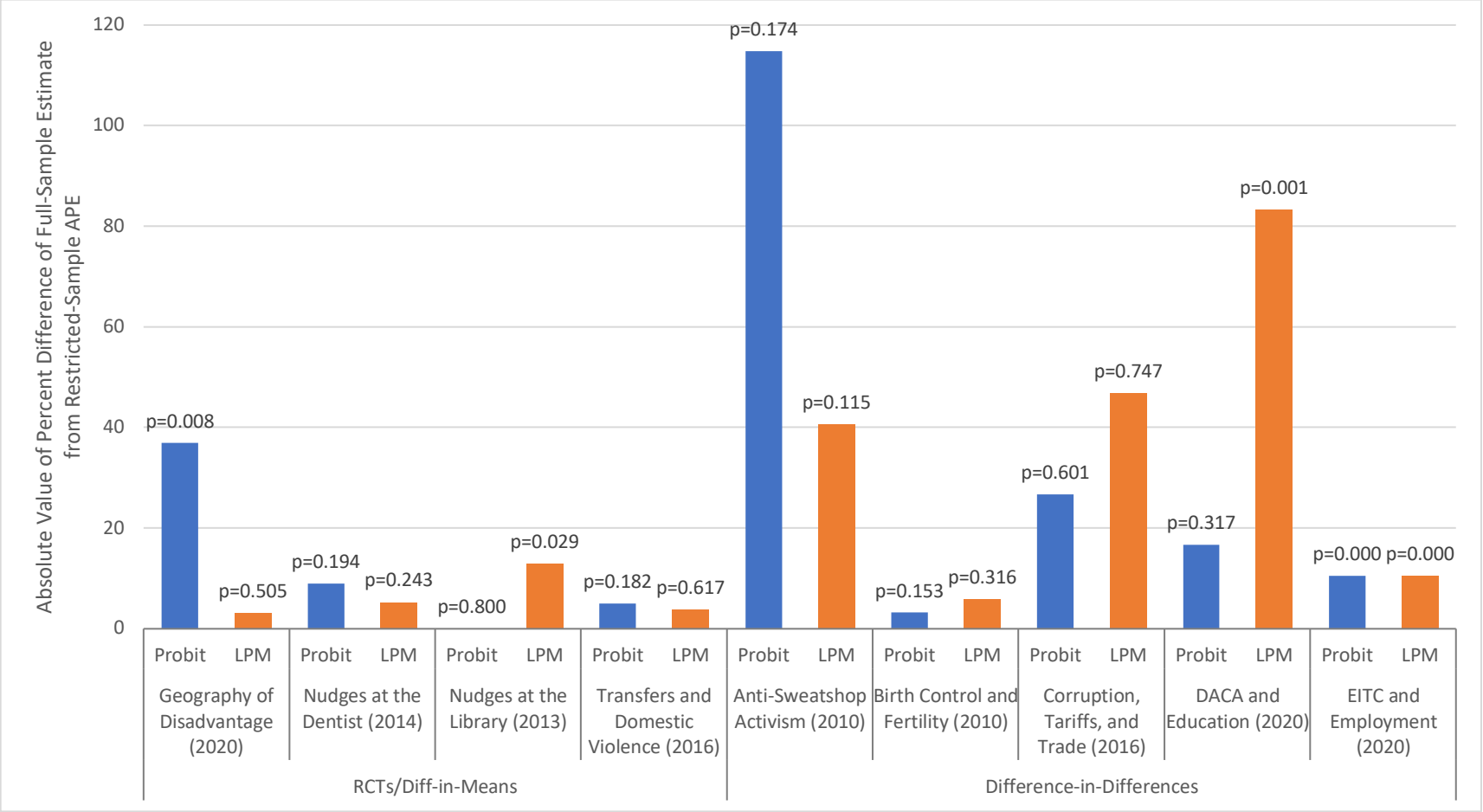

Notes: This figure shows difference-in-differences estimates of being in the treated group relative to the untreated group before and after treatment, using a panel data sample with two groups and two time periods. We estimate the effect of the interaction using four different specifications: (1) the standard LPM estimated over the entire sample, (2) a probit model where the APE is calculated over only treated observations during the post-treatment period, (3) a probit model where the APE is calculated over treated observations in both time periods, and (4) a probit model where the APE is calculated over the entire sample (both groups and time periods). We use Monte Carlo simulations, averaging over 1,000 replications each with sample size 10,000, and assume that the data generating process follows a probit model with the following specification: $\Pr(y = 1 \mid G, T) = \Phi(0.1 + G + 0.3T + 0.5G*T)$. We vary the size of the untreated group (as a share of the overall sample) from 0.01 to 0.99 (in increments of 0.01).

**Figure 3. Counts of Total APEs and the Number Significantly Different from Restricted Sample APEs in Recent Studies**



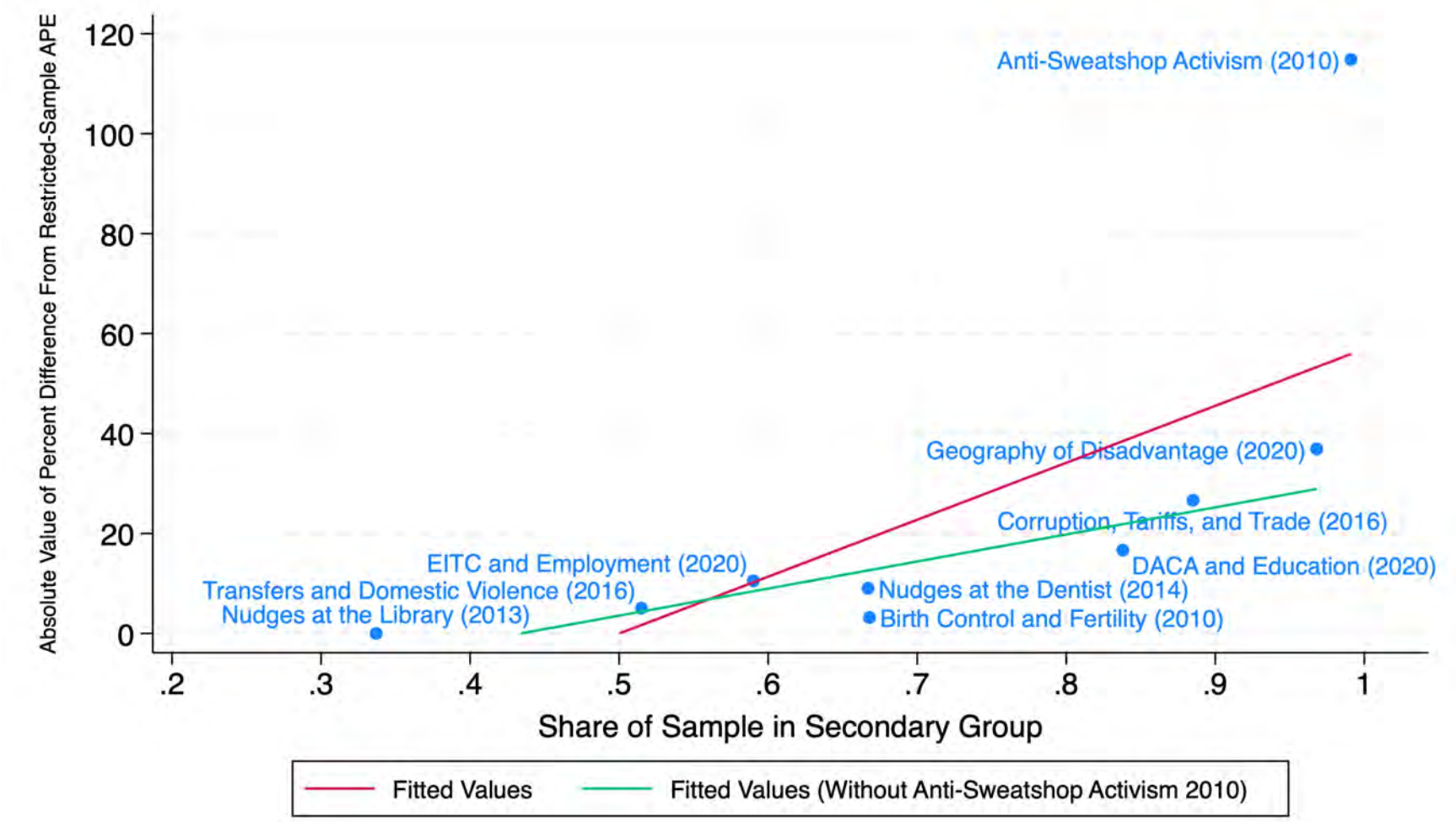

Notes: For each paper, we plot the number of outcomes in each paper for which the difference between the published estimate and the restricted-sample probit APE is statistically significant at the 10% level. The blue bars represent the difference between the full-sample probit APE and the restricted-sample probit APE, while the orange bars represent the difference between the LPM coefficient and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in just the post-treatment period. The standard errors of the differences are bootstrapped using 100 replications. Dark shading represents a difference that is statistically significant at the 10% level, while light shading represents a difference that is not statistically significant at the 10% level.

**Figure 4. Magnitude of Differences in Total APEs of Main Outcome**



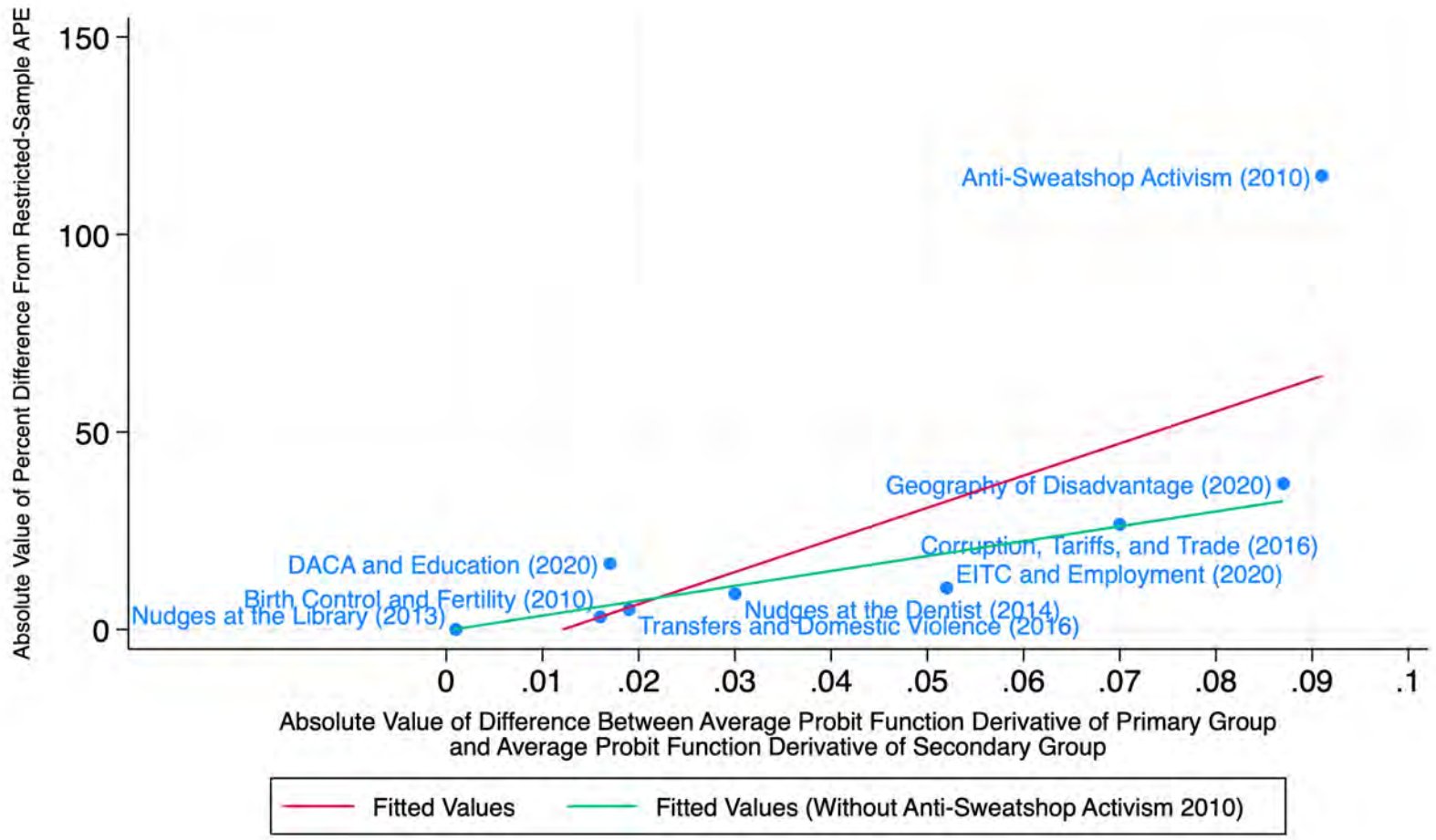

Notes: For each paper, we plot the absolute value of the percent difference between the published estimate and the restricted-sample probit APE. We also show the p-value testing the null hypothesis that the percent difference is equal to 0. The blue bars represent the difference between the full-sample probit APE and the restricted-sample probit APE, while the orange bars represent the difference between the LPM coefficient and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in just the post-treatment period. For each paper, we choose the main outcome based on what the authors focus discussion on most in the paper.

**Figure 5a. Relationship Between Difference in Probit APEs of Main Outcome and the Secondary Group Sample Share**




Notes: For each paper's main outcome, we plot on the y-axis the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE and on the x-axis the share of the sample in the secondary group. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in just the post-treatment period. The red line shows the line of best fit, which has slope 113.9 and standard error 43.9. The green line shows the line of best fit excluding the anti-sweatshop paper, which has slope 54.2 and standard error 11.3.

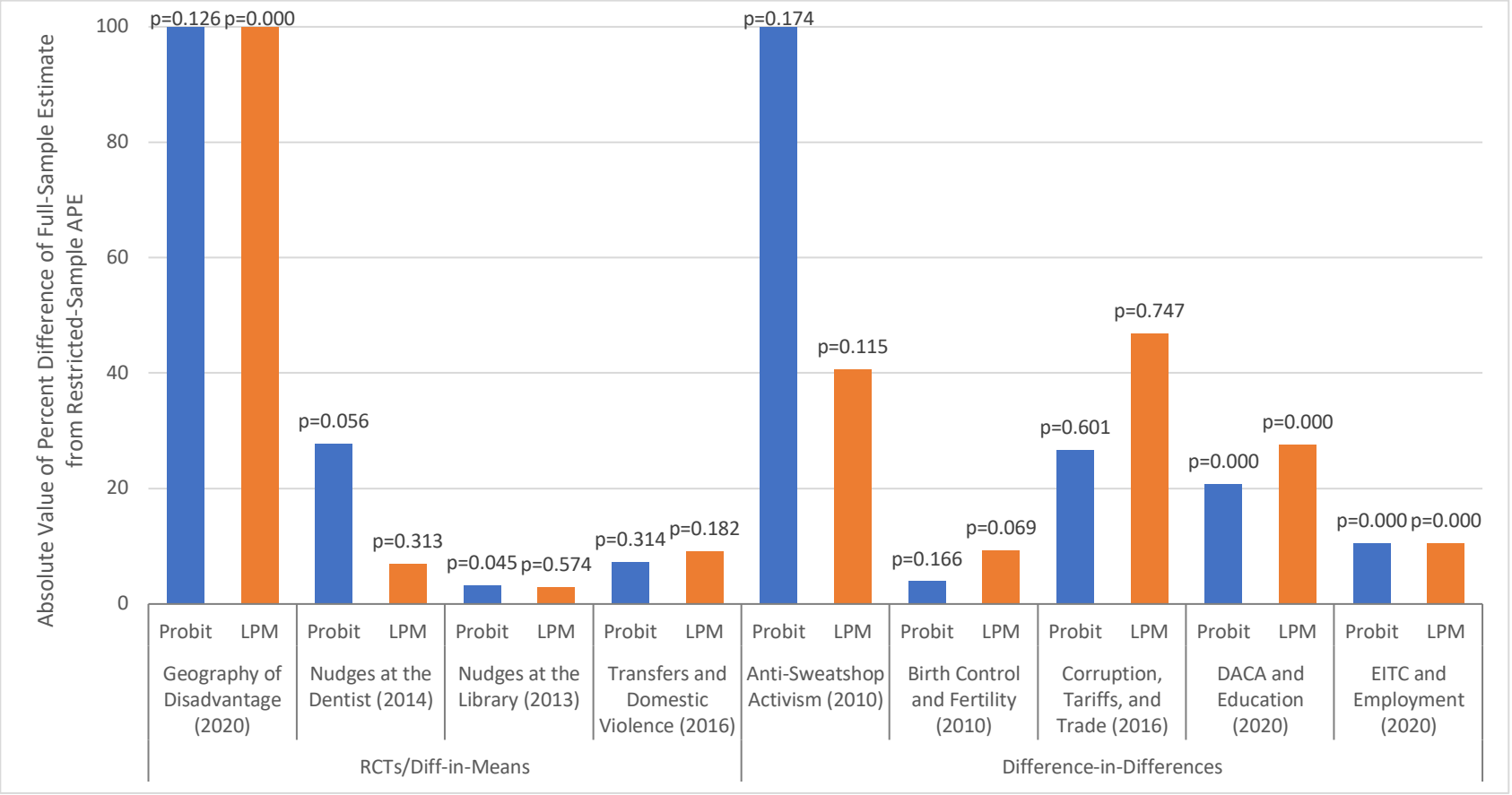
**Figure 5b. Relationship Between Difference in Probit APEs of Main Outcome and the Difference in Average Probit Function Derivatives**




Notes: On the y-axis we plot the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE. On the x-axis we plot the absolute value of the difference between the average probit function derivative of the primary group and the average probit function derivative of the secondary group. For each paper, we choose the main outcome. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in just the post-treatment period. The red line shows the line of best fit, which has slope 813.6 and standard error 262.3. The green line shows the line of best fit excluding the anti-sweatshop paper, which has slope 376.9 and standard error 76.5.

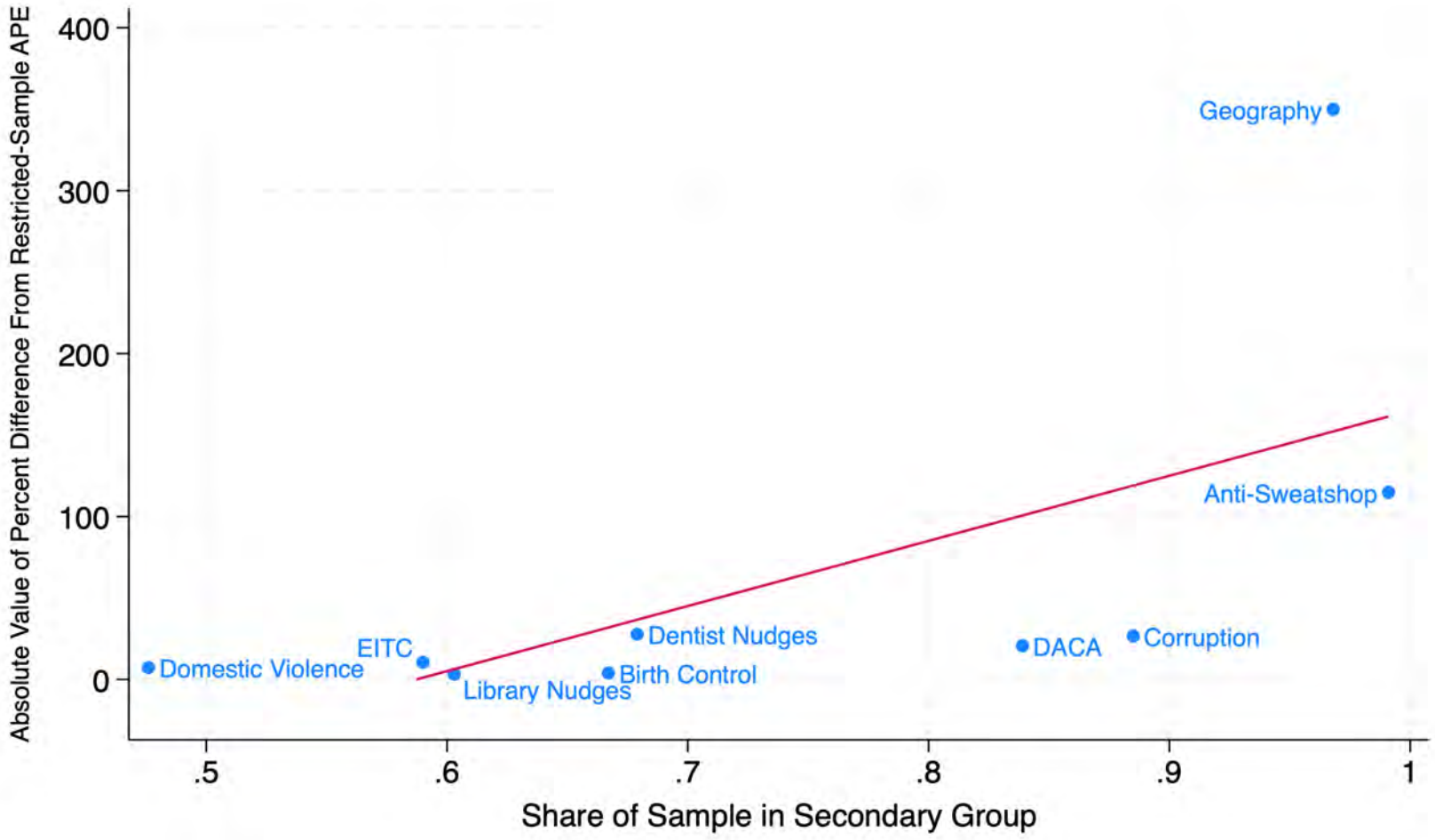
# Appendix Figures and Tables

### Appendix Figure A1. Magnitude of Differences in Total APEs of Outcome with Largest Difference



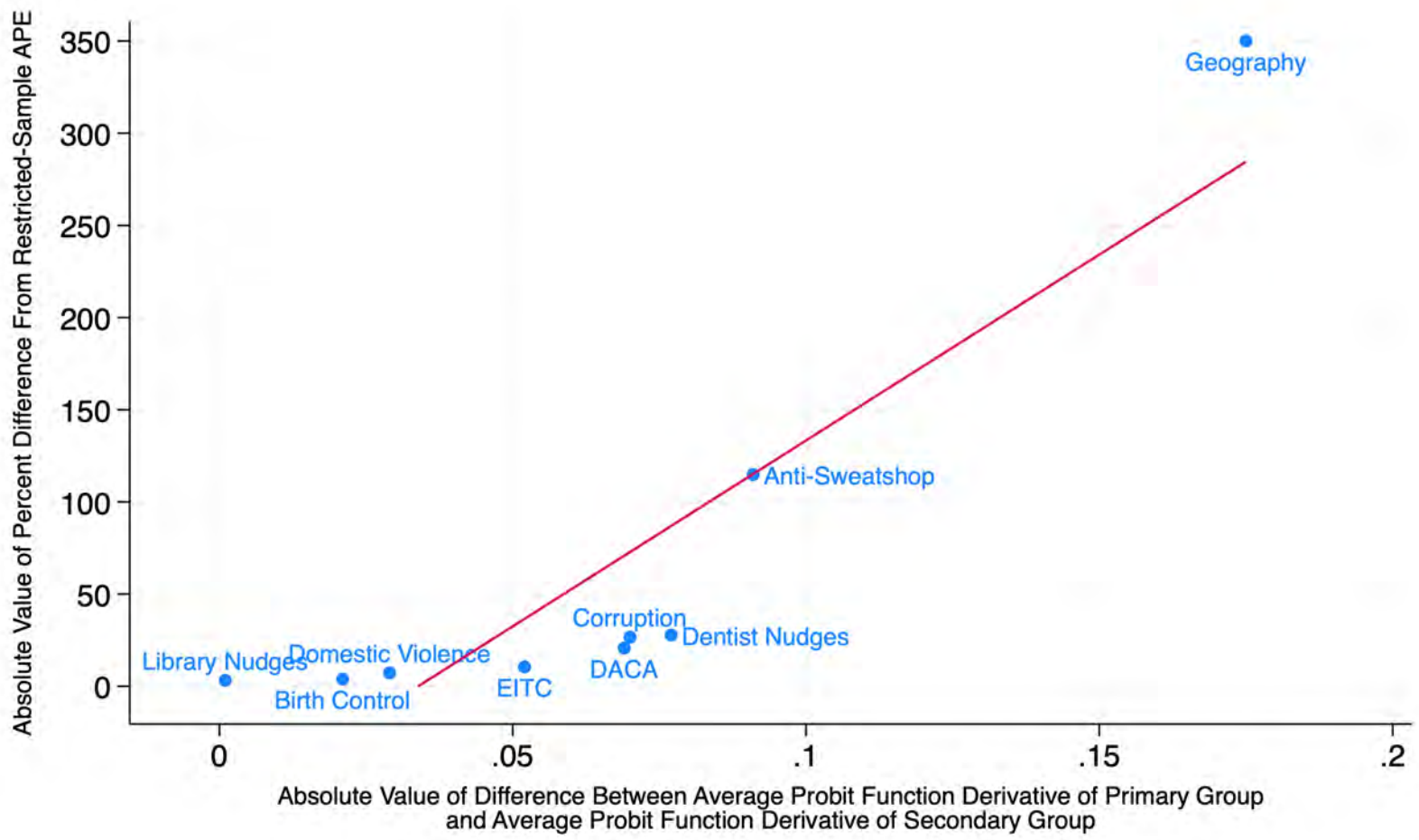

Notes: For each paper, we plot the absolute value of the percent difference between the published estimate and the restricted-sample probit APE. We also show the p-value testing the null hypothesis that the percent difference is equal to 0. The blue bars represent the difference between the full-sample probit APE and the restricted-sample probit APE, while the orange bars represent the difference between the LPM coefficient and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in just the post-treatment period. For each paper, we choose the outcome with the largest difference between full-sample probit APE and restricted-sample probit APE. The y-axis is capped at 100 percent.

**Appendix Figure A2. Relationship Between Difference in Probit APEs of Outcome with Largest Difference and the Secondary Group Sample Share**



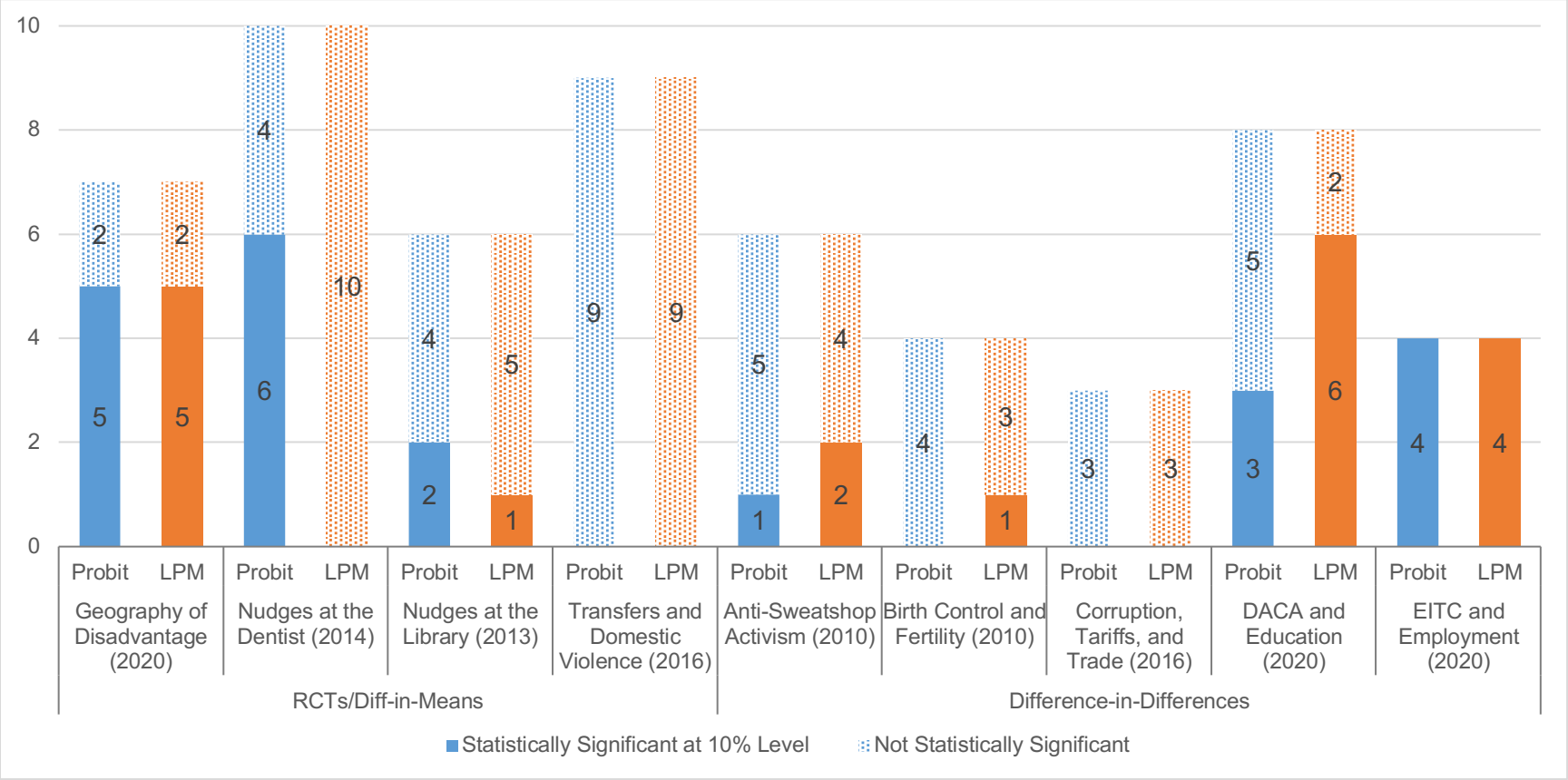

Notes: On the y-axis we plot the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE and on the x-axis we plot the share of the sample in the secondary group. For each paper, we choose the outcome with the largest difference between full-sample probit APE and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in just the post-treatment period. The red line shows the line of best fit, which has slope 399.9 and standard error 179.4.

**Appendix Figure A3. Relationship Between Difference in Probit APEs of Outcome with Largest Difference and the Difference in Average Probit Function Derivatives**



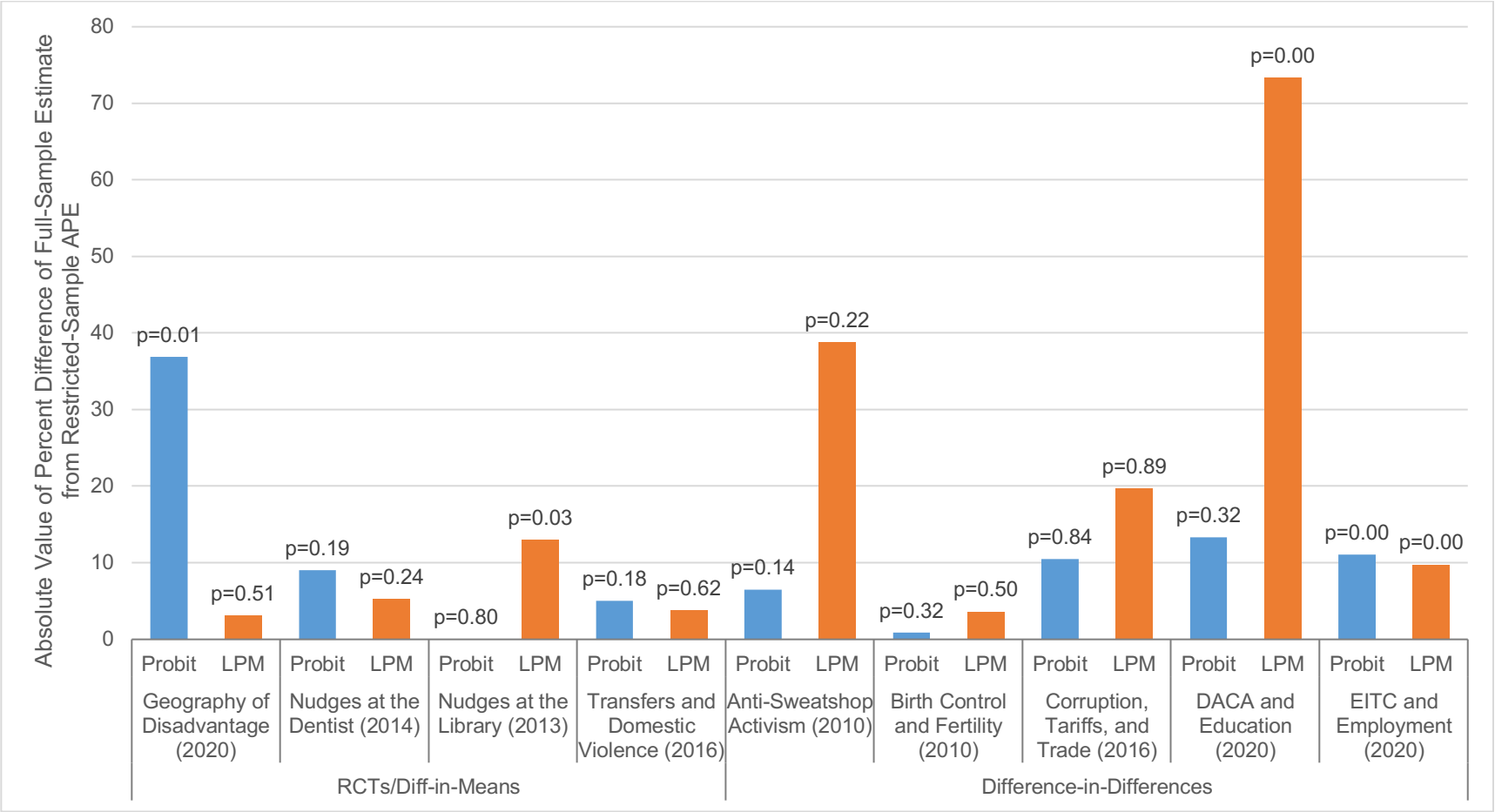

Notes: On the y-axis we plot the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE. On the x-axis we plot the absolute value of the difference between the average probit function derivative of the primary group and the average probit function derivative of the secondary group. For each paper, we choose the outcome with the largest difference between full-sample probit APE and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in just the post-treatment period. The red line shows the line of best fit, which has slope 2016.7 and standard error 365.5.

**Appendix Figure A4. Counts of Total APEs and the Number Significantly Different from Restricted Sample APEs in Recent Studies Using Alternate Restricted-Sample Probit APE**



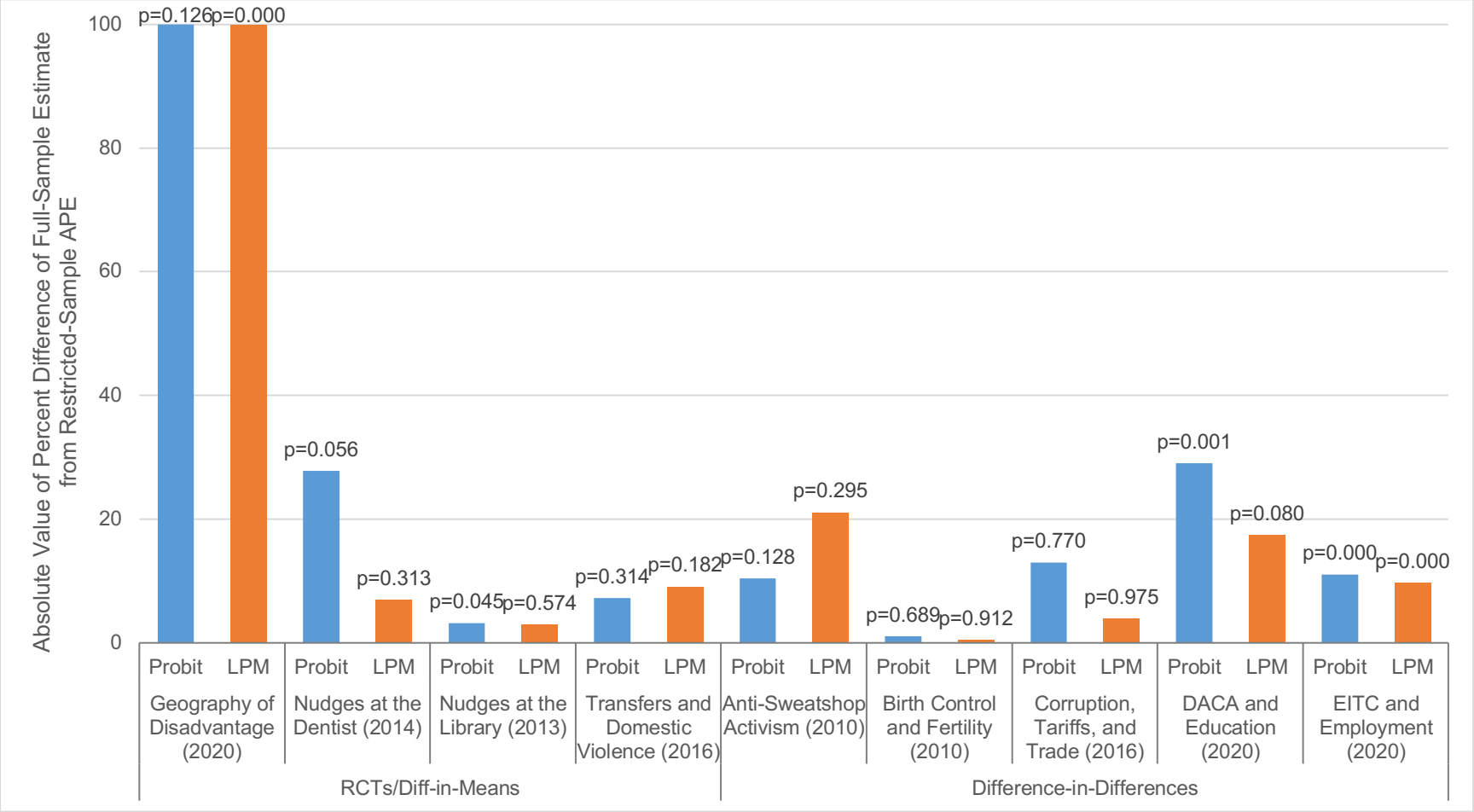

Notes: For each paper, we plot the number of outcomes for which the difference between the published estimate and the restricted-sample probit APE is statistically significant at the 10% level. The blue bars represent the difference between the full-sample probit APE and the restricted-sample probit APE, while the orange bars represent the difference between the LPM coefficient and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in both time periods. The standard errors of the differences are bootstrapped using 100 replications. Dark shading represents a difference that is statistically significant at the 10% level while light shading represents a difference that is not statistically significant at the 10% level.

**Appendix Figure A5. Magnitude of Differences in Total APEs of Main Outcome Using Alternate Restricted-Sample Probit APE**



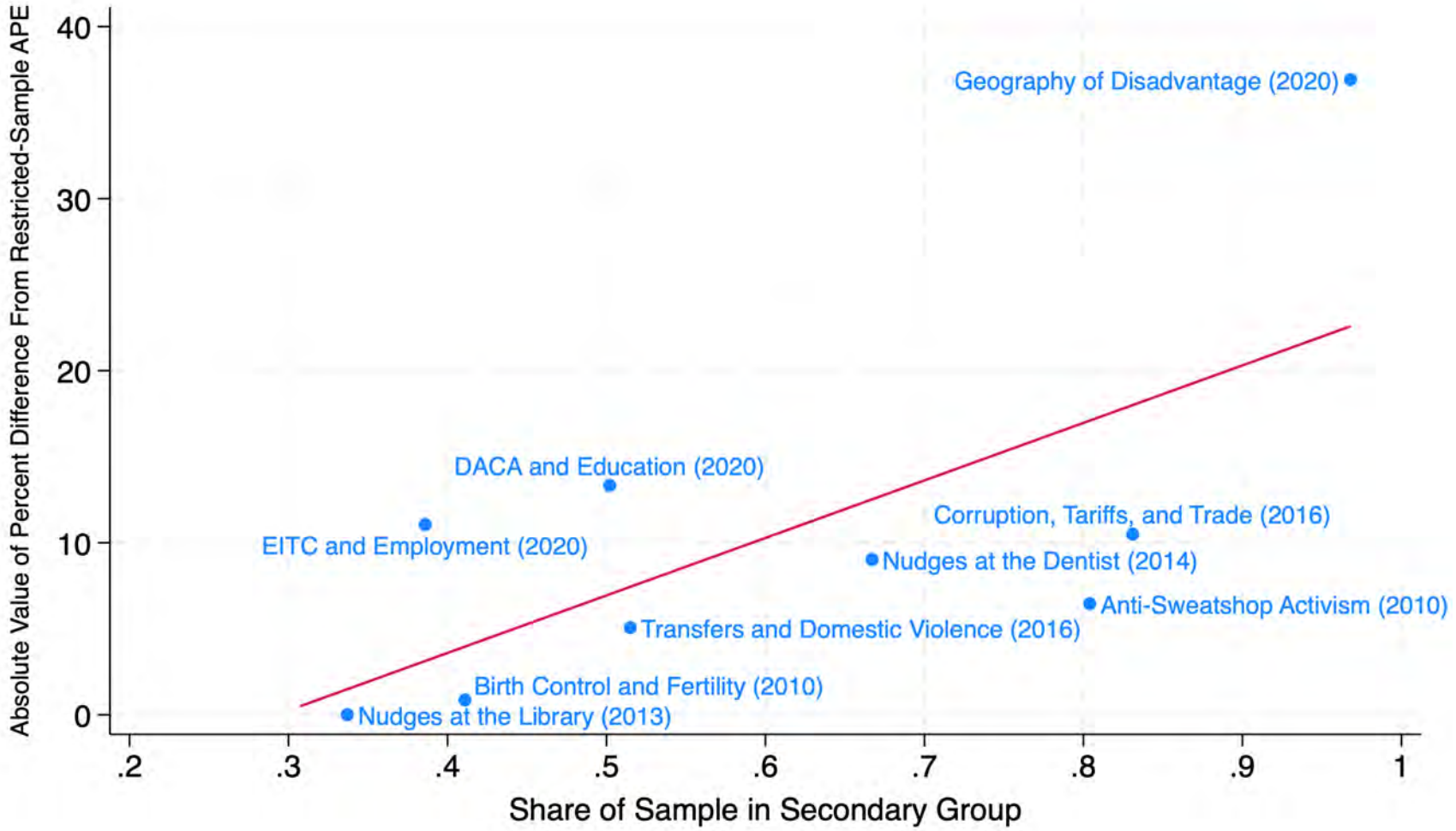

Notes: For each paper, we plot the absolute value of the percent difference between the published estimate and the restricted-sample probit APE. We also show the p-value testing the null hypothesis that the percent difference is equal to 0. The blue bars represent the difference between the full-sample probit APE and the restricted-sample probit APE, while the orange bars represent the difference between the LPM coefficient and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in both time periods. For each paper, we choose the main outcome based on what the authors focus on most in the paper.

**Appendix Figure A6. Magnitude of Differences in Total APEs of Outcome with Largest Difference Using Alternate Restricted-Sample Probit APE**



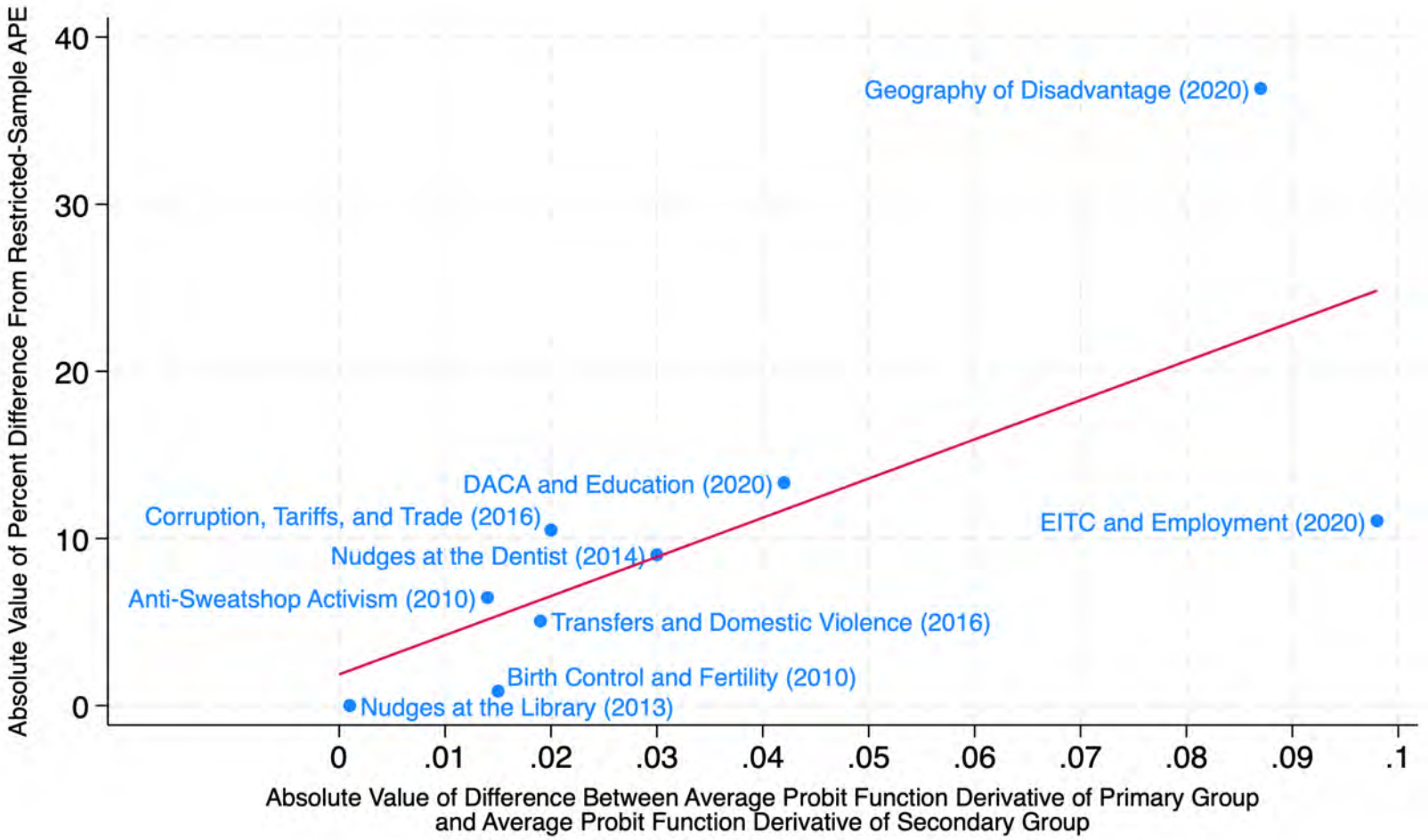

Notes: For each paper, we plot the absolute value of the percent difference between the published estimate and the restricted-sample probit APE. We also show the p-value testing the null hypothesis that the percent difference is equal to 0. The blue bars represent the difference between the full-sample probit APE and the restricted-sample probit APE while the orange bars represent the difference between the LPM coefficient and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in both time periods. For each paper, we choose the outcome with the largest difference between full-sample probit APE and restricted-sample probit APE. The y-axis is capped at 100 percent.

**Appendix Figure A7. Relationship Between Difference in Probit APEs of Main Outcome and the Secondary Group Sample Share Using Alternate Restricted-Sample Probit APE**



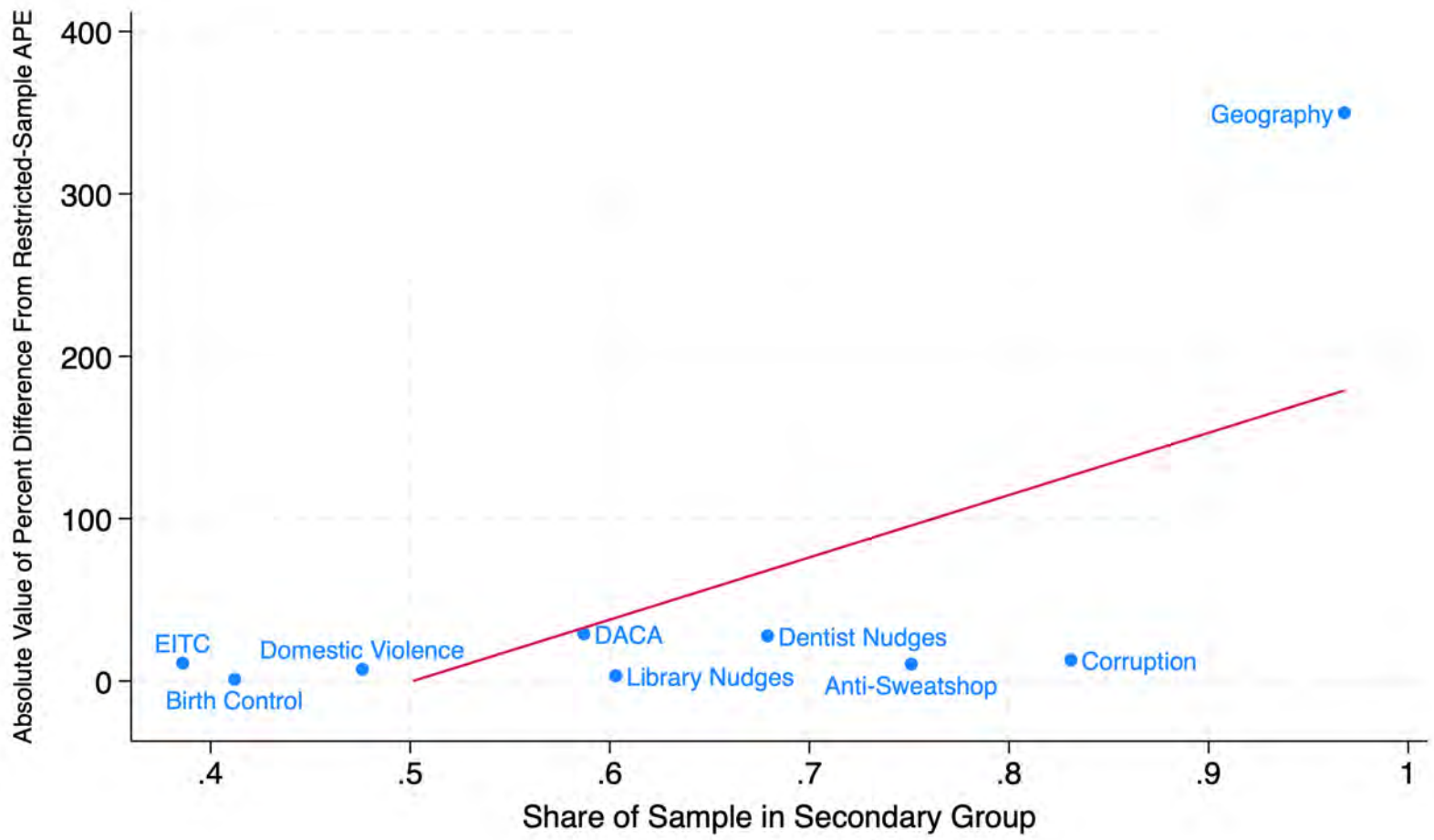

Notes: For each paper's main outcome, we plot on the y-axis the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE and on the x-axis the share of the sample in the secondary group. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in both time periods. The red line shows the line of best fit, which has slope 33.4 and standard error 13.4.

**Appendix Figure A8. Relationship Between Difference in Probit APEs of Main Outcome and the Difference in Average Probit Function Derivatives Using Alternate Restricted-Sample Probit APE**



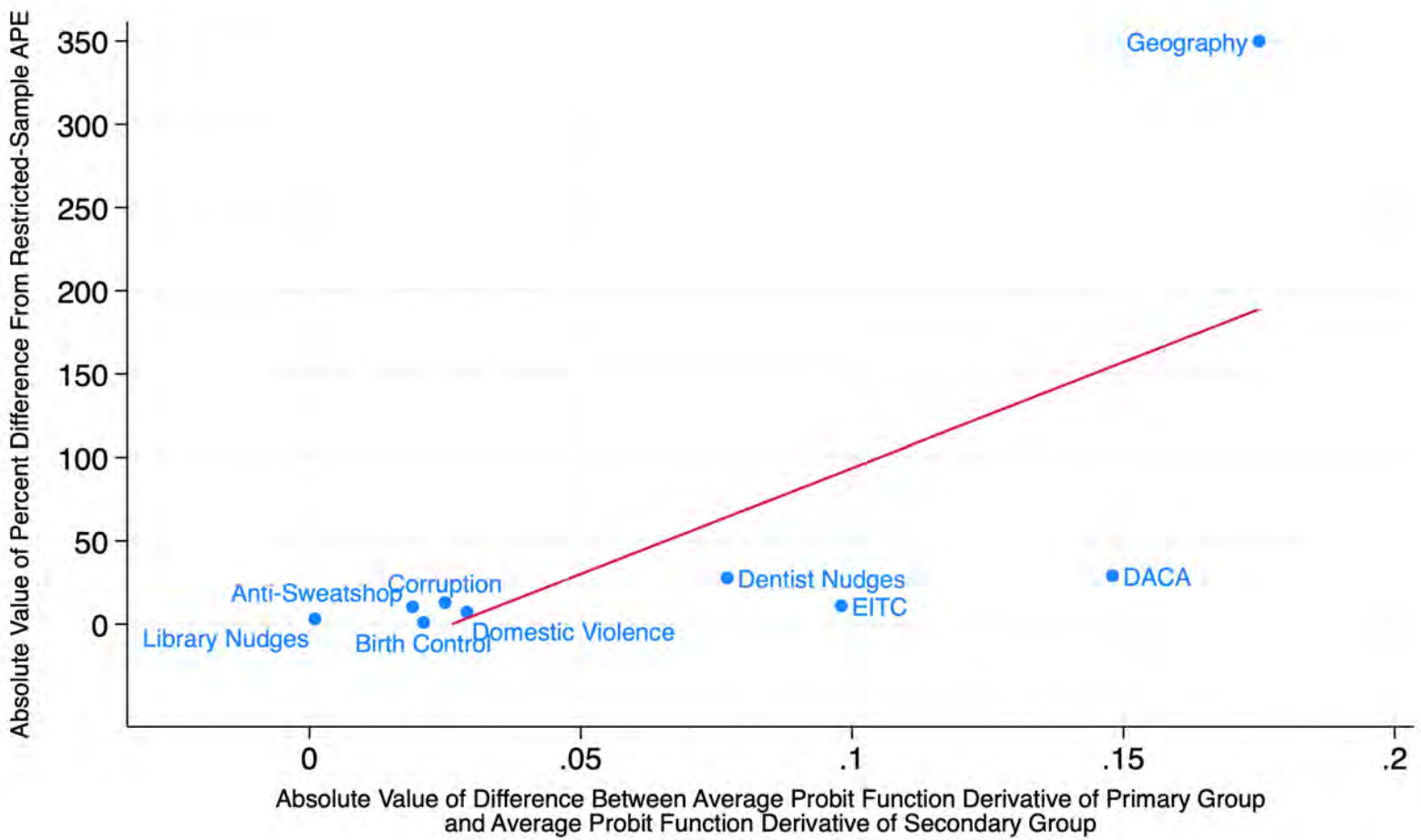

Notes: On the y-axis we plot the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE. On the x-axis we plot the absolute value of the difference between the average probit function derivative of the primary group and the average probit function derivative of the secondary group. For each paper, we choose the main outcome. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in both time periods. The red line shows the line of best fit, which has slope 234.4 and standard error 83.7.

**Appendix Figure A9. Relationship Between Difference in Probit APEs of Outcome with Largest Difference and the Secondary Group Sample Share Using Alternate Restricted-Sample Probit APE**




Notes: On the y-axis we plot the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE and on the x-axis we plot the share of the sample in the secondary group. For each paper, we choose the outcome with the largest difference between full-sample probit APE and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in both time periods. The red line shows the line of best fit, which has slope 383.6 and standard error 163.3. When excluding the geography paper, the line of best fit has slope 21.8 and standard error 25.1.

**Appendix Figure A10. Relationship Between Difference in Probit APEs of Outcome with Largest Difference and the Difference in Average Probit Function Derivatives Using Alternate Restricted-Sample Probit APE**




Notes: On the y-axis we plot the absolute value of the percent difference of the full-sample probit APE from the restricted-sample probit APE. On the x-axis we plot the absolute value of the difference between the average probit function derivative of the primary group and the average probit function derivative of the secondary group. For each paper, we choose the outcome with the largest difference between full-sample probit APE and the restricted-sample probit APE. For difference-in-differences papers, we calculate the restricted-sample probit APE over the treated group in both time periods. The red line shows the line of best fit which has slope 1270.0 (standard error 484.6).