**Finance and Economics Discussion Series**

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

# A Dummy Test of Identification in Models with Bunching

**Carolina Caetano, Gregorio Caetano, Hao Fe, and Eric Nielsen**

**2021-068**

Please cite this paper as:
Caetano, Carolina, Gregorio Caetano, Hao Fe, and Eric Nielsen (2021). "A Dummy Test of Identification in Models with Bunching," Finance and Economics Discussion Series 2021-068. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2021.068.

# A Dummy Test of Identification in Models with Bunching[*]

Carolina Caetano[†], Gregorio Caetano[†], Hao Fe[††], Eric Nielsen[†††]

September 2021

**Abstract**

We propose a simple test of the main identification assumption in models where the treatment variable takes multiple values and has bunching. The test consists of adding an indicator of the bunching point to the estimation model and testing whether the coefficient of this indicator is zero. Although similar in spirit to the test in Caetano (2015), the dummy test has important practical advantages: it is more powerful at detecting endogeneity, and it also detects violations of the functional form assumption. The test does not require exclusion restrictions and can be implemented in many approaches popular in empirical research, including linear, two-way fixed effects, and discrete choice models. We apply the test to the estimation of the effect of a mother's working hours on her child's skills in a panel data context (James-Burdumy 2005).

JEL Codes: C12, C21, C23, C24

## 1 Introduction

Caetano (2015) introduced the idea that confounders tend to be discontinuous at bunching points. This presents the opportunity to detect endogeneity by testing whether the outcome is discontinuous at a bunching point. There is a growing literature applying this test, see e.g. Rozenas et al. (2017), Erhardt (2017), Pang (2017), Bleemer (2018a), Bleemer (2018b), Ferreira et al. (2018), Lavetti and Schmutte (2018), Caetano and Maheshri (2018), De Vito et al. (2019), Caetano et al. (2019), Fe and Sanfelice (2020) and Caetano et al. (2021).

In this paper, we present a test similar in spirit to Caetano (2015)'s discontinuity test (henceforth CDT), but with some important advantages. A key advantage is that it is easy to apply: the test consists of simply adding an indicator variable (dummy) of a bunching point to the model and testing whether the parameter of the indicator is zero. The only requirement is a rank condition (essentially, that there is bunching), so it extends the applicability of CDT to cases where the treatment variable is discrete or mixed. In fact, some papers have used variations of this approach in an informal attempt to implement CDT (e.g. Caetano and Maheshri 2018, Ferreira et al. 2018 and Caetano et al. 2019). Yet, there has been no formal study of this test, which is one of the aims of this paper.

[*]†: University of Georgia. ††: San Diego State University † † †: Federal Reserve Board of Governors. We would like to thank Marinho Bertanha, Brantly Callaway, Bruno Ferman, Leonard Goff, Guido Imbens, Pedro Sant'Anna, Tymon Sloczynski, Firmin Tchatoka as well as seminar participants in several conferences and institutions for helpful conversations and comments. The analysis and conclusions set forth here are those of the authors and do not indicate concurrence by other members of the Federal Reserve Board research staff, the Board of Governors, or the Federal Reserve System.

Another advantage of the dummy test is that it tests all the main identification assumptions of the model at the same time, while CDT tests only exogeneity. This is desirable, since when presenting results, it is preferable to report diagnostic statistics about whether all identifying assumptions are valid, rather than only a subset. In linear models, the dummy test has power to detect violations of both the exogeneity and the linearity conditions. In models with heterogeneous treatment effects, it additionally detects correlated random effects. In linear difference-in-differences models that are estimated with two-way fixed effects regressions, the dummy test detects violations from the "strong parallel trends" assumption, which includes the standard parallel trends assumption plus the uncorrelated treatment effects assumption (e.g., de Chaisemartin and d'Haultfoeuille 2020, Callaway et al. 2021). In nonlinear models, including those estimated by nonlinear regression, GMM, and Maximum Likelihood, it detects violations from the model's specific exogeneity, functional form, and distributional assumptions.

The dummy test is also substantially more powerful than CDT even at detecting endogeneity. The lower power of CDT is due to more than its use of nonparametric estimators – it stems also from the split-sample nature of that test. In fact, we compare the dummy test to the parametric version of CDT, and the dummy test is more powerful.

The array of applications where the dummy test can be used is vast. Bunching is commonly found when the treatment variable is constrained to be above or below a threshold. Constraints can be natural (e.g. when the variable cannot be negative, such as the number of cigarettes smoked), or generated by laws/rules (e.g. minimum and maximum requirements, such as minimum schooling). Bunching is also frequently found at interior points, for example due to changes in policies at known thresholds (e.g. bunching at kinks in the US tax schedule). Extensive lists of examples can be found in Caetano (2015) and Caetano et al. (2020) as well as in the public finance bunching literature (see, e.g. Kleven (2016) and Bertanha et al. (2021)). In discrete choice models, there are often product characteristics that are bunched at zero, such as the number of previous purchases of cars of a given brand (Train and Winston 2007), the quantity in foodstuffs of sugar, fat, gluten, carbohydrates, and salt (Harding and Lovenheim 2017), crime in a neighborhood (Caetano and Maheshri 2018), the number of venues of a given type in a neighborhood (e.g. cafes, stores, parks, see Caetano and Maheshri 2019), and the racial composition of schools and neighborhoods (Caetano and Maheshri 2021). Notably, the dummy test can also be used to assess the validity of some popular selection-on-unobservables strategies such as difference-in-differences approaches using multi-way fixed effects. Examples of empirical papers using such strategies where the treatment variable takes multiple values and has bunching include Nunn (2008), Anderson and Sallee (2011), Forman et al. (2012), Imberman et al. (2012), and Dube and Vargas (2013), among many others.

We apply the dummy test to study the effect of maternal working hours on the skills of the child. In this literature all models are linear, so it is important to test all the main identification assumptions, not only endogeneity. In a panel setting (James-Burdumy 2005), we find evidence that year fixed effects together with a detailed list of control variables are not sufficient to identify the effect of interest, but family and year fixed effects with the same list of controls are, provided the panel is short enough. With a longer panel, the strong parallel trends assumption becomes invalid and the two-way fixed effects strategy is rejected, highlighting the fragility of this strategy in this context, and the importance of using tests such as the one we propose to guide the empirical approach.

Large parts of the empirical research in social and behavioral sciences relies on observational data and non-experimental identification strategies. This test contributes to a growing list of useful tools for sensitivity analyses in models with multi-valued treatments when experimental or quasi-experimental variation is not readily available or may be imperfect (e.g. Altonji et al. 2005, McCrary 2008, Oster 2019, de Chaisemartin and d'Haultfoeuille 2020, Callaway et al. 2021, D'Haultfoeuille et al. 2021). Since the dummy test does not require exclusion restrictions or special data structures, it can be used in the early stages of research as a diagnostic test to assess whether a different identification strategy should be used (perhaps necessitating longitudinal data, instrumental variables, or different identification assumptions such as those explored in some of the papers cited above).

The remainder of the paper is as follows. In Section 2, we formalize the test in the linear case, and discuss its size. In Section 3, we study the power of the test. In Section 4, we compare the dummy test with CDT, and we also discuss the results of a Monte Carlo experiment comparing the tests. We present our application in Section 5. In Section 6, we show how the test can be applied to nonlinear models, and we conclude in Section 7. The Appendix contains proofs and details, as well as various extensions including a section on how interactions of the dummy with controls, and multiple bunching points, can be used to increase power.

## 2 Test Statistic and Size

For simplicity, we focus first on linear models (including heterogeneous treatment effect and difference-in-differences models). However, the ideas translate well into nonlinear models, as detailed in Section 6. We want to identify $\beta$ in the following equation:

$$Y = \beta X + \varepsilon, \tag{1}$$

where $Y$ is the outcome variable, $X$ is the explanatory variable of interest (a scalar), and $\varepsilon$ is the remainder (so this equation is without loss of generality).[1]

Because we are concerned that $X$ and $\varepsilon$ are correlated, we may want to use controls. Let the vector $Z$ include a constant and any controls we may wish to include. To estimate $\beta$, we intend to run an OLS regression of $Y$ on $X$ and $Z$. For $\hat{\beta}$ obtained from this regression to be consistent, one needs to assume

**Assumption 1.** $\mathbb{E}[\varepsilon|X, Z] = \mathbb{E}[\varepsilon|Z] = Z'\lambda$.

This assumption implies $Cov(X, \varepsilon|Z) = 0$. It states that any confounder of $X$ can be absorbed by a linear combination of the elements of $Z$. $Z$ may include fixed effects, lagged measures of $Y$ and $X$, proxy variables (including generated regressors) and any other observed control variables. This setting is therefore rather general. In Appendix C.1, we show that this setting includes heterogeneous treatment effects models and difference-in-differences models that are estimated with multi-way fixed effects. There, we also use the potential outcomes notation, which may be more familiar to some readers.

Let $W = (X, Z')'$, and assume that

**Assumption 2.** $\mathbb{E}[(W', \mathbf{1}(X = 0))(W', \mathbf{1}(X = 0))']$ *is invertible.*

[1] Note that $\varepsilon$ does not need to be centered around zero, hence why a constant is not explicitly included in this equation.

Because $Z$ includes a constant, this rank condition implicitly requires that $0 < \mathbb{P}(X = 0) < 1$, i.e., that $X$ varies and has a bunching point at $X = 0$. Note that there is no restriction in the support beyond Assumption 2. In particular, the distribution of $X$ may be discrete or mixed.

We propose testing Assumption 1 by adding $\mathbf{1}(X = 0)$ to the regression of $Y$ on $X$ and $Z$, and testing whether the coefficient of $\mathbf{1}(X = 0)$ is equal to zero. To increase power, it may also be desirable to add interactions of $\mathbf{1}(X = 0)$ and functions of $Z$ instead. Also, if more than one bunching point is available, it may be advantageous to add more dummies. Here we focus on the simple case where a single dummy is added, and we provide details of these extensions in Appendix D.

Let the sample be $\{(Y_i, X_i, Z_i')'\}_{i=1}^n$, and define $\mathbf{y} = (Y_1, \ldots, Y_n)'$, $\mathbf{d} = (\mathbf{1}(X_1 = 0), \ldots, \mathbf{1}(X_n = 0))'$, and $\mathbf{w}$ the matrix with rows equal to $(X_i, Z_i')'$. For $n$ large enough, Assumption 2 guarantees that we can write the matrix inverses below. Define $M_{\mathbf{w}} = I - \mathbf{w}(\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'$, where $I$ is the $n \times n$ identity matrix. Then the coefficient of $\mathbf{d}$ in a regression of $\mathbf{y}$ onto $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{d}$ is

$$\hat{\theta} = (\mathbf{d}'M_{\mathbf{w}}\mathbf{d})^{-1}\mathbf{d}'M_{\mathbf{w}}\mathbf{y}.$$

The dummy test statistic is simply the $t$-statistic of the test that the coefficient of $\mathbf{1}(X = 0)$ is significant. Specifically, the test statistic is $\hat{\theta}/SE(\hat{\theta})$, where $SE(\hat{\theta})$ is the estimator of the standard deviation of $\hat{\theta}$, which will depend on the assumptions and method of estimation. Thus, at the $\alpha$ significance level, we reject the null hypothesis

$$H_0: \quad \text{Assumption 1 holds}$$

if $|\hat{\theta}/SE(\hat{\theta})| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\cdot 100$th quantile of the standard normal distribution.

Technically, the dummy test is identical to a specification test in which we add an additional term, $\mathbf{1}(X = 0)$, to the regression, and then test if the coefficient of this new term is equal to zero. Such tests are ubiquitous in practice. Establishing that the size is asymptotically correct is simply a matter of proving the convergence in distribution of the OLS estimator of the dummy coefficient in the augmented regression, and the consistency of the corresponding standard error estimator. These convergence results have been established for a host of combinations of data structures and assumptions about the data. Instead of choosing a specific structure and repeating such results here, we refer the reader directly to the relevant papers. For classical cases, see White (1980) for cross-sectional data with independent but not identically distributed observations, and see Arellano et al. (1987) for panel data with clustered errors. Subsequently, many papers have established the asymptotic behavior of the OLS coefficients and standard errors under variations in the data structure and relaxations of the assumptions of the model. For example, asymptotic results for OLS regressions with generated covariates are established in Newey and McFadden (1994) and Newey (1994). The literature on spatial and panel data is also rich in different specifications, cluster definitions, variance models and covariance estimation techniques for which asymptotic results have been established (e.g. Lee 2007, Bester et al. 2011, Bonhomme and Manresa 2015, Bester et al. 2016, de Chaisemartin and d'Haultfoeuille 2018, de Chaisemartin and d'Haultfoeuille 2020).

# 3 Test Power

In this section, we study the power of the test. Local power analyses and other determinants of power which depend on the variance of $\hat{\theta}$ follow trivially from the asymptotic results of the specific setting (e.g., see the papers cited at the end of the previous section). We focus instead on the magnitude of $\hat{\theta}$, which is the main determinant of power as the sample size increases. Specifically, we examine what $\hat{\theta}$ identifies when Assumption 1 does not hold.

We can write, without loss of generality,

$$\mathbb{E}[\varepsilon|X,Z] = \Gamma(X,Z) + \Delta(Z)\mathbf{1}(X=0), \tag{2}$$

where $\Gamma(X,Z)$ is continuous in $X$ at $X=0$ for all $Z$.[2] This equation categorizes violations from Assumption 1 as either continuous ($\Gamma(X,Z) \neq Z'\lambda$) or discontinuous at $X=0$ ($\Delta(Z) \neq 0$). In Appendix B, we develop a model in which the bunching at $X=0$ is generated by a constraint that $X$ cannot be negative. This example is rather general, and fits most applications where bunching is at one extreme of the support of $X$'s distribution. There, $\Gamma$ and $\Delta$ have a structural interpretation within the model.[3]

Let $\mathbf{\Gamma} = (\Gamma(X_1,Z_1),\ldots,\Gamma(X_n,Z_n))'$, $\mathbf{\Delta}_0 = (\Delta(Z_1)\mathbf{1}(X_1=0),\ldots,\Delta(Z_n)\mathbf{1}(X_n=0))'$, and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)'$. Then the estimated coefficient of $\mathbf{1}(X=0)$ is

$$\hat{\theta} = (\mathbf{d}'M_{\mathbf{w}}\mathbf{d})^{-1}\mathbf{d}'M_{\mathbf{w}}\mathbf{\Gamma} + (\mathbf{d}'M_{\mathbf{w}}\mathbf{d})^{-1}\mathbf{d}'M_{\mathbf{w}}\mathbf{\Delta}_0 + (\mathbf{d}'M_{\mathbf{w}}\mathbf{d})^{-1}\mathbf{d}'M_{\mathbf{w}}\boldsymbol{\epsilon}.$$

The last term of $\hat{\theta}$ is asymptotically negligible.[4] In Appendix A, we show that, under standard assumptions such as random sampling and the existence of moments,

$$\hat{\theta} \to_p \frac{\mathbb{E}[\Gamma(0,Z)|X=0] - \mathbf{\Gamma}_0^*}{1-\mathbf{d}^*} + \frac{\mathbb{E}[\Delta(Z)|X=0] - \mathbf{\Delta}_0^*}{1-\mathbf{d}^*}, \tag{3}$$

where, letting $m_{ZX} := p\lim_{n\to\infty} n^{-1}(\mathbf{z}'(I-\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\mathbf{z})$, $\mathbf{\Gamma}_0^* := \mathbb{E}[Z|X=0]'m_{ZX}^{-1}p\lim_{n\to\infty} n^{-1}\mathbf{z}'(I-\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\mathbf{\Gamma}$ is the predicted value of $\Gamma(X,Z)$ in a regression on $X$ and $Z$ at $X=0$. Analogously, $\mathbf{\Delta}_0^* := \mathbb{E}[Z|X=0]'m_{ZX}^{-1}\mathbb{E}[Z\Delta(Z)\mathbf{1}(X=0)]$ is the asymptotic limit of the predicted value of $\Delta(Z)\mathbf{1}(X=0)$ from a regression on $X$ and $Z$ at $X=0$.

The power is therefore dependent on two factors. The first factor, $\mathbb{E}[\Gamma(0,Z)|X=0] - \mathbf{\Gamma}_0^*$, depends on the continuous nonlinearities. If $\Gamma(X,Z) = \alpha X + Z'\lambda$, (i.e. there is no misspecification, but there is linear endogeneity, through $\alpha X$) then this term will be zero.[5] In every other case, this term will be

[2] Any function $f(X,Z)$ can be written without loss of generality as the sum of a continuous function in $X$ at $X=0$ and the discontinuity in $X$ at $X=0$.

[3] In particular, in that model, (a) if there is endogeneity, then $\Delta(Z) \neq 0$, and thus there is a discontinuity in the unobservables generated by the constraint; (b) a discontinuity in the treatment function will also affect $\Delta(Z)$; and (c) $\Gamma$ is affected by continuous nonlinearities both in the treatment function and, if there is endogeneity, in the effect of the confounder on the outcome.

[4] This holds for any data structure, method, and choice of $Z$ under the assumptions that guarantee the consistency of that method. For example, if the $\epsilon_i$ are independent but not identically distributed and we use the Eicker-White standard errors, then the negligibility of the last term follows by White (1980)'s Theorem 1, under Assumptions 2-4 in that paper (replacing White (1980)'s $X_i$ and $\varepsilon_i$ with $(W_i', \mathbf{1}(X_i=0))'$ and $\epsilon_i$, respectively, and noting that Assumption 1 in that paper holds by (2)).

[5] Nevertheless, since $\alpha$ and $\Delta(Z)$ tend to be determined by the same factors, in this case the endogeneity will usually

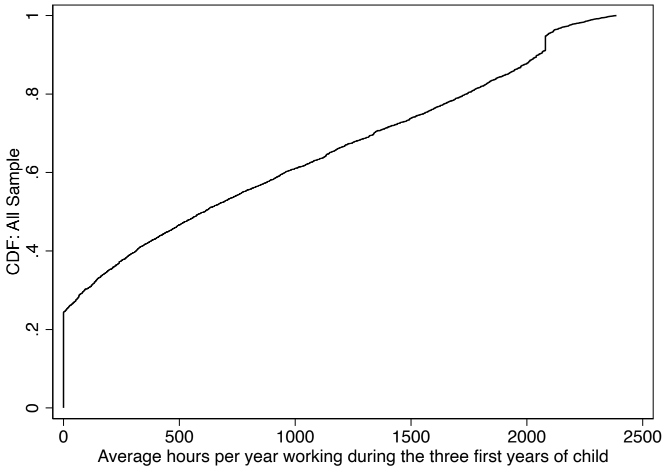
different from zero. As we show in Appendix B, nonlinearities in $\Gamma$ may appear because the treatment function is misspecified, or because there is nonlinear endogeneity. In particular, this term detects action of confounders that only affect the outcome for values of $X$ away from the bunching point, since this tends to generate continuous nonlinearities in $\mathbb{E}[Y|X,Z]$.[6]

The second factor, $\mathbb{E}[\Delta(Z)|X=0] - \boldsymbol{\Delta}_0^*$, depends on the size of the discontinuities. Such discontinuities appear if the treatment function is discontinuous at $X=0$ or if there is endogeneity and the unobservables are discontinuous at the bunching point. As argued by Caetano (2015), and shown in Appendix B for a constrained choice model, discontinuities in unobservables are ubiquitous.

The term $\mathbf{d}^* := \mathbb{E}[Z|X=0]' m_{ZX}^{-1} \mathbb{E}[Z\mathbf{1}(X=0)]$ in the denominator is the asymptotic limit of the predicted value of $\mathbf{1}(X=0)$ from a regression on $X$ and $Z$ at $X=0$. In Appendix A, we show that $0 < 1 - \mathbf{d}^* \leq 1$ and thus that the difference in the numerators are further magnified.

To illustrate the sources of power of the dummy test, consider our application. We are interested in the effects of the number of hours a mother works during the first three years of the child on the child's skills. There is a pronounced bunching of 25% of mothers at zero hours, which can be seen in Figure 1.

Figure 1: Evidence of Bunching in Maternal Working Hours



Note: The figure shows the empirical cumulative density function of the mother's average hours working per year during the first three years of the child's life for our full sample ($N = 3,383$). Source: NLSY79. See Section 5 for details about the application.

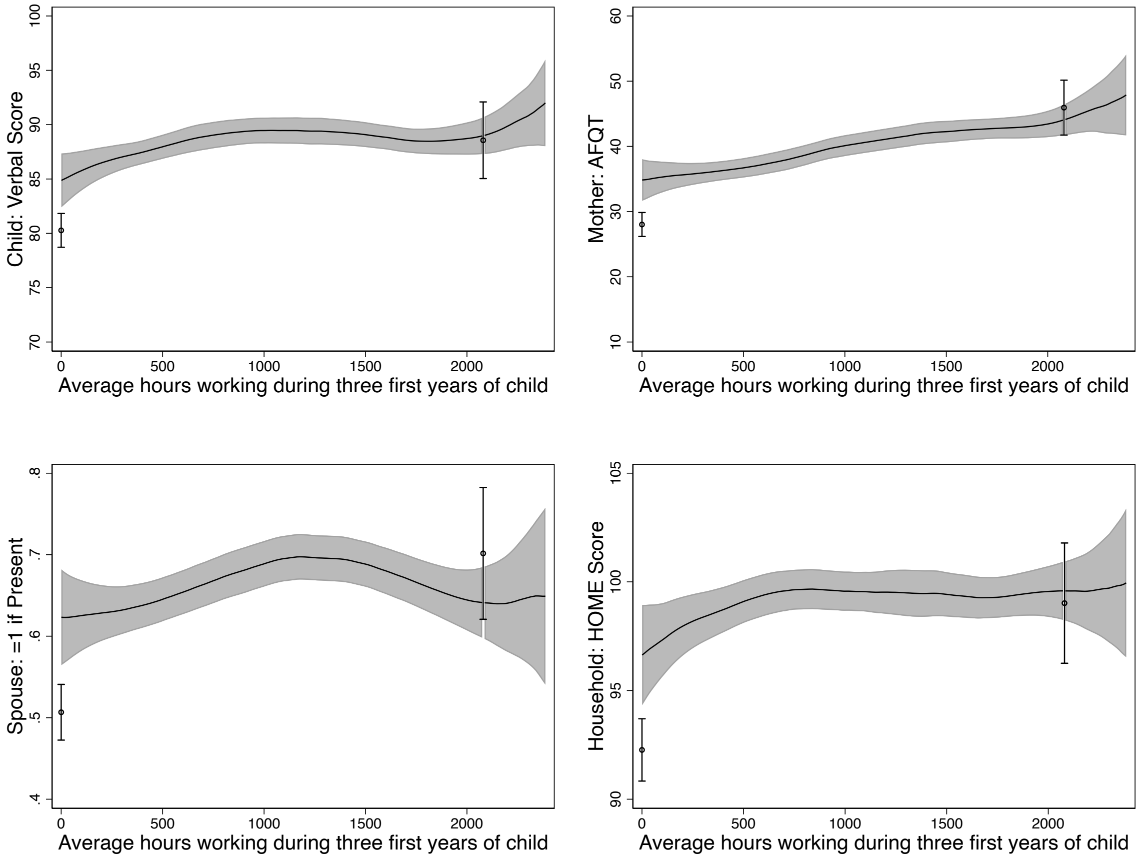
The top left panel in Figure 2 shows the local linear fit of the expected verbal score of the child (our outcome variable) for each positive level of working hours of the mother, as well as the average test score among those who are bunched. The evident discontinuity at zero has only two possible (non-exclusive) explanations: the effect of working hours on skills is discontinuous at zero hours, or the confounders are discontinuous at zero hours.

Indeed, the vast majority of observable confounders are discontinuous at zero hours. The other panels in Figure 2 are constructed similarly to the top left panel already discussed. These panels show

be detected by the second term in equation (3). This can be seen in the example in Appendix B and in the Monte Carlo study (Appendix F).

[6] For example, suppose that $Y = \beta X + Z'\gamma + \delta\zeta + \varepsilon$, where $\mathbb{E}[\zeta|X,Z] = 0$ if $X < 10$, and $\mathbb{E}[\zeta|X,Z] = a(X-10)$ if $X \geq 10$. Then $\mathbb{E}[Y|X,Z]$ is a piecewise linear function of $X$ for $X > 0$, with a kink at $X = 10$. So, while the confounder $\zeta$ does not vary discontinuously at $X = 0$, the dummy test can still detect it because of non-linearities.

Figure 2: Evidence that $\varepsilon$ is Discontinuous at Bunching Points



Note: This figure shows the local linear regression of observables on $X$ (average hours working per year during the child's first three years) along with the 95% confidence interval. The bandwidth is 300 hours. At $X = 0$ and $X = 2,080$, the average along with the 95% confidence interval is also shown. $N = 3,383$. Source: NLSY79. See Section 5 for details about the application.

discontinuities in the mother's Armed Forces Qualifying Test (AFQT) score, a pre-market measure of her academic skills, the presence of the spouse in the household in the year the child took the test, and the Home Observation Measurement of the Environment (HOME) score.[7] Moreover, we find that those children bunched at zero are systematically negatively selected, in the sense that the observables that are positively correlated with $Y$ (verbal test scores) tend to be discontinuously lower at $X = 0$. This is consistent with what we found in the top left panel of Figure 2 for the outcome variable $Y$. Because discontinuities at $X = 0$ are so prevalent in observables, we expect that they should also be prevalent in unobservables. Thus, if there is endogeneity, we expect $\Delta(Z) \neq 0$ in this application.

The models in this literature (e.g. James-Burdumy 2005 and the references therein) rely not only on exogeneity as the main identification assumption; they also assume that the model is linear. Any continuous nonlinearities resulting from the nonlinearity of the true model are reflected in $\Gamma$. For instance, it is possible that one additional hour of work becomes more or less costly for the development of the child's skill the longer hours the mother works. Additionally, it is possible that there are confounders that affect the outcome only after the mother works enough hours (e.g. quality of child care).

[7] The HOME score measures the quality of the home environment of the child for cognitive and emotional development (Bradley and Caldwell 1984; Bradley et al. 1992).

The linearity assumption also indirectly rules out the possibility that the effect of the hours the mother works is discontinuous at zero hours. Indeed, it is plausible that the effect is continuous, as working 0 hours per year in the first three years of the child should have a similar effect on the child's skill at age 4 to working, say, 1 hour per year. In any case, a discontinuity in the treatment effect would affect $\Delta(Z)$, and thus be detected by the dummy test.

Figure 1 also shows bunching of 3% of the sample at 2,080 hours, which is equivalent to 40 hours per week for 52 weeks. This opens the possibility of a multiple dummy test, by including $\mathbf{1}(X = 0)$ and $\mathbf{1}(X = 2,080)$ in the regression, and performing a joint test of whether the coefficients of both dummies are equal to zero, as described in Appendix D.2. However, Figure 2 does not show a corresponding discontinuity in the outcome or observables at that threshold. This suggests that both the unobservables as well as the treatment effect are likely to be continuous at 2,080 hours. As discussed in Appendix D.2, in this instance, the multiple dummy test is advisable only if the amount of bunching at $X = 2,080$ is substantially larger than at $X = 0$, which is not the case here.

# 4 Comparison with Caetano (2015)'s Discontinuity Test

In this section, we compare the power of the dummy test and Caetano (2015)'s Discontinuity Test (CDT). CDT identifies the quantity $\lim_{x\downarrow 0}\mathbb{E}[\mathbb{E}[Y|X=0,Z]-Y|X=x]$. In our context, this is equivalent to

$$\lim_{x\downarrow 0}\mathbb{E}[\Delta(Z)|X=x], \tag{4}$$

provided certain conditions on $\Gamma$ hold (e.g. bounded above by an integrable function). Thus, the power of CDT comes entirely from the discontinuities. In contrast, the dummy test can detect both discontinuities and continuous nonlinearities ($\Gamma(X,Z) \neq \alpha X + Z'\gamma$).

Supposing $\Gamma(X,Z) = \alpha X + Z'\lambda$, the power of both tests depend entirely on the discontinuities. First, we consider the estimated quantities. CDT identifies an average of the discontinuities among the values of $Z$ near the bunching point. The dummy test identifies a more complex quantity (the second term in equation (3)). On the one hand, it subtracts from the average of the discontinuities the part of those discontinuities which is linearly predicted by $X$ and $Z$. On the other hand, it divides this term by $(1 - \mathbf{d}^*)$, a number between 0 and 1. Neither quantity always dominates the other.

In contrast, the standard errors of the estimators of both quantities are very different. CDT uses nonparametric estimators, so the rate of convergence of its test statistic is much slower than that of the dummy test ($\sqrt{nh}$ vs. $\sqrt{n}$, where $h$ is the bandwidth in the local linear regression in CDT). Therefore, the resulting power of the dummy test will usually be larger because the standard errors of the estimators will tend to be much smaller.

The stark difference in power between CDT and the dummy test can be seen in the Monte Carlo simulations in Appendix F. The results there reflect what is expected from the theory: the dummy test detects continuous misspecification (while CDT does not), and has substantially more power to detect endogeneity.

The variance advantage of the dummy test over CDT is not only due to the use of parametric versus nonparametric estimators. To show this, we develop a parametric version of CDT. We refer to this test as the Linear CDT. In model (1), the Linear CDT estimates (4) in two steps:[8]

1. Regress $Y$ onto $Z$ using only observations such that $X = 0$; let the coefficients be $\hat{\lambda}$. Calculate $Q = Z'\hat{\lambda} - Y$ in the entire sample.

2. Regress $Q$ onto $X$ using only observations such that $X > 0$. The intercept of this regression is $\hat{\theta}_{LCDT}$.

If Assumption 1 holds, then the first step is an estimator of $\lambda$, so $Q \approx -\beta X - \epsilon$, where $\epsilon = Y - \mathbb{E}[Y|X, Z]$. Thus, step 2 consistently estimates the true intercept, zero. The Linear CDT is identical to CDT with linear instead of nonparametric regressions in the first and second steps.

The Linear CDT has power to detect misspecification as well as endogeneity, and does not suffer the loss of power from the nonparametric estimation. However, like CDT, it is still a split-sample test, as steps 1 and 2 are estimated on different subsamples of the data. In Appendix F.3, we consider the performance of the Linear CDT in our Monte Carlo study. Although it is substantially more powerful than CDT, it is less powerful than the dummy test.

## 5 Application

We showcase the test using the application in James-Burdumy (2005), which estimates the effect of maternal working hours on children's skills. We assemble the same data, from the National Longitudinal Survey of Youth (NLSY),[9] and we consider both the original sample and an extended sample augmented to include data from more recent survey rounds. In the notation of our paper, $Y$ is the child's verbal test score (Peabody Picture Vocabulary Test), measured around age four, and $X$ is the yearly average number of working hours of the mother in the three years following the child's birth.[10]

The NLSY allows us to observe many covariates that help control for confounders, but controlling only for these covariates might not be sufficient, thus leading to bias in the effect of interest. James-Burdumy (2005) improves on the previous literature by adding time-invariant family fixed effects to these detailed control specifications. Intuitively, the paper aims to compare the test scores of two siblings whose mother worked different hours during their respective first three years of life. The siblings were born in different years, and so the test scores are observed in different calendar years, 1986 and 1988, depending on the sibling. Because family and year fixed effects are used, the identification strategy is a conditional (on observed controls) version of difference-in-differences, with two years and many groups (families). Naturally, there is still the concern that there are other confounders varying with the child within the family, such as factors affecting labor supply and test scores that may change across children during their first three years of life (e.g., spouse's presence, hours of work, quality of child care).

We also consider an extended sample where we include siblings whose test scores are observed in any of the years 1986, 1988, 1990 and 1992. This version of difference-in-differences also compares siblings'

[8] We formalize the Linear CDT in Appendix E.

[9] Specifically, we link maternal work history data during the first three years of a child's life from the National Longitudinal Surveys of Youth 1979 (NLSY79) to the children's skill measures from the Children of the National Longitudinal Surveys (CNLSY).

[10] We start the period in the fourth month after the month of the birth of the child to avoid measurement error related to differences in maternity leave.

test scores when they are observed farther apart from each other, which extends the sample substantially, but may lead to further sources of bias due to violations of the strong parallel trends assumption. Indeed, not only do the unobservables of the family have more scope to change over time (a violation of the standard parallel trends assumption), but the treatment effects are more likely to be heterogeneous (a violation of the uncorrelated treatment effects assumption, which is also necessary for strong parallel trends). Fortunately, the dummy test detects both of these types of confounders (see Appendix C.1.2 for details).

Table 1 presents the results for different versions of the dummy test, different samples, and different identification strategies. The first two columns show results for the original sample from James-Burdumy (2005), while the last two columns show the results for the extended sample. The specification labeled "Diff" refers to the one used by James-Burdumy (2005) with the most detailed set of controls, including year fixed effects.[11] The specification labeled "DiD" adds to these controls family fixed effects, which is the best specification in that paper.[12] In the first two rows of the table, we implement the univariate version of the test, while in the next two rows we implement the multivariate version (Appendix D.1) by allowing for heterogeneity in the coefficient of the dummy variable by whether the spouse is both present and has a high school degree. Specifically, instead of the dummy $\mathbf{1}(X = 0)$, we add two dummies $\mathbf{1}(X = 0, \mathcal{Z})$ and $\mathbf{1}(X = 0, \mathcal{Z}^c)$, where $\mathcal{Z}$ indicates that the spouse is present and has a high school degree, and $\mathcal{Z}^c$ represents all other possibilities, and perform a joint test of whether the two coefficients are equal to zero. We choose to use the presence and level of education of the spouse as the source of heterogeneity because a mother's decision to work likely depends on whether there is a spouse present who is capable of earning enough money on their own to support the family.

The first two columns show that in the original context of James-Burdumy (2005) we strongly reject Assumption 1 for the Diff specification, but we do not reject Assumption 1 for the DiD specification. Thus, the dummy test suggests that James-Burdumy (2005)'s approach of considering a second difference in her analysis is key to controlling for most confounders.

In the next columns of the table, we extend the sample and conduct the analogous comparisons under presumably stronger assumptions (because of the wider range of comparison among siblings across years, as discussed above). Reassuringly, we reject even the DiD identification strategy in this case.

The table also shows that, in this context, the multivariate test (bottom rows) tends to have a bit more power to detect violations from Assumption 1 than the univariate test (top rows). However, the conclusions are very similar irrespective of the version of the dummy test one uses.

[11] The list of controls includes the child's gender, birth order, and age; the mother's age at the child's birth, highest education level, and average wage in the child's first three years of life; whether the spouse is present; the spouse's income and highest education level; the number of children in the household with ages 0-2, 3-5, 6-11 and above 12; and region of residence and survey year fixed effects.

[12] James-Burdumy (2005) also provides an IV approach, but argues that this DiD specification is the preferred one.

Table 1: F-statistics and P-Values of Dummy Tests

| | Sample<br>Bunching Location | Original<br>X=0 | | Extended<br>X=0 | |
|---|---|---|---|---|---|
| Identification Strategy | | Diff | DiD | Diff | DiD |
| Univariate | F statistic | 10.358 | 0.000 | 16.071 | 2.428 |
| | p-value | 0.001 | 0.997 | 0.000 | 0.119 |
| Multivariate | F statistic | 6.485 | 0.987 | 18.088 | 2.471 |
| | p-value | 0.002 | 0.373 | 0.000 | 0.085 |
| | N | 1867 | 1172 | 3383 | 2545 |

Note: This table shows the F statistics and p-values of the univariate and multivariate dummy tests for different samples and identification strategies. The samples are either the original one used in James-Burdumy (2005), or the extended sample discussed in the main text. The identification strategies are either Differences (Diff) or Difference-in-Differences (DiD), where the former includes several controls plus year FEs (see Footnote 11), while the latter adds family FEs as well. In the multivariate case, we allow for heterogeneity in the coefficient of the dummy by whether the spouse is both present and has a high school degree.

We conclude that most of the sources of bias in the original sample and application seem to come from confounders that vary across families, which are absorbed by the family fixed effects. By focusing on siblings whose test scores were observed within a narrow range of years, James-Burdumy (2005) seems to have successfully controlled for confounders with the DiD identification strategy. By contrast, an extended version of that DiD identification strategy where siblings' test scores are allowed to be observed within a wider range of years does not seem to be valid.

## 6 Nonlinear Models

The dummy test can be implemented in nonlinear models. If the model allows for the inclusion of the dummy and the identification of its coefficient under the null, the test can be performed. The gamut of such models is very large, and it is not possible, as far as we know, to characterize every identification strategy under the same conceptual umbrella. In this section, we show how the dummy test can be applied to a wide range of nonlinear models that are estimated with extremum estimators. This includes most classical models which are estimated by Maximum Likelihood or GMM, such as nonlinear regression, probit, and discrete choice models, among others.

**Assumption 3.** *Suppose that, under some condition $A$, the parameter $\gamma_0$ can be identified as*

$$\gamma_0 = \underset{\gamma \in \Lambda}{\operatorname{argmax}}\, M_0(Y, X, Z; \gamma).$$

This assumption states that $\gamma_0$ is identified as the argument which maximizes a function $M_0$ within a parameter set $\gamma$ when condition $A$ holds. We want to test whether assumption $A$ holds. We will now extend the model to a more general one which includes the dummies and nests the original model under assumption $A$.

**Assumption 4.** *Suppose that there exists a function $Q_0(Y, X, Z, \mathbf{1}(X = 0); \gamma, \delta)$ such that if assumption A holds,*

$$(\gamma_0, 0) = \underset{\gamma\in\Lambda,\delta\in\Omega}{\operatorname{argmax}}\, Q_0(Y, X, Z, \mathbf{1}(X = 0); \gamma, \delta),$$

*and*

$$Q_0(Y, X, Z, \mathbf{1}(X = 0); \gamma, 0) = M_0(Y, X, Z; \gamma).$$

The following theorem establishes that one can test assumption $A$ by testing whether $\delta = 0$ using a $t$ or $F$ test, depending on whether $\delta$ is a scalar or a multivariate vector.

**Theorem 6.1.** *Let $\hat{Q}_n$ be an estimator of $Q_0$. If Assumptions 3 and 4, as well as the conditions of Theorem 3.1 in Newey and McFadden (1994) hold,*[13] *then if condition A holds,*

$$\sqrt{n\hat{V}}\begin{pmatrix} \hat{\gamma} - \gamma_0 \\ \hat{\delta} \end{pmatrix} \to_d \mathcal{N}(\mathbf{0}, I),$$

*where $I$ is the identity matrix, and $\hat{V}$ is the consistent estimator of the asymptotic variance of the coefficients built using Theorem 4.1 in Newey and McFadden (1994).*

The proof of Theorem 6.1 is trivial and is thus omitted. Note that the setup above is true for any specification test based on the inclusion of an additional variable into the model which should not be there if the identification assumption holds. We propose specifically the inclusion of the dummy because of the discontinuities in confounders that are often found at bunching points, as we discuss in Section 3.

In Appendix C.2, we discuss assumptions, power, and implementation details in the context of some well known models fitting this setting: standard nonlinear models which are estimated with GMM (Appendix C.2.1), probit (Appendix C.2.2), and discrete choice models (Appendix C.2.3).

## 7 Conclusion

We propose a simple test of identification when the treatment variable takes multiple values and has a bunching point. The test is easy to implement: it consists of adding a dummy of the bunching point to the model and testing if the coefficient of the dummy is equal to zero. To increase power, one may also interact the dummy with controls, or include dummies of additional bunching points. The dummy test is similar in spirit to Caetano (2015)'s discontinuity test, but it is more powerful at detecting endogeneity, and it also detects misspecification. The test can be used to validate identification strategies or diagnose problems, and it has advantages over Caetano (2015)'s discontinuity test on both accounts.

The test can be naturally extended for a multivariate treatment vector with bunching points at all coordinates, as implemented in Caetano and Maheshri (2018) and Caetano et al. (2019). We conjecture that this test can also be extended to other contexts where bunching has been used for testing, analogously to what has been done by Caetano et al. (2016) for control function approaches and by Khalil and Yildiz (2019) for treatment variables without bunching.

[13] Note that implicit in these conditions is often a requirement that $\mathbf{1}(X = 0)$ is not a part of $Z$, that is, the model is continuous in $X$ at $X = 0$. It is not always necessary for the model to be continuous if the discontinuity does not invalidate the rank condition on the extended model (e.g. the model is $Y = \beta X/(1 + \alpha\mathbf{1}(X = 0)) + \varepsilon$, and the extended model includes the dummy additively).

# References

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184.

Anderson, S. T. and Sallee, J. M. (2011). Using loopholes to reveal the marginal cost of regulation: The case of fuel-economy standards. *American Economic Review*, 101(4):1375–1409.

Arellano, M. et al. (1987). Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

Berry, S., Linton, O. B., and Pakes, A. (2004). Limit theorems for estimating the parameters of differentiated product demand systems. *The Review of Economic Studies*, 71(3):613–654.

Bertanha, M., McCallum, A. H., and Seegert, N. (2021). Better bunching, nicer notching. *arXiv preprint arXiv:2101.01170*.

Bester, C. A., Conley, T. G., and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.

Bester, C. A., Conley, T. G., Hansen, C. B., and Vogelsang, T. J. (2016). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. *Econometric Theory*, 32(1):154.

Bleemer, Z. (2018a). The effect of selective public research university enrollment: Evidence from california. *Working Paper*.

Bleemer, Z. (2018b). Top percent policies and the return to postsecondary selectivity. *Available at SSRN 3272618*.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bradley, R. H. and Caldwell, B. M. (1984). The home inventory and family demographics. *Developmental Psychology*, 20(2):315.

Bradley, R. H., Caldwell, B. M., Brisby, J., Magee, M., Whiteside, L., and Rock, S. L. (1992). The home inventory: a new scale for families of pre-and early adolescent children with disabilities. *Research in developmental disabilities*, 13(4):313–333.

Caetano, C. (2015). A test of exogeneity without instrumental variables in models with bunching. *Econometrica*, 83(4):1581–1600.

Caetano, C., Caetano, G., and Nielsen, E. (2020). Correcting Endogeneity Bias in Models with Bunching. *Working Paper*. Available here.

Caetano, C., Caetano, G., and Nielsen, E. (2021). Should children do more enrichment activities? Leveraging bunching to correct for endogeneity. *Working Paper*. Available here.

Caetano, C., Rothe, C., and Yıldız, N. (2016). A discontinuity test for identification in triangular nonseparable models. *Journal of Econometrics*, 193(1):113–122.

Caetano, G., Kinsler, J., and Teng, H. (2019). Towards causal estimates of children's time allocation on skill development. *Journal of Applied Econometrics*, 34(4):588–605.

Caetano, G. and Maheshri, V. (2018). Identifying Dynamic Spillovers of Crime with a Causal Approach to Model Selection. *Quantitative Economics*, 9(1):343–394.

Caetano, G. and Maheshri, V. (2019). Gender segregation within neighborhoods. *Regional Science and Urban Economics*, 77:253–263.

Caetano, G. and Maheshri, V. (2021). Explaining Recent Trends in US School Segregation. Technical report, Forthcoming.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. (2021). Dose-response difference in differences: Identification. *Working Paper*.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory*. Springer - New York.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Difference-in-differences estimators of intertemporal treatment effects. *Available at SSRN 3731856*.

de Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

de Chaisemartin, C. and d'Haultfoeuille, X. (2018). Fuzzy differences-in-differences. *The Review of Economic Studies*, 85(2):999–1028.

De Vito, A., Jacob, M., and Müller, M. A. (2019). Avoiding taxes to fix the tax code. *Available at SSRN 3364387*.

D'Haultfoeuille, X., Hoderlein, S., and Sasaki, Y. (2021). Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. *arXiv preprint arXiv:2104.14458*.

Dubé, J.-P., Fox, J. T., and Su, C.-L. (2012). Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267.

Dube, O. and Vargas, J. F. (2013). Commodity price shocks and civil conflict: Evidence from colombia. *The review of economic studies*, 80(4):1384–1421.

Erhardt, E. C. (2017). Microfinance beyond self-employment: Evidence for firms in bulgaria. *Labour economics*, 47:75–95.

Fe, H. and Sanfelice, V. (2020). How bad is crime for business? evidence from consumer behavior. *Center for Health Economics and Policy Studies Working Paper*.

Ferreira, D., Ferreira, M. A., and Mariano, B. (2018). Creditor control rights and board independence. *The Journal of Finance*, 73(5):2385–2423.

Forman, C., Goldfarb, A., and Greenstein, S. (2012). The internet and local wages: A puzzle. *American Economic Review*, 102(1):556–75.

Harding, M. and Lovenheim, M. (2017). The effect of prices on nutrition: comparing the impact of product-and nutrient-specific taxes. *Journal of Health Economics*, 53:53–71.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imberman, S. A., Kugler, A. D., and Sacerdote, B. I. (2012). Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, 102(5):2048–82.

James-Burdumy, S. (2005). The effect of maternal labor force participation on child development. *Journal of Labor Economics*, 23(1):177–211.

Khalil, U. and Yildiz, N. (2019). A Test of Selection on Observables Assumption Using a Discontinuously Distributed Covariate. working paper.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–464.

Lavetti, K. and Schmutte, I. M. (2018). Estimating compensating wage differentials with endogenous job mobility. *Working paper.*

Lee, L.-F. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2):333–374.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714.

McFadden, D. et al. (1973). *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, University of California.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy*, 9(4):513–548.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Nunn, N. (2008). The long-term effects of africa's slave trades. *The Quarterly Journal of Economics*, 123(1):139–176.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Pang, J. (2017). Do subways improve labor market outcomes for low-skilled workers. *working paper, Syracuse University.*

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rozenas, A., Schutte, S., and Zhukov, Y. (2017). The political legacy of violence: The long-term impact of stalin's repression in ukraine. *The Journal of Politics*, 79(4):1147–1161.

Słoczyński, T. (2020). Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *Forthcoming, Review of Economics and Statistics.*

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge university press.

Train, K. E. and Winston, C. (2007). Vehicle choice behavior and the declining market share of us automakers. *International economic review*, 48(4):1469–1496.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

# A Proofs of the claims in Section 3

First, we prove the statements made prior to equation (3). Let $\hat{m}_{ZX} = n^{-1}(\mathbf{z}'(I - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\mathbf{z})$, then

$$(\mathbf{d}'M_{\mathbf{w}}\mathbf{d})^{-1}\mathbf{d}'M_{\mathbf{w}}\boldsymbol{\Delta}_0 = \frac{\mathbf{d}'\boldsymbol{\Delta}_0 - \left(\frac{1}{n}\sum_{i=1}^{n} Z_i\mathbf{1}(X_i = 0)\right)\hat{m}_{ZX}^{-1}\mathbf{z}'\boldsymbol{\Delta}_0}{\mathbf{d}'\mathbf{d} - \left(\frac{1}{n}\sum_{i=1}^{n} Z_i\mathbf{1}(X_i = 0)\right)\hat{m}_{ZX}^{-1}\mathbf{z}'\mathbf{d}}.$$

Note that the coefficient vector of $Z$ in a regression of a variable $Q$ onto $X$ and $Z$ is $\hat{m}_{ZX}^{-1}\mathbf{z}'M_{\mathbf{x}}\mathbf{q}$, where $M_{\mathbf{x}} = I - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$ and $\mathbf{q} = (Q_1, \dots, Q_n)'$. Moreover, the predicted value of such a regression at $X = 0$ is $\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\mathbf{1}(X_i = 0)\right)\hat{m}_{ZX}^{-1}\mathbf{z}'M_{\mathbf{x}}\mathbf{q}$. However, note that since $X$ is orthogonal to $\mathbf{1}(X = 0)$, $M_{\mathbf{x}}\mathbf{d} = \mathbf{d}$, and $M_{\mathbf{x}}\boldsymbol{\Delta}_0 = \boldsymbol{\Delta}_0$. Therefore, the second term in the numerator is the prediction of $\Delta(Z)\mathbf{1}(X = 0)$ in a regression onto $X$ and $Z$ at $X = 0$, and the second term in the denominator is the prediction of $\mathbf{1}(X = 0)$ in a regression onto $X$ and $Z$ at $X = 0$. The denominator is a quadratic form, and by Assumption 2, for $n$ large enough, it is positive with probability equal to one. Analogous interpretations can be made for the first term in (3).

Next, we prove equation (3) itself. The convergence in probability of $\mathbf{d}'\mathbf{d}/n$, $\mathbf{d}'\boldsymbol{\Delta}_0/n$, $\mathbf{d}'\mathbf{z}/n$, $\mathbf{z}'\boldsymbol{\Delta}_0/n$ and $\hat{m}_{ZX}$ can be shown using whichever laws of large numbers are applicable for the specific data structure and $Z$ specification. For example, under the setting in White (1980) (i.e. independent but not identically distributed observations, and $\mathbb{E}[|\Delta(Z_i)|^{2+\alpha}] \leq A$, $\mathbb{E}[||Z_i||^{2+\alpha}] \leq A$ and $\mathbb{E}[||Z_i\Delta(Z_i)||^{2+\alpha}] \leq A$ for some $\alpha, A > 0$), the results are obtained by Brunk-Chung's Strong Law of Large Numbers (see Chow and Teicher (1997), Theorem 10.1.3, for r=1). By the continuous mapping theorem,

$$(\mathbf{d}'M_{\mathbf{w}}\mathbf{d})^{-1}\mathbf{d}'M_{\mathbf{w}}\boldsymbol{\Delta}_0 \to_p \frac{\mathbb{E}[\Delta(Z)|X = 0] - \mathbb{E}[Z|X = 0]'m_{ZX}^{-1}\mathbb{E}[Z\Delta(Z)\mathbf{1}(X = 0)]}{1 - \mathbb{E}[Z|X = 0]'m_{ZX}^{-1}\mathbb{E}[Z\mathbf{1}(X = 0)]}.$$

The convergence of the first term to $(\mathbb{E}[\Gamma(0, Z)|X = 0] - \boldsymbol{\Gamma}_0^*)/(1 - \mathbf{d}^*)$ is established analogously.

# B A model of constrained choice

In this section, we present a model where $X$ is the result of a constrained problem. We show that it is possible to interpret $\Gamma$ and $\Delta$ structurally. Importantly, this example provides intuition that endogeneity and nonlinearities typically affect both $\Gamma$ and $\Delta$, but in different ways.

Consider the case that $X$ cannot be negative. The choice of $X$ is assumed to result from a combination of observable, $Z$, and unobservable, $\eta$, factors under a constrained problem. The solution to this problem thus yields

$$X = \max\{0, h(Z, \eta)\}, \tag{5}$$

where $\eta$ is scalar and $h$ is strictly monotonic in $\eta$ (suppose it is increasing, for the sake of the exposition). Assume that $0 < \mathbb{P}(h(\eta; Z) < 0) < 1$, so that the constraint is binding to a subset of the population. Define $X^* = h(Z, \eta)$ as the "latent" or "desired" choice absent the non-negativity constraint.

Let the model be

$$Y = g(X, Z) + m(Z)\eta + \nu,$$

where $g$ is continuous in $X$ at $X = 0$, and $\mathbb{E}[\nu|X, Z, \eta] = 0$. Since the intention is to estimate the effect of $X$ on $Y$ by regressing $Y$ onto $X$ and $Z$, there are two concerning problems. First, there may be misspecification of the functional form: $g(X,Z) \neq \beta X + Z'\lambda$. Second, there may be endogeneity: $m(Z) \neq 0$.

Define the notation $D_f(Z) = f(0,Z) - \lim_{x\downarrow 0} f(x,Z)$, and assuming the limits below exist, we can write

$$\begin{aligned}\mathbb{E}[Y|X,Z] = {} & [g(X,Z) - D_g(Z)\mathbf{1}(X=0)] + m(Z)[h^{-1}(X;Z) - D_{h^{-1}}(Z)\mathbf{1}(X=0)] \\ & + D_g(Z)\mathbf{1}(X=0) + m(Z)[\mathbb{E}[h^{-1}(X^*;Z)|X^* \leq 0, Z] - \lim_{x\downarrow 0} h^{-1}(x,Z)]\mathbf{1}(X=0).\end{aligned}$$

Note that $\mathbb{E}[h^{-1}(X^*;Z)|X^* \leq 0, Z] - \lim_{x\downarrow 0} h^{-1}(x,Z) \leq 0$ a.s., and $\mathbb{P}(\mathbb{E}[h^{-1}(X^*;Z)|X^* \leq 0, Z] - \lim_{x\downarrow 0} h^{-1}(x,Z) < 0) > 0$, thus if $g$ is continuous in $X$ at zero, there will be a discontinuity in $\mathbb{E}[Y|X,Z]$ if and only if $m(Z) \neq 0$ (i.e. if and only if there is endogeneity). If, additionally, $g$ is discontinuous, this may increase or decrease the discontinuity in the outcome depending on whether the sign of both discontinuities are equal or different, respectively.

This model satisfies the structure in (2). Here, $\Gamma(X,Z) = -\beta X + [g(X,Z) - D_g(Z)\mathbf{1}(X=0)] + m(Z)[h^{-1}(X;Z) - D_{h^{-1}}(Z)\mathbf{1}(X=0)]$, which depends on the continuous nonlinearities in $g$, and the continuous nonlinearities in $h$ if there is endogeneity. $\Delta(Z) = D_g(Z)\mathbf{1}(X=0) + m(Z)[\mathbb{E}[h^{-1}(X^*;Z)|X^* \leq 0, Z] - \lim_{x\downarrow 0} h^{-1}(x,Z)]\mathbf{1}(X=0)$, which depends on the discontinuity of $g$ at $X=0$, and on whether there is endogeneity. Therefore, endogeneity affects both $\Gamma$ (via the nonlinearity of $h$) and $\Delta$. The nonlinearity in $g$ affects $\Gamma$, and if additionally $g$ is discontinuous, it also affects $\Delta$.

To understand the power when there are no nonlinearities, suppose that $h(Z,\eta) = Z'\pi + \eta$, $g(X,Z) = \beta X + Z'\gamma$, and $m(Z) = \delta$. Then, $\Gamma(X,Z) = \delta X + Z'(\gamma - \pi\delta)$, and $\Delta(Z) = \delta\mathbb{E}[X^*|X^* \leq 0, Z]$. In this case, the first component of the power of equation (3) is equal to zero. The power then depends entirely on the discontinuity. That is, it depends on the magnitude of $\delta$, and on how binding the constraint on the choice of $X$ is, which is expressed on how negative is $\mathbb{E}[X^*|X^* \leq 0, Z]$.

# C Examples of linear and nonlinear models

In this section we discuss some popular models, and how the dummy test may be applied in those contexts.

## C.1 Linear models

In this section we show that heterogeneous treatment effect models, as well as difference-in-differences models estimated with two-way fixed effect regressions, fit the linear framework in equation (1), and discuss the meaning of Assumption 1 in those contexts. Therefore, the standard linear dummy test can be used.

### C.1.1 Heterogeneous treatment effects

The linear setting from equation (1) includes a standard heterogeneous treatment effects model with multivalued treatment, when one wishes to identify the average treatment effect. Suppose that

$$Y_i = \beta_i X_i + U_i. \tag{6}$$

This model can be written as equation (1), where $\beta = \mathbb{E}[\beta_i]$ is the average treatment effect, and $\varepsilon_i = U_i + (\beta_i - \beta)X_i$. Supposing that the covariate vector $Z_i$ includes only a constant, Assumption 1 requires that both $U_i$ and $\beta_i$ are mean independent of $X_i$, which are the standard assumptions for identification of average treatment effects.[14] If $Z_i$ includes other variables besides a constant, Assumption 1 is equivalent to $\mathbb{E}[U_i|X_i, Z_i] = Z_i'\lambda$ and $\mathbb{E}[\beta_i|X_i, Z_i] = \mathbb{E}[\beta_i]$, which are the necessary conditions for the identification of average treatment effects from a regression of $Y$ on $X$ with controls.[15]

### C.1.2 Difference-in-differences

The linear setting from equation (1) also applies to testing whether average treatment effects are identified by difference-in-differences in a two-way fixed effects regression. (In fact, the arguments below also hold analogously for one-way and multi-way fixed effects, as well as for pooled cross-section data with group fixed effects.) Here we discuss the standard case with two periods where nobody is treated in the first period, but the test can also be applied with multiple periods with or without staggered treatment adoption. Suppose that the treatment variable of individual $i$ in time $t$ is $D_{i,t}$, which assumes multiple values, and the potential outcome under each treatment level $d$ is $Y_{i,t}(d)$. We intend to run a regression of the observed outcome $Y_{i,t}$ onto $D_{i,t}$ using individual and time fixed-effects. This is equivalent to our model, where $X_{i,t} = D_{i,t}$ and $Z_{i,t} = (1, \mathbf{1}(i = 1), \ldots, \mathbf{1}(i = N-1), \mathbf{1}(t = 1))'$.[16]

The potential outcome of treatment $d$ is given by

$$Y_{i,t}(d) = \alpha + \beta_{i,t} d + \gamma_i + \delta_t + U_{i,t}, \tag{7}$$

where $\gamma_i$ is an unobservable that does not vary in time, and $\delta_t$ is an unobservable which does not vary per individual. Here, the treatment effect of one additional unit is the same for all $d$, that is, $Y_{i,t}(d) - Y_{i,t}(d-1) = \beta_{i,t}$, and we are interested in identifying $\beta = \mathbb{E}[\beta_{i,t}]$.

Define $\lambda = (\alpha + \gamma_N + \delta_0, \gamma_1 - \gamma_N, \ldots, \gamma_{N-1} - \gamma_N, \delta_1 - \delta_0)'$. Then $\varepsilon_{i,t} = Z_{i,t}'\lambda + U_{i,t} + (\beta_{i,t} - \beta)D_{i,t}$. Assumption 1 in this context is therefore equivalent to $\mathbb{E}[U_{i,t}|X_{i,t}, Z_{i,t}] = 0$ (exogeneity) and $\mathbb{E}[\beta_{i,t}|X_{i,t}, Z_{i,t}] = \mathbb{E}[\beta_{i,t}]$ (uncorrelated random effects).[17] In this model, these conditions correspond

[14] In this model, Assumption 1 is equivalent to ignorability of $X$ (e.g., Rosenbaum and Rubin 1983 and Imbens 2000). In the notation of the potential outcomes model, $X_i = D_i$, $\beta_i = Y_i(1) - Y_i(0)$ and $U_i = Y_i(0)$. Then, $\beta = \mathbb{E}[Y_i(1) - Y_i(0)]$ is the average treatment effect of one additional treatment unit, and $\varepsilon_i = Y_i(0) + ([Y_i(1) - Y_i(0)] - \mathbb{E}[Y_i(1) - Y_i(0)])D_i$. Assumption 1 in this model is thus equivalent to exogeneity ($\mathbb{E}[Y_i(0)|D_i, Z_i] = \mathbb{E}[Y_i(0)]$) and uncorrelated treatment effects ($\mathbb{E}[Y_i(1) - Y_i(0)|D_i, Z_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$).

[15] The often assumed conditional ignorability condition (here equivalent to $\mathbb{E}[U_i|X_i, Z_i] = \mathbb{E}[U_i|Z_i]$, and $\mathbb{E}[\beta_i|X_i, Z_i] = \mathbb{E}[\beta_i|Z_i]$) is weaker than Assumption 1, but it is not sufficient for the identification of average treatment effects in linear regression models (see Słoczyński 2020).

[16] Defining $Z$ in this manner is, of course, an abuse of notation. The vector of fixed effect dummies is not, technically, a random variable.

[17] Here we prove only that Assumption 1 implies $\mathbb{E}[U_{i,t}|X_{i,t}, Z_{i,t}] = 0$, as, other than that, the statement is trivial. Note

to the "strong parallel trends" assumption discussed in Callaway et al. (2021) as the main necessary condition for the identification of average treatment effects in difference-in-differences models.[18] See also related Assumptions 4, 5 and 7 in de Chaisemartin and d'Haultfoeuille (2020) for the case with discrete ordered treatment.

Note that the dummy test in the two-way fixed effects model is also a test of the linearity assumption in model (7). Indeed, if linearity does not hold, the two-way fixed effects regression cannot identify average treatment effects (see Corollary 2 in de Chaisemartin and D'Haultfœuille 2020 and Theorem 3 in Callaway et al. 2021). In this case, identification of interesting quantities may be done with other approaches, such as the ones suggested by the papers cited above, as well as by D'Haultfoeuille et al. (2021).

## C.2 Nonlinear models

Here we discuss implementation details of some well known examples of nonlinear models within the setting discussed in Section 6. We use several results in Newey and McFadden (1994), which we abbreviate as N-MF.

### C.2.1 Standard nonlinear model

Suppose that

$$Y = g(X, Z; \gamma_0) + U,$$

where $g$ is a known function, and $Y$, $X$ (a scalar) and $Z$ are observable variables.

The parameter $\gamma_0$ is identifiable if $\mathbb{E}[U|X, Z] = 0$.[19] If this condition does not hold, then without loss of generality

$$\mathbb{E}[Y|X, Z] = g(X, Z; \gamma) + \Gamma(X, Z) + \Delta(Z)\mathbf{1}(X = 0).$$

One can perform the test by including the dummy $\mathbf{1}(X = 0)$ into the nonlinear or GMM regression of $Y$ onto $X$ and $Z$. Here, we propose testing the identification assumptions by estimating the model $Y = g(X, Z; \gamma_0) + \delta\mathbf{1}(X = 0) + \nu$ as if $\mathbb{E}[\nu|X, Z] = 0$, and testing if $\delta$ is equal to zero. Barring very specific functional shapes of $\Gamma$, this test also has the power to detect $\Gamma \neq 0$ because nonlinearities – whether caused by misspecification or endogeneity – are at least partially absorbed by the dummy variable.

that $\mathbb{E}[U_{i,t}|Z_{i,t}] = \sum_j a_i\mathbf{1}(j = i) + \sum_s b_s\mathbf{1}(s = t) + \sum_j\sum_s c_{js}\mathbf{1}(j = i)\mathbf{1}(s = t) = a_i + b_t + \sum_j\sum_s c_{js}\mathbf{1}(j = i)\mathbf{1}(s = t)$. Without loss of generality, assume $a_i = 0$ and $b_t = 0$ (since $\gamma_i$ and $\delta_t$ are in the model). If Assumption 1 holds, then $\mathbb{E}[U_{i,t}|X_{i,t}, Z_{i,t}] = \mathbb{E}[U_{i,t}|Z_{i,t}] = Z'_{i,t}\lambda$, implies that $c_{it} = 0$, because the terms in the last sum are not a linear combination of the elements of $Z_{i,t}$.

[18] The strong parallel trends assumption states that $\mathbb{E}[Y_{i,t}(d) - Y_{i,t-1}(0)|D_{i,t} = d] = \mathbb{E}[Y_{i,t}(d) - Y_{i,t-1}(0)]$ for all $d$. That strong parallel trends implies Assumption 1 under linearity is trivial given equation (7) and Footnote 17. Conversely, here we show the stronger result that Assumption 1 always implies strong parallel trends under model (1), not just in model (7):

$$\begin{aligned}\mathbb{E}[Y_{i,t}(d) - Y_{i,t-1}(0)|D_{i,t} = d'] &= \mathbb{E}[\mathbb{E}[Y_{i,t}(d)|D_{i,t} = d', Z_{i,t}] - \mathbb{E}[Y_{i,t-1}(0)|D_{i,t} = d', Z_{i,t-1}]|D_{i,t} = d'] \\ &= \beta d + \mathbb{E}[\mathbb{E}[\varepsilon_{i,t}|D_{i,t} = d', Z_{i,t}] - \mathbb{E}[\varepsilon_{i,t-1}|D_{i,t} = d', Z_{i,t-1}]|D_{i,t} = d'] \\ &= \beta d + \mathbb{E}[Z_{i,t} - Z_{i,t-1}|D_{i,t} = d']'\lambda = \beta d + \delta_t - \delta_{t-1},\end{aligned}$$

which does not vary with $d'$.

[19] Plus some rank condition that is specific to the identification method. For example, if we intend to identify $\gamma_0$ with GMM, see Lemma 2.3 in N-MF.

If we intend to estimate $\gamma_0$ with a GMM regression, the test is correctly sized under the rank and regularity conditions in Lemma 2.3, and the regularity conditions in Theorems 3.4 and 4.5 in N-MF.

### C.2.2 Probit model

Suppose that

$$Y = \mathbf{1}(\beta X + Z'\lambda + U > 0).$$

In this model, $\beta$ is identifiable if $U|X, Z \sim \mathcal{N}(Z'\pi, \sigma^2)$[20] and is usually estimated with Maximum Likelihood. To test the identification conditions, we propose estimating instead the model

$$Y = \mathbf{1}(\beta X + Z'\gamma + \delta \mathbf{1}(X = 0) + U > 0),\ U|X, Z \sim \mathcal{N}(Z'\pi, \sigma^2),$$

and testing whether $\delta \neq 0$. If Assumption 2 holds and $X$ and $Z$ have finite fourth moments, the test is correctly sized by Theorems 3.3 and 4.4 in N-MF (see example 1.2 in pages 2147 and 2159).

This test has the power to detect nonlinearities in the original model, as well as violations of the normality or homoskedasticity assumptions. Most importantly, it can detect discontinuities in confounders at $X = 0$. In the extreme case, suppose that linearity, normality and homoskedasticity hold, but that there is endogeneity which causes a discontinuity in $U$. That is, suppose $U|X, Z \sim \mathcal{N}(\tau X + Z'\pi + \kappa \mathbf{1}(X = 0), \sigma^2)$. Then the estimator of the coefficient of the dummy in the extended model is, in fact, an estimator of $\kappa$.

### C.2.3 Discrete choice model

Consider a standard selection on unobservables discrete choice model where individuals, indexed by $i$, choose an option $j$ in choice set $\mathbb{J}$ in order to maximize their utility. Specifically, individuals solve the optimization problem

$$\max_{j \in \mathbb{J}} V_{ij} = V(X_{ij}, Z_{ij}; \beta, \lambda) + \xi_j + \epsilon_{ij}, \tag{8}$$

for a known function $V(\cdot)$, where $X_{ij}$ (a scalar) and $Z_{ij}$ are often understood as the characteristics of the product that individual $i$ obtains if they choose option $j$, and $\beta$ and $\gamma$ are understood as the corresponding preference parameters. The term $\xi_j$ is often interpreted as the mean utility that individuals obtain from unobservable characteristics of option $j$. We observe $X$, $Z$ and the share of people who choose each alternative option, and we are interested in identifying $\beta$.[21]

One can identify $\beta$ under an exogeneity assumption, $\mathbb{E}[\epsilon_{ij}|X_{ij}, Z_{ij}, \xi_j] = \mathbb{E}[\epsilon_{ij}|Z_{ij}, \xi_j]$, functional form assumptions on $V$, and distributional assumptions about the idiosyncratic error, $\epsilon_{ij}$.[22] For instance, it

[20] Plus the rank condition that $\mathbb{E}[(X, Z')'(X, Z)]$ is invertible. Note also that in the context of Theorem 6.1, $\gamma_0 = (\beta, \gamma' + \pi')'$ in this model.

[21] With individual-level data on people's choices, it is also common to allow for heterogeneous preferences, $\beta_i$, depending on individual-level observables. In this case, $\beta_i$ is modelled as a function of the elements of $Z_{ij}$ which do not vary with $j$, and the function $V$ can then be reparameterized. Thus, for example, if $\beta_i = \alpha + \pi' Z_{1i}$ (for a subvector $Z_{1i}$ of $Z_{ij}$), then we can redefine $\beta = (\alpha, \pi')'$ (see Nevo 2000). Therefore, the heterogeneous preferences setting fits equation (8), and we can test the specification of $V$ and $\beta_i$ jointly.

[22] Plus a rank condition requiring that the aggregated market shares be continuously differentiable, and the matrix of the derivatives be invertible (see Assumption 2 in Berry et al. 2004). Note also that, in the context of Theorem 6.1,

is common in applications to assume that $V$ is linear and $\epsilon_{ij}$ is i.i.d. Extreme Value Type I. Estimation is usually done with simulation-assisted methods, as in Berry et al. (1995) or any of the alternative methods developed thereafter in this extensive literature (e.g. Dubé et al. 2012).

If the identification condition does not hold, then without loss of generality, we can write

$$\mathbb{E}[\epsilon_{ij}|X_{ij}, Z_{ij}, \xi_j] = \Gamma(X_{ij}, Z_{ij}) + \Delta(Z_{ij})\mathbf{1}(X_{ij} = 0),$$

where $\Gamma$ is continuous in $X_{ij}$ (see Footnote 2). If $X_{ij}$ has bunching at zero, this justifies our choice to test the exogeneity condition by including the dummy $\mathbf{1}(X_{ij} = 0)$ in the model as an additive term in equation (8). More specifically we can use the augmented parametric function $\tilde{V}(X_{ij}, Z_{ij}, 1(X_{ij} = 0); \beta, \lambda, \delta) = V(X_{ij}, Z_{ij}; \beta, \lambda) + \delta 1(X_{ij} = 0)$, then estimate $\delta$ jointly with the other parameters with the same estimation approach discussed in the last paragraph. The test is correctly sized under the assumptions of the specific estimation method used. For example, if using the method in Berry et al. (1995), the correct size follows from Theorem 2 and the discussion about standard errors that follows it in Berry et al. (2004), under their assumptions A1-A6 and B1-B5.

The dummy test detects violations from the exogeneity, functional form of $V$, and distributional assumptions. If the dummy test in this scenario is rejected, it motivates the use of different function form or distributional assumptions, or the use of control function approaches such as those discussed in Chapter 13 of Train (2009).

If one wants to identify $\beta$ in a model without the unobservable term $\xi_j$, as in McFadden et al. (1973), then estimation is less computationally burdensome. The dummy test in this setting can be done similarly by including the dummy additively into the utility specification. The correct size of the test follows from the results in Chapter 10 of Train (2009), which depend on the estimation method (e.g. maximum simulated likelihood, method of simulated moments or method of simulated scores). In this case, if the dummy test is rejected, it motivates the inclusion of $\xi_j$, as discussed above.

# D Extensions: using multiple dummies

Depending on the context, it is possible to extend the framework discussed above to use multiple dummies. Below, we consider two scenarios which are not mutually exclusive: using interactions with covariates and additional bunching points.

## D.1 Allowing for heterogeneous effects of the dummy

It is clear in equation (2) that the discontinuities may vary with $Z$. For example, if $Z \in \{z_1, \ldots, z_L\}$ has finite support, equation (2) can be rewritten as

$$\mathbb{E}[\varepsilon|X, Z] = \Gamma(X, Z) + \sum_{l=1}^{L} \Delta(z_l)\mathbf{1}(X = 0, Z = z_l).$$

We can conceive a multivariate version of the dummy test in which, instead of adding only the dummy $\mathbf{1}(X = 0)$ to the model, we add all the dummies $\mathbf{1}(X = 0, Z = z_l)$, $l = 1, \ldots, L$, and perform a joint

$\gamma_0 = (\beta, \lambda')'$ in this model.

$F-$test of whether the coefficients of these dummies are all equal to zero. Thus, in the model above, the coefficients of the dummies estimate the $\Delta(z_l)$. In a situation where $\Delta(z_l)$ differs enough across the $z_l$, this multivariate test will prove to be more powerful than the univariate test, as shown in the Monte Carlo simulations.

In general, when $Z$ has arbitrary support $\mathcal{Z}$, partitioned into subsets $\mathcal{Z}_1, \dots \mathcal{Z}_L$, we propose testing Assumption 1 by including the dummies $\mathbf{1}(X = 0, Z \in \mathcal{Z}_1), \dots, \mathbf{1}(X = 0, Z \in \mathcal{Z}_L)$ in the regression and performing an $F$-test of whether the coefficients of the dummies are all equal to zero.

The fundamental rank condition for this approach is

**Assumption 5.** *$\mathbb{E}[(W', \mathbf{1}(X = 0, Z \in \mathcal{Z}_1), \dots, \mathbf{1}(X = 0, Z \in \mathcal{Z}_L))(W', \mathbf{1}(X = 0, Z \in \mathcal{Z}_1), \dots, \mathbf{1}(X = 0, Z \in \mathcal{Z}_L))']$ is invertible.*

This rank condition is testable and indirectly requires that $0 < \mathbb{P}(X = 0, Z \in \mathcal{Z}_l) < 1$ for all $l = 1 \dots, L$. As with the simple dummy test, this version is also technically identical to a specification test, so the same assumptions and results applicable in the simple case also apply here.

The heterogeneity of the discontinuities in the covariates can be leveraged not only by using multiple dummies, but also by interacting the dummy of the bunching point with functions of the controls. For example, letting $Z_1$ be a non-binary element in the vector $Z$, one may add $Z_1 \cdot \mathbf{1}(X = 0)$ and $Z_1^2 \cdot \mathbf{1}(X = 0)$ to the regression. Moreover, both of these approaches (interacting and multiple dummies) may be combined.

Although the use of multiple dummies can increase the power in some cases, it may also lead to a less powerful test if some of the $\mathbb{P}(X = 0, Z \in \mathcal{Z}_l)$ are too close to 0 (Assumption 5), or if there is not enough heterogeneity in the correlation between confounders and $Y$ at $X = 0$ across different values of $Z$. Thus, in practice, a targeted division of the support on a few characteristics for which heterogeneity in the discontinuities is high is likely to better balance the gain in power from the heterogeneity with the possible loss in precision by the inclusion of another dummy.

## D.2 Multiple bunching points

Figure 1 also shows evidence of some bunching in maternal hours worked at $X = 2,080$, which is the total number of yearly hours of someone who works 40 hours per week every week of the year. When there is a second bunching point, the dummy of that point can also be added to the regression, and a joint test that the coefficients of both dummies are equal to zero performed. This test is similar to the test in the previous section, and the technical results are established analogously. The approach can be also immediately extended to cases when there are more than two bunching points.

The power of the joint test is particularly higher when the confounder has sufficiently different correlations with $Y$ at the different bunching points.[23] However, there are some instances in which the single dummy test performs better. If the size of the discontinuity in the confounders in the second bunching point is small, or if there is little bunching, then the increase in power from detecting such confounder may not compensate for the loss of power resulting from the inclusion of the additional

[23] This is analogous to the previous extension, where more power is obtained when the confounder has sufficiently different correlations with $Y$ for different values of $Z$. In this case, the heterogeneity is along different values of $X$ instead of $Z$. See the Monte Carlo simulation results in Section F, specifically Panel (d) in Figure 3.

dummy. To clarify this point, consider the modification of equation (2) for the case with the two bunching points in our application:

$$\mathbb{E}[\varepsilon|X,Z] = \Gamma(X,Z) + \Delta_1(Z)\mathbf{1}(X=0) + \Delta_2(Z)\mathbf{1}(X=2,080).$$

The coefficient of each dummy depends on the magnitude of the $\Delta$ function which multiplies it. However, the variances of these estimators depend on both $P(X=0)$ and $P(X=2080)$. The magnitude of $\Delta_2(Z)$ must be large enough to make it worthwhile to add a second dummy, particularly if $P(X=2080)$ is small.

In our application, $P(X=2,080) \approx 0.03$, which is small but sufficient for testing given our sample. Nevertheless, the empirical evidence is that $\Delta_2(Z)$ is very small. Figure 2 shows no discontinuities at $X=2,080$ in any of the plots. In fact, as shown in the top left panel, there seems to be no discontinuity in the outcome (verbal score) at $X=2,080$ either, which is direct evidence that $\Delta_2(Z)\approx 0$. This is in stark contrast to the discontinuities at $X=0$. Therefore, we conclude that it is better in our application to test using only the bunching point at $X=0$.

# E Linear CDT details

In Section 4, we introduced a parametric version of Caetano (2015)'s Discontinuity Test, which we named Linear CDT. Here we provide the details.

Define $\mathbf{y}_+ = (Y_1\mathbf{1}(X_1>0),\ldots,Y_n\mathbf{1}(X_n>0))'$, $\mathbf{y}_0 = (Y_1\mathbf{1}(X_1=0),\ldots,Y_n\mathbf{1}(X_n=0))'$, $\mathbf{x}_+, \mathbf{z}_+$ and $\mathbf{z}_0$ the matrices with rows equal to $(\mathbf{1}(X_i>0), X_i)$, $\mathbf{1}(X_i>0)Z_i'$ and $\mathbf{1}(X_i=0)Z_i'$ respectively. Supposing the appropriate rank conditions hold, then

$$\hat{\theta}_{LCDT} = e_1'(\mathbf{x}_+'\mathbf{x}_+)^{-1}\mathbf{x}_+'[\mathbf{z}_+(\mathbf{z}_0'\mathbf{z}_0)^{-1}\mathbf{z}_0'\mathbf{y}_0 - \mathbf{y}_+], \tag{9}$$

where $e_1 = (1,0)'$.

Let $\boldsymbol{\epsilon}_+ = (\epsilon_1\mathbf{1}(X_1>0),\ldots,\epsilon_n\mathbf{1}(X_n>0))'$ and $\boldsymbol{\epsilon}_0 = (\epsilon_1\mathbf{1}(X_1=0),\ldots,\epsilon_n\mathbf{1}(X_n=0))'$, if Assumption 1 holds, then

$$\hat{\theta}_{LCDT} = e_1'(\mathbf{x}_+'\mathbf{x}_+)^{-1}\mathbf{x}_+'[\mathbf{z}_+(\mathbf{z}_0'\mathbf{z}_0)^{-1}\mathbf{z}_0'\boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon}_+].$$

Let $\Sigma_+ = \mathbb{E}[\boldsymbol{\epsilon}_+\boldsymbol{\epsilon}_+'|\mathbf{x},\mathbf{z}]$, $\Sigma_0 = \mathbb{E}[\boldsymbol{\epsilon}_0\boldsymbol{\epsilon}_0'|\mathbf{x},\mathbf{z}]$, and $\Omega_0 = (\mathbf{z}_0'\mathbf{z}_0)^{-1}\mathbf{z}_0'\Sigma_0\mathbf{z}_0(\mathbf{z}_0'\mathbf{z}_0)^{-1}$. Then,

$$Var(\hat{\theta}_{LCDT}|\mathbf{x},\mathbf{z}) = e_1'(\mathbf{x}_+'\mathbf{x}_+)^{-1}\mathbf{x}_+'\left(\Sigma_0 + \mathbf{z}_+\Omega_0\mathbf{z}_+'\right)\mathbf{x}_+(\mathbf{x}_+'\mathbf{x}_+)^{-1}e_1.$$

The first term $(e_1'(\mathbf{x}_+'\mathbf{x}_+)^{-1}\mathbf{x}_+'\Sigma_0\mathbf{x}_+(\mathbf{x}_+'\mathbf{x}_+)^{-1}e_1)$ is the Eicker-White variance of the constant in a regression of $\mathbf{z}_+\lambda - \mathbf{y}_+$ onto $\mathbf{x}_+$ if $\lambda$ were known. The second term $(e_1'(\mathbf{x}_+'\mathbf{x}_+)^{-1}\mathbf{x}_+'\mathbf{z}_+\Omega_0\mathbf{z}_+'\mathbf{x}_+(\mathbf{x}_+'\mathbf{x}_+)^{-1}e_1)$ is the penalty due to the fact that $\lambda$ is in fact estimated, where $\Omega_0$ is the Eicker-White covariance matrix of the coefficients estimated in the first-step regression.

Let $SE(\hat{\theta}_{LCDT})$ be the square-root of the estimator of $Var(\hat{\theta}_{LCDT}|\mathbf{x},\mathbf{z})/n$. The test statistic is therefore $\hat{\theta}_{LCDT}/SE(\hat{\theta}_{LCDT})$. Under standard assumptions that allow the application of a Central Limit Theorem, such as the existence of moments and independence or stationarity of the data, this statistic can be compared to the standard normal distribution.

# F Monte Carlo

We perform a set of Monte Carlo simulations to contrast the finite-sample properties of the dummy test, the multivariate dummy test discussed in Section D.1, and Caetano (2015)'s discontinuity test (CDT).

## F.1 Set up

For each of the 5,000 iterations of the Monte Carlo, we draw $(Z, \eta, \epsilon)$ randomly $N$ times, where $(Z, \eta) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$ and $\epsilon \sim \mathcal{N}(0, 1)$. Next, we define $X$ and $Y$ as follows:

$$\begin{aligned} X^* &= 1 + .5Z + \eta \\ X &= \max\{X^*, 0\} \\ Y &= 2 + X + \phi X^2 + 2Z + (\mu + \rho \mathbf{1}(Z \leq 0) - \rho \mathbf{1}(Z > 0))\eta + \epsilon \end{aligned} \tag{10}$$

This specification yields bunching rates of around 25% for each iteration, which is approximately the bunching rate of maternal labor supply in the empirical application (Figure 1).

In all cases, we compare the performance of three tests:

1. Univariate dummy test: we run a linear regression of $Y$ on $X$, $Z$ and $\mathbf{1}(X = 0)$ and test whether the coefficient of $\mathbf{1}(X = 0)$ is equal to zero. This is the test discussed in Sections 2-3.

2. Multivariate dummy test: we run a linear regression of $Y$ on $X$, $Z$, $\mathbf{1}(X = 0, Z < 0)$ and $\mathbf{1}(X = 0, Z \geq 0)$ and jointly test whether the coefficients of $\mathbf{1}(X = 0, Z < 0)$ and $\mathbf{1}(X = 0, Z \geq 0)$ are equal to zero. This is the test discussed in Appendix D.1.

3. CDT: we perform Caetano (2015)'s discontinuity test assuming $Z$ enters the equation linearly but allowing $X$ to enter the equation nonparametrically.[24]

## F.2 Main results

We consider three sets of Monte Carlo experiments corresponding to different values of the parameters of equation (10), and show the results in Figure 3. For each of the three tests discussed above, we calculate the proportion of the 5,000 iterations for which we reject the null hypothesis at the 5% level of significance.

The first set of Monte Carlo simulations studies the size and power of the tests under no misspecification of the functional form of the effect of $X$ ($\phi = 0$) and under no heterogeneity in the effect of $\eta$

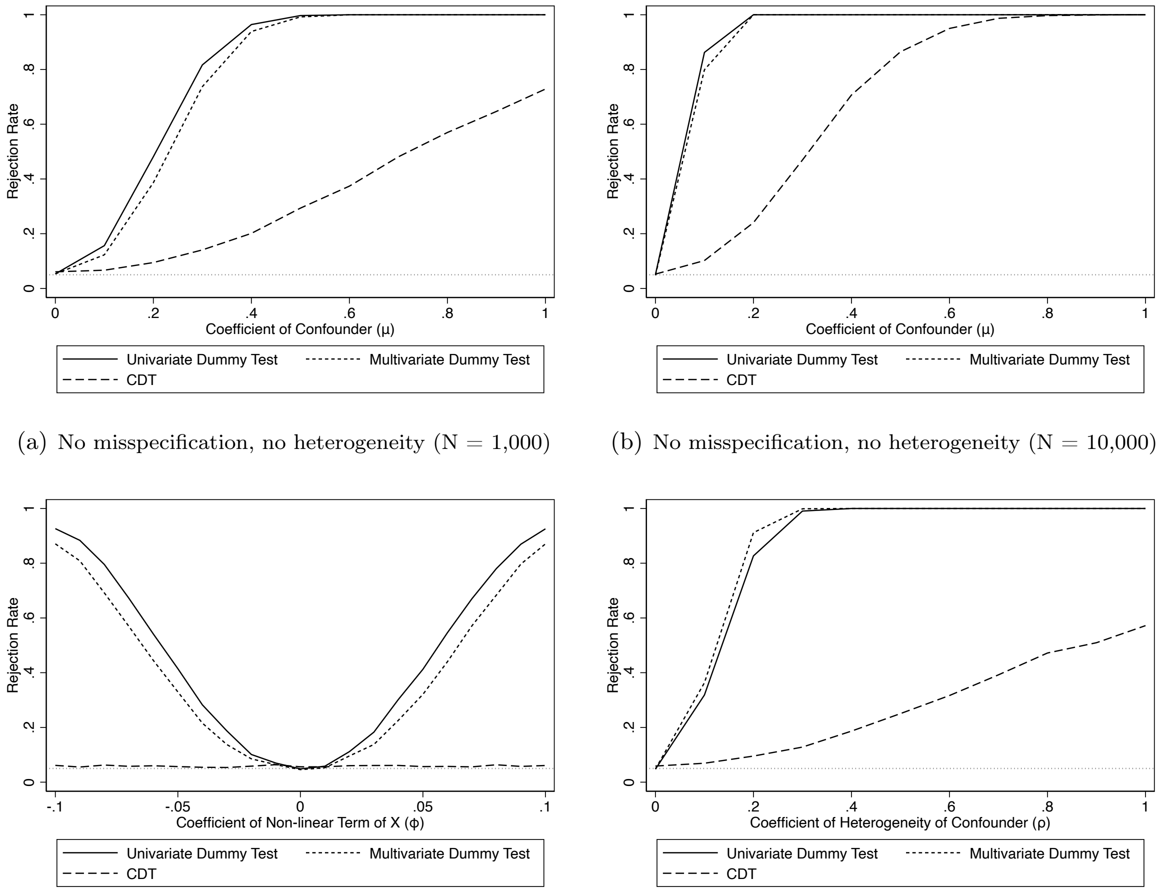
[24] We report results for the triangular kernel with bandwidths $h = 0.4$ for $N = 1,000$, and $h = 0.2$ for $N = 10,000$. On average, across all iterations, there are about 30 (150) observations in the bandwidth for $N = 1,000$ ($N = 10,000$). The triangular kernel is known to be the best kernel for boundary estimation, and the reported bandwidths are the largest that still yield a correctly sized test. Nevertheless, the relative performance of CDT and the dummy tests do not change using different kernels or bandwidths of any size. See, for example, Figure 4 in Appendix F.3 for the uniform kernel, infinite bandwidth case (which is equivalent to the Linear CDT).

on $Y$ ($\rho = 0$). Thus, the OLS estimate of the coefficient of $X$ could be biased only because the baseline endogeneity parameter $\mu$ is different from zero in equation (10). The Monte Carlo results are shown in Panels (a) ($N = 1,000$) and (b) ($N = 10,000$) of Figure 3. The point in the far left of the plots in both panels ($\mu = 0$) shows the size of the tests (i.e., the rejection rate under the null of no endogeneity or misspecification). As expected, for all tests, this rate is close to 5% (dotted line in each plot), thus showing that all three tests are correctly sized. As $\mu$ increases, the rejection rates increase for each test, but more so for the dummy tests. With a sample of $N = 1,000$, both dummy tests rejects the null 100% of the time at $\mu = 0.6$. In contrast, at that level of endogeneity, CDT rejects the null only 40% of the time.

The second set of Monte Carlo simulations assumes $\mu = \rho = 0$ (no endogeneity), and varies the influence of the quadratic term $X^2$, $\phi$, from negative (concave) to positive (convex) in equation (10). The results of this exercise can be seen in Panel (c) for $N = 1,000$. The rejection rates of CDT remain around 5%, as expected, since this test is designed to not detect misspecification. In contrast, the rejection rates of both dummy tests increase steeply as $\phi$ moves away from zero in either direction.

Note that $\rho = 0$ in Panels (a)-(c), and as expected the multivariate dummy test performs a little worse than the univariate dummy test in these cases. The third set of Monte Carlo simulations assumes $\phi = \mu = 0$ and varies $\rho$ in equation (10), thus allowing the influence of $\eta$ on $Y$ to be heterogeneous in $Z$. The results of this exercise can be seen in Panel (d) for $N = 1,000$, and, as expected, the multivariate dummy test performs better than the univariate dummy test.

Figure 3: Monte Carlo Results



(a) No misspecification, no heterogeneity (N = 1,000)

(b) No misspecification, no heterogeneity (N = 10,000)

(c) Misspecification, no endogeneity (N = 1,000)

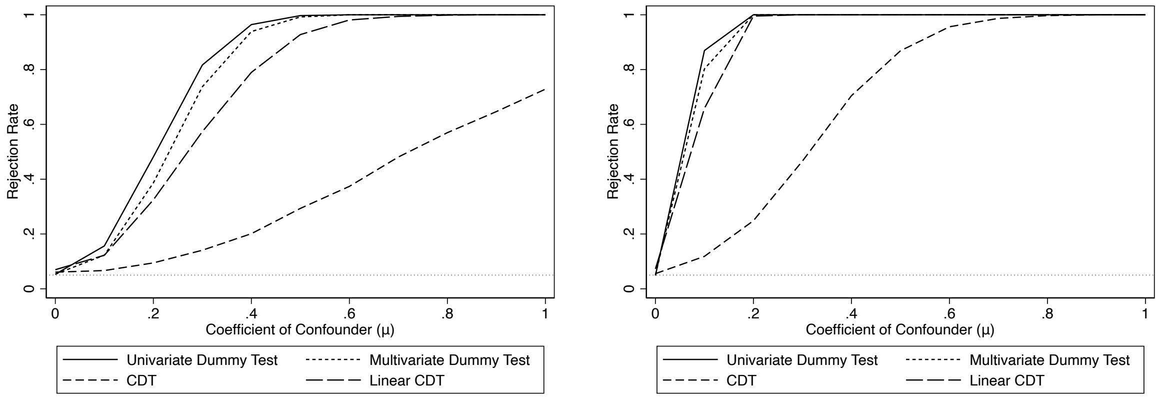
(d) No misspecification, heterogeneity (N = 1,000)

Note: Each panel compares the rejection rates of the null hypothesis that Assumption 1 holds for Caetano (2015)'s discontinuity test (CDT) as well as the univariate (Sections 2-3) and multivariate (Section D.1) dummy tests. These panels make different assumptions about the parameters of equation (10). Panels (a) and (b) assume no misspecification nor heterogeneity in the endogeneity ($\phi = \rho = 0$), with $N = 1,000$ in Panel (a) and $N = 10,000$ in Panel (b). Panel (c) assumes only misspecification and no endogeneity ($\mu = \rho = 0$), with $N = 1,000$. Panel (d) assumes no misspecification and no baseline endogeneity ($\phi = \mu = 0$), but varies the degree of heterogeneity in the endogeneity ($\rho$), with $N = 1,000$. For each parameter value, we perform 5,000 Monte Carlo iterations.

## F.3 Comparison with the Linear CDT

For completeness, we also compare the Linear CDT, discussed in Section 4 and Appendix E, with the dummy tests and CDT. The results can be seen in Figure 4, which consider the case with no misspecification nor heterogeneity in the endogeneity ($\phi = \rho = 0$ in equation (10)). As expected, the Linear CDT is more powerful to detect endogeneity than the CDT, because of its parametric rate of convergence, yet the Linear CDT is still less powerful than the dummy tests.

Figure 4: Monte Carlo Results



(a) No misspecification, no heterogeneity (N = 1,000)

(b) No misspecification, no heterogeneity (N = 10,000)

Note: Each panel compares the rejection rates of the null hypothesis (that Assumption 1 holds) for Caetano (2015)'s discontinuity test (CDT), the Linear CDT (Section 4 and Appendix E) and the univariate (Sections 2-3) and multivariate (Section D.1) dummy tests. Both panels assume no misspecification nor heterogeneity in the endogeneity ($\phi = \rho = 0$ in equation (10)), with $N = 1,000$ in Panel (a) and $N = 10,000$ in Panel (b).