Size and power of difference-in-differences studies in financial economics: An approximate permutation test*

Sebastian Bunnenberg†
Institute of Banking and Finance
Leibniz University Hannover

Steffen Meyer
Institute of Money and International Finance
Leibniz University Hannover

December 22, 2016

Abstract

Researchers use difference-in-differences models to evaluate the causal effects of policy changes. As the empirical correlation across firms and time is usually unknown, estimating consistent standard errors is difficult and statistical inferences may be biased. We suggest an approximate permutation test using simulated interventions to reveal the empirical error distribution of estimated policy effects. In contrast to existing econometric corrections, such as single- or double-clustering, our approach does not impose any parametric form on the data. In comparison to alternative parametric tests, our procedure maintains correct size with simulated and real-world interventions. Simultaneously, it improves power.

JEL classification: C33, C54, E61, G18

Keywords: Difference-in-differences, policy studies, clustered standard errors, approximate permutation tests.

*We thank Amit Goyal, Alexandra Niessen-Ruenzi, Matthias Rumpf, Marta Szymanowska, and Stephen Taylor for constructive feedback and helpful comments.

†Corresponding author. *E-mail address:* sebastian.bunnenberg@finance.uni-hannover.de.

1. Introduction

"Political economy is the science which prescribes rules and regulations for such a production, distribution, and consumption of wealth as to render the citizens good and happy."

Ely (1886, p. 531)

Difference-in-differences (DID) models are currently among the most popular methods in economic research: In a textual analysis, The Economist (2016) found "difference-in-differences" to be one of the most frequently used key word in the abstracts of all National Bureau of Economic Research (NBER) working papers since 2010. In the same period, more than 130 studies—roughly 5% of the total output—published in the top three finance and financial economics journals[1] apply a DID model. For empirical research, such as policy evaluation, DID models are supposed to correctly discriminate ineffective from effective interventions, i.e. combine correct size with high power.[2]

Bertrand et al. (2004) are among the surprisingly few who scrutinize the discriminatory abilities of DID models. Imposing effective and ineffective simulated interventions (laws) on Current Population Survey (CPS) data from 1979 to 1999, they find reliable inferences in terms of size, provided they collapse the data, block bootstrap the data or use appropriately clustered standard error (SE) estimates. Petersen (2009) and Thompson (2011) suggest to cluster SE by multiple dimensions simultaneously, such as by firm and time, which allows to keep the panel structure of the data unaltered.[3] Using the same dataset as Bertrand et al. (2004), Cameron et al. (2011) find that SE double-clustered by state and year provide a correct size using data of individuals. The latter is a mandatory condition in empirical research.[4]

A review of all 137 DID articles published in the top journals in finance since 2010 (Table 1) shows a wide heterogeneity with respect to SE estimation in the literature. While cross-sectionally single-clustered SE are most frequently used, it is arguable if this estimator is truly dominant from a statistical point of view. What is more, it is unlikely that a "gold standard" SE estimator will ever evolve, since an unbiased SE estimation would require to correctly identify the residual correlation structure.

[Insert Table 1 near here]

[1] *The Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies*.

[2] As common, we define *size* as the ability to not reject the null hypothesis of ineffectiveness for ineffective policies and *power* as the ability to reject the null hypothesis of ineffectiveness for effective policies

[3] This does not apply to bootstrapped data, such as the bootstrap-t procedure or the wild bootstrap suggested by Cameron et al. (2008).

[4] In financial economics, Siegel and Choudhury (2012), for instance, criticize the SE estimator used by Bertrand et al. (2002), because it ignores serial correlation and overestimates statistical significance. Gow et al. (2010) find a similar issue in their review of empirical studies in accounting.

However, a high power may be just as important: Richard T. Ely (1886, Science)—a founding father of the American Economic Association—defines economics as an active force in society. Economists should assume an active role, e.g. by suggesting and evaluating interventions to achieve welfare-enhancing equilibria.[5] Evaluating truly effective interventions to be ineffective may likewise destroy welfare. Previous results on power, though, are mostly disappointing: Given that size was correctly estimated, on average only one in three effective laws could be correctly identified in Bertrand et al. (2004).

We suggest a simple approximate permutation test for DID settings. As key contribution, this test avoids imposing any parametric form on the data, and thereby eliminates the danger of inconsistent SE estimates. Our testing procedure extracts the error distributions of estimated law effects by imposing permuted placebo laws on the dependent variable.[6] We find that our test exhibits dominant discriminatory abilities between simulated ineffective and effective laws when compared to parametric procedures based on White (1980), single-clustered SE, and double-clustered SE. We finally replicate an analysis on the impact of U.S. Securities and Exchange Commission (SEC) regulation SHO on stock returns by Diether et al. (2009) and verify that our test yields consistent results.

We also complement the findings of Bertrand et al. (2004) and Cameron et al. (2011) using a financial markets dataset. Our sample are monthly total returns of all stocks between January 1970 and December 2014 from the Center for Research in Security Prices (CRSP) US Stock Database. The advantage of using financial data over CPS data, for example, is that we observe monthly data over a long term per stock, and a clearly identified panel structure. We find that size and power strongly depend on the parametric form placed on the data. Cross-sectional aggregation of data yields conservative—often even overcritical—size estimates, but with disappointing results on power. This implies that parametric testing procedures in DID settings have a limited ability to correctly identify effective interventions.

Finally, we demonstrate how factor models of stock returns can be integrated in DID settings. This is a particular issue in financial economics: The computational effort may easily become huge, as it grows with the number of factors in the market model and cross-sectional units in the dataset. As such, simultaneously estimating a DID-model and factor loadings for each unit may become computationally infeasible. We two alternative two-stage settings that require drastically less computing time and that improve the efficiency

[5] Two examples emphasize the importance of these tasks: The European Commission institutionalized the "better regulation initiative", see http://ec.europa.eu/smart-regulation/index_en.htm, and the Securities and Exchange Commission ran a controlled trial before eliminating all short sale price tests, see Diether et al. (2009) and Fang et al. (2016).

[6] This procedure shares some conceptual ideas with procedures financial economists have lately applied to separate skill from luck in the performance measurement of mutual fund managers, see Kosowski et al. (2006) and Fama and French (2010).

of estimated law effects. We find that using factor model estimations as control variables increases power while maintaining a correct size in our approximate permutation test. In total, our study provides guidance on how to systematically improve inference statistics and hypothesis testing in DID settings while maintaining a correct size.

The remainder of this paper is organized as follows: in section 2, we describe our approximate permutation test; in section 3, we use simulated placebo laws to identify efficient DID specifications and compare our test with various parametric tests using different SE estimators; in section 4, we derive two alternative ways to integrate factor models into DID settings, and empirically analyze these improvements in terms of size and power; in section 5, we present the application of our test to SEC regulation SHO and section 6 concludes.

2. An approximate permutation test for DID models

Whenever researchers apply DID models, they usually only observe a dependent variable that is—at least potentially—affected by a law to be tested. In our notation, we use the superscript $^+$ to denote this quality, such as in Y_{it}^+. Any DID model on Y_{it}^+ contains at least three variables: a cross sectional dummy D_i^{law}, which is 1 in all periods for those firms that are affected by the law and 0 otherwise; a time dummy D_t^{law}, which is 1 for all firms in those periods in which the law is in effect and 0 otherwise; and the law variable $D_i^{law} \times D_t^{law}$, which is the interaction of the first two variables:

$$Y_{it}^+ = a + D_i^{law} + D_t^{law} + \lambda D_i^{law} \times D_t^{law} + \varepsilon_{it} \tag{1}$$

We further use the excess return er_{it}^+ of stock i in month t as dependent variable Y_{it}^+. Accordingly, constant a captures the mean excess return over all firms that are not affected by the law, in all months before the law is introduced. The dummy D_i^{law} (D_t^{law}) controls for unobserved heterogeneity between firms (months) affected by the law and those that are not. The coefficient λ of the interaction term $D_i^{law} \times D_t^{law}$ measures the impact of the law on the dependent variable, and its statistical significance determines if we consider a law to be effective or not.

Let Y_{it}^c denote an ideal counterfactual of Y_{it}^+. Such a conterfactual contains the realizations of the dependent variable if the law to be tested had never been discussed nor introduced. Usually, researchers are unable to observe Y_{it}^c. Therefore, we estimate it as

$$\hat{Y}_{it}^c = Y_{it}^+ - \hat{\lambda} D_i^{law} \times D_t^{law} \tag{2}$$

$\hat{Y}_{it}^c$ is obtained by subtracting an estimated law impact $\hat{\lambda}$ from those observations of Y_{it}^+

that are subject to the law to be tested. Any estimation error in $\hat{\lambda}$ will also affect $\hat{Y}_{it}^{c}$.[7] Yet, the in-sample effect of the law to be tested on $\hat{Y}_{it}^{c}$ is zero by definition. Our further procedure relies on this particular quality.

The SE estimators of White (1980) and Newey and West (1987) as well as any clustered SE estimator impose a particular correlation structure onto the residuals of Equation 1. However, the true correlation structure in ε_{it} is typically unknown. Therefore, researchers do not know which SE estimator best represents the empirical correlation structure and provides unbiased SE estimates. The choice of any SE estimator therefore entails placing a particular parametric form on the data, and statistical significance of the estimated law effect λ may highly vary with the clustering strategy, requiring researchers to interpret oppositional test outcomes.

We propose a non-parametric approach for inferential tests of law effects that imposes no specific restriction on the data while fully maintaining its structure.[8] Our approach is an approximate permutation test as suggested by Dwass (1957). Kosowski et al. (2006) and Fama and French (2010) use a similar approach for inferences on the measured skill of mutual fund managers. A related technique has been applied by Huang (2008) in an out-of-sample context. In contrast to these studies, we permute over the law dummies D_i^{law} and D_t^{law} and estimate their effects on Y_{it}^{c}.

Let the sample used to estimate Equation 1 contain observations of Y_{it}^{+} from a set of cross-sectional units f, e.g. firms, over a set of periods p. We define these sets and their cardinalities to be

$$F = \{f_1, f_2, \ldots, f_i, \ldots, f_N\} \qquad P = \{p_1, p_2, \ldots, p_t, \ldots, p_T\}$$
$$|F| = N \qquad |P| = T$$

Any law to be tested is represented by the interaction of two dummies D_i^{law} and D_t^{law}. Both dummies equal one for each a subset of the N cross-sectional units and the T time periods. We define these two subsets as follows:

$$F^{law} = \{f_i | D_i^{law} = 1\} \subset F \qquad P^{law} = \{p_t | D_t^{law} = 1\} \subset P$$
$$|F^{law}| = M \qquad |P^{law}| = S$$

For our testing procedure, we generate subsets F_k^{sim} and P_k^{sim} from the sets F and P and

[7] In Section 3, we suggest a statistical approach to identify efficient specifications based on Equation 1 that include FE on a more granular level, such as firm or month FE.

[8] Another commonly used non-parametric approach to estimate SE is to bootstrap the model. However, this only allows maintaining one dimension of the data, either the cross-section or the time.

define the dummy variables $D^{sim}_{i(k)}$ and $D^{sim}_{t(k)}$ as

$$F^{sim}_k \subset F \ \forall \ k = 1, 2, \ldots, K \qquad P^{sim}_k \subset P \ \forall \ k = 1, 2, \ldots, K$$
$$D^{sim}_{i(k)} = 1 \ \forall \ f_i \in F^{sim}_k \qquad D^{sim}_{t(k)} = 1 \ \forall \ p_t \in P^{sim}_k$$

So far, we have not specified how to generate the subsets F^{sim}_k and P^{sim}_k. Basically, researches face only weak restrictions in answering this question, and our approach allows for answers that are specific for the research question at hand and that depend on the nature of the law to be tested. However, some advice can be given: For F^{sim}_k and randomized trials, it should usually suffice to randomly draw M firms out of F without replacement for each k. If researchers can identify any structure in the formation of the treatment and control group, however, it may be appropriate to adjust the permutation procedure for F^{sim}_k accordingly. For P^{sim}_k, one may either assume that simulated laws are always in effect during the same periods as the law to be tested, which implies $D^{sim}_{t(k)} = D^{law}_t \ \forall \ k, t$. Alternatively, the kth simulated law may be assigned a randomly drawn start date $s_{(k)} \in P$ and we have $D^{sim}_{t(k)} = 1 \ \forall \ t \geq s_{(k)}$.[9]

Finally, we estimate the following model to quantify the impact of the kth simulated law:

$$\hat{Y}^c_{it} = a_{(k)} + D^{sim}_{i(k)} + D^{sim}_{t(k)} + \kappa_{(k)} D^{sim}_{i(k)} \times D^{sim}_{t(k)} + \eta_{it(k)} \tag{3}$$

Ideally, one could sample all possible permutations of the subsets F^{sim}_k and P^{sim}_k, which would allow for an exact significance test. However, this procedure frequently requires an impossible effort: For 100 firms of which 50 are affected by a law to be tested, there already exist more than 10^{29} permutations.

We therefore suggest to sample only a part of the possible permutations. Precisely, we reveal the distribution of the estimation error of λ in Equation 1 through $K = 5,000$ different simulated laws and estimations of their effect $\kappa_{(k)}$. Under the assumption that $E(\kappa_{(k)}) = 0 \ \forall \ k$, the estimation of Equation 3 exclusively reflects estimation errors. We can then conduct a hypothesis test by comparing the estimated effect of a law to tested with the distribution obtained from the simulated laws. If the coefficient of a law to be tested is outside the 126th smallest and 126th highest value of these 5,000 coefficients, we consider the tested law to be statistically significant at the 5%-level.

[9] For example, one could draw s from a discrete uniform distribution $U\{a, b\}$ with $a = 1 + \Delta$, $b = T - \Delta$, and $\Delta < 0.5 \cdot T$ such that the simulated laws start during a window centered within the sample and at least Δ periods away from the beginning and the end of the sample.

3. Discriminatory abilities of DID models with simulated placebo laws

3.1. The dataset

Our dataset is obtained from the CRSP US Stock Database. We include monthly total returns of all stocks between January 1970 and December 2014 in our sample, and exclude all stocks with negative prices, stocks with a minimal price equal or below USD 5 as penny stocks, stocks whose returns are defined to be the 100 most extreme return observations in absolute value, and stocks with less than 60 return observations. Finally, we map each stock to one of 49 industry sectors, as defined by French (2016), and exclude all stocks that belong to the sector "other" as this sector contains very heterogeneous businesses. The final sample consists of 575,621 monthly stock returns from 3,230 companies. Our market portfolio is the CRSP value weighted portfolio of all NYSE, AMEX, and NASDAQ stocks. This data and that on additional market factors (SMB, HML, and MOM) are from French (2016), who follows Fama and French (1993) and Carhart (1997). We also use the monthly returns of 49 value-weighted industry portfolios from the same source to analyze how data aggregation impacts estimated law effects and their inferences.

3.2. Simulating placebo and effective laws

To simulate laws, we assign each stock to one of 49 industry sectors as defined by French (2016) according to the stock's SIC code, excluding the sector "other" and all stocks therein. Subsequently, we sample 24 of the 48 remaining industries without replacement, which are affected by the law, while the remaining industries are not. The start date of the law is randomly drawn from a uniform distribution between January 1985 and December 1999 for every industry (random start). Finally, the law may have no effect on the returns (placebo law), or an additive effect (effective law): for each return observation that is affected by the law, we draw the effect from a normal distribution with mean 2% p.a. and standard deviation of 0.5% p.a.

$$er_{it}^{+} = er_{it} + \left[(1+\xi_{it})^{\frac{1}{12}} - 1\right] D_i^{law} \times D_t^{law}, \tag{4}$$

where $er_{it} = r_{it} - r_{ft}$ is the return r of stock i in month t in excess of the contemporaneous risk-free rate r_{ft}, and er_{it}^{+} is the excess return from including the impact of simulated (placebo or effective) laws. The interaction term $D_i^{law} \times D_t^{law}$ is 1 if stock i is affected by the law in month t and 0 otherwise. Variable ξ_{it} captures the random impact of the simulated law

on firm i in month t: 0 for placebo laws and an i.i.d. draw from a normal distribution $N(0.02, 0.005^2)$ for effective laws. Therefore, for a placebo law, we have $er_{it}^{+} = er_{it}\ \forall\ i, t$. As we simulate interventions, we are able to observe both the depend variable including the law effect, er_{it}^{+}, and its ideal counterfactual, $er_{it}^{c} = er_{it}$. We stress that this is usually not the case in empirical research.

3.3. Efficient model specification from a statistical perspective

The specification in Equation 1 can be altered in several ways. As an alternative control for unobserved cross-sectional heterogeneity, one might substitute constant a and the dummy variable D_i^{law} with industry FE.[10] In the time dimension, D_t^{law} can be substituted with year dummies as an alternative control for unobserved heterogeneity over time. Finally, a control variable for market climate can be included, such as the excess return of the market, $er_{mt} = r_{mt} - r_{ft}$.[11] These different options result in eight potential specifications for the DID model.

We choose the baseline model for our further analyses these specifications by the efficiency, as indicated by the dispersion of estimated law impacts. For this purpose, we simulate 500 placebo laws and estimate all eight DID specifications for each of them. As suggested by Bertrand et al. (2004), we apply this procedure to stock excess returns as well as to the excess returns of 48 value-weighted industry portfolios as per French (2016). Table 2 reports descriptive statistics of the estimated law coefficients using firm level data in Panel A and industry level data in Panel B. We provide details on the respective specifications in the three bottom rows of Table 2.

[Insert Table 2 near here]

All specifications reported in Table 2 are unbiased, as the mean and the median of the estimated coefficients are close to 0. At firm level, the specifications strongly differ with respect to the shape of the error distribution: industry FEs are crucial to obtaining normally distributed estimates. For industry level data, this does not apply. Otherwise, we find the distributions to be mostly symmetrical and with limited excess kurtosis.

Aggregating the data to industry level does not necessarily improve efficiency of the estimated law coefficients. We observe the smallest dispersion of estimated law effects for a

[10] We refrain from adding firm and/or month FE, because they would drastically increase the computational effort to estimate Equation 1. One could absorb firm and/or month FE by demeaning the model before the estimation. However, the according estimates are less efficient than those we present here. We gladly provide results upon request.

[11] Adding the factors of Fama and French (1993) and Carhart (1997) in addition to er_{mt} decreases the efficiency of the estimated law effects due to all these factors being multiple controls for time FE.

DID model using industry, and year FE on the level of stocks. In this setting, the dispersion of the effects is 25 bp less than the lowest dispersion found at industry level in Panel B.

Overall, including industry FE and year FE in the DID specification supports correct inferences by reducing the estimation error of the law coefficient. Adding er_m as a control variable, however, seems to be of minor importance. In summary, we find that a DID model including industry and year FEs performs best on the firm and industry levels simultaneously. While this conclusion stems from a strictly econometric perspective and not from economic reasoning, we still require a baseline specification for all further tests. Using the most efficient one here seems sensible.

3.4. Comparing inferential tests of estimated law effects

In this section we compare how our approximate permutation tests performs in terms of size and power compared to single and double clustered SE. Our baseline model follows directly from our previous results: the LHS-variable is er_{it}^{+}, the excess return of stock i in month t including the effect of simulated laws. We control for industry fixed effects by including industry dummies, D_i^{ind}, and for time fixed-effects by including year fixed-effects, D_t^{yr}, and include the law variable $D_i^{law} \times D_t^{law}$:

$$er_{it}^{+} = D_i^{ind} + D_t^{yr} + \lambda D_i^{law} \times D_t^{law} + \varepsilon_{it} \tag{5}$$

The coefficients in Equation 5 are estimated by OLS.[12] The feasible SE estimator depends on the cross- and autocorrelation in er_{it}. We conduct significance tests using heteroskedasticity robust SEs, according to White (1980), single clustered SEs clustered by firm, by industry, by month, and by industry-month cell, as well as double clustered standard errors clustered by firm and month, and by industry and month.[13] Again, we use 500 simulated placebo and effective laws, and to each we apply all of these procedures and compare their power and size in Table 3. We also compute implicit effect levels for a correct power in one-sided tests at the 5% level, that is, the required additive effect in percent p.a. to reject the H_0 of $\lambda_0 \leq 0$ in 95% of the cases. Finally, we also run approximate permutation test and compute the implicit effect level for this test as well.

[Insert Table 3 near here]

At firm level, the rejection rates for placebo laws confirm the issue on size documented

[12] In certain applications, other estimators such as maximum likelihood may be more adequate. We focus on OLS because it is the most common choice and it is analytical.

[13] For an excellent review of the size of various bootstrap procedures, see Cameron et al. (2008).

in Bertrand et al. (2004): SEs single clustered by firm or industry or double clustered by industry and month too often reject the null hypothesis for placebo laws, though these laws are truly ineffective. However, and in contrast to Bertrand et al. (2004), SEs clustered by month, industry-month-cell, and double-clustered SEs by firm and month perform well in terms of size and may even be overly critical.

Table 3 also shows that the power of all procedures is not as good at firm level: those SE estimators that have a high power (i.e., close to 90%) are unreliable in terms of size, and those that provide a correct size identify only one in four truly effective laws as statistically significant. The required additive effect to achieve a power of 95% is more than 4% p.a. in these cases.

Similar to Bertrand et al. (2004), size improves if we aggregate the data, and is mostly not an issue at industry level. Although power does improve, it is still at less than 50%. The implicit effect remains 4% p.a. at least; a magnitude that may even not require sophisticated econometric techniques to be detected. Falsely rejecting effective policies may result in welfare losses comparable to those of falsely accepting an ineffective one.

Our approximate permutation test provides a superior ability to discriminate between placebo and effective laws compared against all SE estimators we apply. The size is 5% at firm and industry level, making the testing procedure adequately critical. At firm level, power more than doubles with an implicit effect level that is 1.6% p.a. lower, against the next best SE estimator that is at least correct on size. At industry level, power slightly decreases, while the implicit effect level is still reduced by 0.3% p.a.

4. Application 1: DID regressions and factor models

4.1. Estimating factor models in DID settings

To further improve the efficiency and, thereby, the power of our DID setting, we exploit knowledge on the return generating process of the dependent variable. A multivariate factor model, as proposed by Fama and French (1993) and Carhart (1997), is commonly used to explain the variation of stock returns. For the unaltered excess returns, er_{it}, this model is:

$$er_{it} = \alpha_i + \beta_{1i} er_{mt} + \beta_{2i} smb_t + \beta_{3i} hml_t + \beta_{4i} mom_t + u_{it} \tag{6}$$

Empirical researchers typically use Equation 6 for individual stocks, to estimate their exposure to non-diversifiable market risk (er_{mt}), as well as to other market factors that explain the cross-section of stock returns (smb_t, hml_t, and mom_t).

To correctly exploit the structure of such a factor model for a DID regression, one would

have to jointly estimate firm-specific coefficients $\alpha_i, \beta_{1i}, \ldots, \beta_{4i}$ and law effect λ by adding the respective interactions to the DID model. If dummy variable D_i identify single firms, the following model simultaneously produces these estimates:

$$\begin{aligned} er_{it}^{+} = D_i + D_i \times er_{mt} + D_i \times smb_t + D_i \times hml_t + D_i \times mom_t \\ + D_t^{law} + \lambda D_i^{law} \times D_t^{law} + \varepsilon_{it} \end{aligned} \tag{7}$$

However, this approach may quickly cause an unfeasible computational effort:[14] it requires $5 \cdot N + 2$ coefficients to be estimated, where N is the number of entities in the sample. For our dataset, we would have to estimate more than 16,000 coefficients for each DID-regression at firm level. This also induces the risk of overfitting and, thus, unstable results.

Therefore, we suggest a two-step procedure by first estimating the factor model for each firm separately and then employing this information in the DID model. However, if we desire to test if a given law is truly effective, the model in Equation 6 cannot be directly estimated for single stocks: an effective law constitutes a structural break for all firms affected by the law, which, in turn, biases the estimates of α_i. To correctly estimate the factor model for single firms, we refer to Equation 2 and eliminate any law effect in the returns er_{it}^{+}.

$$\hat{er}_{it}^{c} = er_{it}^{+} - \hat{\lambda} D_i^{law} \times D_t^{law} \tag{8}$$

$\hat{er}_{it}^{c}$ is the estimated excess return of stock i in month t, excluding the estimated effect of the law to be tested. $\hat{\lambda}$ is an OLS estimate of this effect from a DID model that includes year FE and industry FE as in Equation 5. A cleaned variable such as $\hat{er}_{it}^{c}$ contains no effect from the particular law used for "cleaning", and may, therefore, be used to estimate factor models. Next, we estimate Equation 6 using $\hat{er}_{it}^{c}$ as LHS variable:

$$er_{it}^{c} = \alpha_i^{c} + \beta_{1i}^{c} er_{mt} + \beta_{2i}^{c} smb_t + \beta_{3i}^{c} hml_t + \beta_{4i}^{c} mom_t + u_{it}^{c} \tag{9}$$

We suggest two different approaches to incorporate the estimates of Equation 9 in the DID model. In the first approach, we use the factor model to filter the excess returns er_{it}^{+}. We then estimate the DID model with filtered returns, η_{it}, as dependent variable:

$$\begin{aligned} \eta_{it}^{c} &= er_{it}^{+} - (\hat{\beta}_{1i}^{c} er_{mt} + \hat{\beta}_{2i}^{c} smb_t + \hat{\beta}_{3i}^{c} hml_t + \hat{\beta}_{4i}^{c} mom_t) \\ &= a + D_i^{law} + D_t^{law} + \lambda D_i^{law} \times D_t^{law} + \varepsilon_{it} \end{aligned} \tag{10}$$

[14] The computational effort further increases if one allows the factor loadings to change with the law. For the market excess return er_{mt}, this results in the interaction terms $D_i \times er_{mt} + D_i \times D_t^{law} \times er_{mt}$.

$\hat{\beta}^c_{1i}, \ldots, \hat{\beta}^c_{4i}$ are OLS-estimates of the according variables in Equation 9, fitted for each firm separately. The filtered returns η_{it} contain the variation in er^+_{it} that cannot be explained by static loadings against the market factors and include any level shift from the pre-law to the post-law period.

In the second alternative, we use the factor model in Equation 9 to estimate control variables for our DID setting (henceforth "factor-model controls approach").

$$\begin{aligned} \hat{rp}^c_{it} &= \hat{\beta}^c_{1i} er_{mt} + \hat{\beta}^c_{2i} smb_t + \hat{\beta}^c_{3i} hml_t + \hat{\beta}^c_{4i} mom_t \\ er^+_{it} &= \hat{\alpha}^c_i + \hat{rp}^c_{it} + D^{law}_i + D^{law}_t + \lambda D^{law}_i \times D^{law}_t + \varepsilon_{it} \end{aligned} \tag{11}$$

The coefficients $\hat{\alpha}_i, \hat{\beta}^c_{1i}, \ldots, \hat{\beta}^c_{4i}$ are OLS estimates of the respective coefficients in Equation 9. This second alternative induces an errors-in-variables problem on the RHS of Equation 11. However, given that conditional mean independence holds in Equation 9, this measurement error will only affect the variance of λ, i.e., the efficiency of our estimation, while the OLS-estimator $\hat{\lambda}$ converges in probability towards λ and is therefore, unbiased. The dominance of this model in empirical studies in finance and our results both indicate that this is indeed the case.[15]

In both alternatives, we want to compare the efficiency for different levels of FE. Our analysis starts with the simplest DID setting: for filtered returns, the baseline model considers cross-sectional FE for affected and unaffected stocks, D^{law}_i, and time FE before and after introduction of the law D^{law}_t. Both dummies can then be substituted by the more granular industry and year FE, separately or jointly. For the factor model controls approach, $\hat{\alpha}_i$ and $\hat{rp}_{it}$ together capture cross-sectional heterogeneity, and we drop D^{law}_i and disregard industry FE in this case. We still include D^{law}_t and allow it to be substituted with year FE. This results in six specifications to be tested. Table 4 presents descriptive statistics of the estimated coefficients of 500 placebo laws in each of these settings, estimated at firm level (Panel A) and at industry level (Panel B).

[Insert Table 4 near here]

As in Table 2, all DID settings estimate the effect of placebo laws, in an unbiased manner, with mean and median estimates close to 0. Furthermore, all settings are fairly normally distributed, as indicated by skewness and kurtosis. These findings apply to firm as well as to industry level.

Using filtered returns, η_{it} instead of excess returns, er^+_{it}, as dependent variable visibly

[15] Filtering the dependent variable as for Equation 10 also induces an error in variable bias. Here, this bias occurs in the dependent variable and has the same effect on the estimated law coefficients.

improves the efficiency of the DID estimations at firm level in three out of four specifications. As such, the standard deviation of the estimated coefficients decreases by at least 50 bp in columns (1) to (3) in comparison to Table 2. However, the DID model with industry and year FEs in column (4) performs equally well with excess returns as with filtered returns. At firm level, all the specifications in columns (1) to (4) provide a slightly better efficiency, although the differences are less pronounced.

Using alphas and factor risk premiums as control variables further improves efficiency at firm level: Here, the dispersion of the estimated placebo effects decreases once more to a level below that of all previous specifications. At industry level, these controls still provide a benefit over using er_{it}^{+} as dependent variable, with standard deviations decreasing by at least 8 bp. Compared to filtered returns, however, the latter provide a higher efficiency at industry level.

For filtered returns, we use the specification in column (4), which includes industry and year FEs. Not only does this model provide the highest efficiency in both panels, it is also fully consistent with the model used in Table 3. For factor model estimates, we proceed with column (6). While the year FE slightly reduces efficiency at firm level – although by an insignificant margin –, this specification is closer to the others than the one according to column (5).

4.2. Size and power of DID procedures integrating factor models

We now study how integrating factor models in the DID setting affects rejection rates of placebo and effective laws. We start with using filtered returns, η_{it}, instead of excess returns, er_{it}^{+}, as dependent variable. In the previous section, we have shown that filtering the dependent variable with a four factor model according to Carhart (1997) reduces the dispersion of the estimated coefficients of placebo laws. Accordingly, this should have a positive effect on the discriminatory ability of t-tests. In Table 5, we repeat the analysis on size and power of parametric and non-parametric inferences using the filtered returns as LHS variable.

[Insert Table 5 near here]

At firm level, filtering has little effect on size. While the other SE estimators reject the H_0 of ineffectiveness too often for placebo laws and, therefore, fail to meet scientific standards, SEs single clustered by month, industry-month cell, and SEs double clustered by firm and month provide an even overcritical size. However, the non-parametric test based on placebo laws provides a correct size.

Power increases by 8% to 28% for the overcritical SE estimators; however, it hardly exceeds 30%. The implicit effect level decreases by at least 60 bp for these SE estimators, and falls to 4% p.a. Inferences based on estimated placebo laws now perform slightly weaker in comparison to excess returns as dependent variable: power decreases by 14% and the implicit effect level increases by 18 bp. This drop in power results from inducing an estimation error in the dependent variable by filtering returns.

At industry level, size is hardly affected by filtering returns, as all SE estimators are at least close the correct size of 5%, if not overcritical. With respect to power, filtering increases the proportion of rejected ineffectiveness by at least 5% for all estimators. Implicit effect levels decrease by at least 40 bp, which brings them down to 3.5% p.a. for industry clustered SE. Hypothesis tests using the distribution of placebo law effect perform similarly well.

Finally, in Table 6 we study size and power of DID specifications with factor model estimates as controls, by including the alpha and the risk premium on the RHS.

[Insert Table 6 near here]

Using factor model estimates as controls has ambivalent consequences on our results: for some SE estimators, size decreases dramatically, such as for single clustered SE by firm at firm level, or single clustered SE by industry at industry level. At firm level, however, those SE estimators which were able to produce a correct size in the previous analyses still do so with factor model controls, namely single clustered SEs by month or industry-month-cell, as well as double clustered SE by firm and month. It is also worth noting that these estimators double in power, bringing implicit effect levels down to around 3% p.a. On the firm level, inferences based on the placebo laws show again a high ability to discriminate between truly ineffective and truly effective laws, with a power of 61% and an implicit effect level of 2.8% p.a.

The results are similar at industry level. Again, size decreases, while power often improves visibly. Inferences using placebo laws perform better than all other SE estimators, and provide a correct size with implicit effect levels of 3.5% p.a. In total, adding factor model variables that explain the dynamics of single stocks or industries over time greatly improves the ability of DID models to detect law effects at firm level.

5. Application 2: SEC Regulation SHO

In 2004, the Securities and Exchange Commission (SEC) announced a trial of a new regulatory framework concerning short-selling on US stock markets:[16] The SEC temporarily

[16]Securities Exchange Act Release No. 50104 (July 28, 2004), 69 FR 48032 (August 6, 2004).

suspended price tests for short sales of a randomly selected subset of the Russell 3000 companies. This real-life experiment has been studied in several studies, such as Diether et al. (2009) and Fang et al. (2016). Amongst other issues, Diether et al. (2009) investigate if the average returns of pilot and control stocks differ significantly during the announcement and the beginning of the trial using a DID model with SE clustered by firm and day. They find no significant differences on either the NYSE nor the NASDAQ.

We replicate this particular analysis to demonstrate how our approximate permutation tests performs in comparison to parametric tests using various SE estimators. For this purpose, we follow the sampling procedure of Diether et al. (2009) and select all stocks constituting the Russell 3000 index in 2004 and 2005 that are primarily traded either on the NYSE or on the NASDAQ. Among these, we identify the pilot stocks using the aforementioned SEC release. We sample daily returns from July 26, 2004 to May 4, 2005 for these stocks from CRSP and exclude stocks with incomplete return histories, penny stocks (i.e. stocks with a minimum price of less than 1 USD in that period) and large stocks (i.e. stocks with an average price of more than 100 USD in that period). Our final sample consists of 2,500 stocks and closely matches that of Diether et al. (2009) in proportions.[17]

Next, we further follow Diether et al. (2009) and define two dummy variables D_t^{ann} and D_t^{event} that capture the five trading day period around the announcement of the pilot program on July 28, 2004 and its effective start on May 02, 2005. A third dummy variable, D_i^{pilot}, captures if a stock is a pilot stock or not. To analyze the impact of the pilot program on the average stock returns during the announcement and event window, we estimate the following DID model by exchange:

$$r_{it} = a + b_1 \cdot D_t^{ann} + b_2 \cdot D_t^{ann} \times D_i^{pilot} + b_3 \cdot D_t^{event} + b_4 \cdot D_t^{ann} \times D_i^{pilot} \tag{12}$$

We apply our approximate permutation test to this setting by initially estimating stock returns that are free from any effect of the pilot program as proposed in Equation 2: We estimate Equation 12 and compute "clean" returns $\hat{r}_{it}^c$ as

$$\hat{r}_{it}^c = r_{it} - \hat{b}_2 \cdot D_t^{ann} \times D_i^{pilot} - \hat{b}_4 D_t^{ann} \times D_i^{pilot}$$

The coefficients $\hat{b}_2$ and $\hat{b}_4$ are OLS estimates of the according coefficients in Equation 12 using our original sample. We then randomly draw 833 out of the 2,500 stocks in our sample and consider these as pilot stocks. Next, we estimate Equation 12 using $\hat{r}_{it}^c$ as LHS variable and the randomly drawn pilot stocks, keeping the days of the announcement and the event unchanged. We repeat this process 5,000 times and use the resulting data to reveal the

[17] We gladly provide additional information on the sample composition upon request.

distribution of the estimation error in the relevant coefficients b_2 and $b4$ and to conduct hypothesis tests.

In Table 7, we compare the estimates on average returns of pilot and control stocks during the announcement and event window, the differences between pilot and control stocks, and the p-values of these differences. The p-values have been estimated with various SE estimators as indicated in the first column and our approximate permutation test. We also include the numerical results in Table IV of Diether et al. (2009) to document that our sample closely resembles theirs.

[Insert Table 7 near here]

The average returns of pilot and control stocks in our sample are extremely similar to those of Diether et al. (2009). This indicates that our sample only little differs from theirs. On both exchanges, there are no significant differences between average returns of pilot and control stocks for the announcement windows as well as for the event window. This finding is robust with respect to the SE estimator used and accords with the analysis of Diether et al. (2009). Our approximate permutation test also confirms that the average returns of pilot stocks do not significantly differ from those of the control stocks.

We stress that the choice of clustered SE used here is subjective. Researchers may find arguments for other clustering dimensions, such as clustering by industry to consider spillover effects as discussed by Boehmer et al. (2015), clustering by week to consider volatility clusters in the date, or any combination of these dimension in multi-way clustered SE. It is even possible that some of these clustering options result in significant differences, which would imply a serious dilemma in the interpretation of the results. Our approximate permutation test overcomes such issues by allowing to conduct hypothesis tests that are free from any potential bias due to misspecification of the correlation structure in the date. The application to the regulation SHO pilot program documents that it performs at least as well as any parametric test.

6. Conclusion

Difference-in-differences models are among the standard tools to empirically evaluate the effectiveness of policies or regulations. In this context, correctly estimating SE is crucial for unbiased inferences, but non-trivial. To account for serial correlation in the cross-section and/or over time, the existing literature recommends either to aggregate data (Bertrand et al. 2004) or to cluster standard errors (Petersen 2009; Cameron et al. 2011; Thompson 2011). However, these procedures may be overcritical or exhibit a low power, which both

are undesirable features in policy evaluation: Statistically rejecting a truly effective policy can destroy or prevent gains in economic welfare.

We suggest a simple approximate permutation test that does not impose any restrictions on the residual correlation. The test is based on permuting the variable of interest—the intervention—and estimating a counterfactual dependent variable. We evaluate our procedure by imposing simulated placebo and effective laws (Bertrand et al. 2004) on monthly excess returns of all stocks between January 1970 and December 2014 from the CRSP Stocks Database, and find this test dominates clustered SE in terms of size and power. Clustered SE tend to be biased: White and cross-sectionally single-clustered SE are undercritical in size, while the remaining estimators are overcritical and require effect levels between 4.6% and 6% p.a. for a correct power. In contrast, the approximate permutation test accepts the null of ineffectiveness for placebo laws at the given significance level—i.e. is correct in size—and identifies simulated laws with effects of 3% p.a. We also replicate an analysis of the impact of SEC regulation SHO on stock returns (Diether et al. 2009) with the approximate permutation test and find consistent results when compared to double clustered SE, with the important add-on that this inference is robust to any correlation in the residuals.

In total, our procedure allows researchers to conduct hypothesis tests of estimated DID effects without imposing any parametric form on the data. Due to this feature, it is very unlikely to produce biased inference that may result from erroneous assumptions in clustered SE estimation. Additionally, this test shows higher discriminatory abilities between effective and ineffective laws than the parametric tests.

References

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249–275.

Bertrand, Marianne, Paras Mehta, and Sendhil Mullainathan. 2002. "Ferreting Out Tunneling: An Application to Indian Business Groups". *Quarterly Journal of Economics* 117 (1): 141–148.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang. 2015. "Potential Pilot Problems: Treatment Spillovers in Financial Regulatory Experiments". *Columbia Business School Research Paper*.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors". *Review of Economics and Statistics* 90 (3): 414–427.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference With Multiway Clustering". *Journal of Business & Economic Statistics* 29 (2): 238–249.

Carhart, Mark M. 1997. "On persistence in mutual fund performance". *Journal of Finance* 52 (1): 57–82.

Diether, Karl B., Kuan-Hui Lee, and Ingrid M. Werner. 2009. "It's SHO Time! Short-Sale Price Tests and Market Quality". *Journal of Finance* 64 (1): 37–73.

Dwass, Meyer. 1957. "Modified Randomization Tests for Nonparametric Hypotheses". *The Annals of Mathematical Statistics* 28 (1): 181–187.

Economist. 2016. "Economists are prone to fads, and the latest is machine learning". Visited on 11/26/2016. http://www.economist.com/news/finance-and-economics/21710800-big-data-have-led-latest-craze-economic-research-economists-are-prone.

Ely, Richard T. 1886. "Ethics and Economics". *Science* ns-7 (175S): 529–533.

Fama, Eugene F., and Kenneth R. French. 1993. "Common risk factors in the returns on stocks and bonds". *Journal of Financial Economics* 33 (1): 3–56.

Fama, Eugene F., and Kenneth R. French. 2010. "Luck versus skill in the cross-section of mutual fund returns". *Journal of Finance* 65 (5): 1915–1947.

Fang, Vivian W., Allen H. Huang, and Jonathan M. Karpoff. 2016. "Short Selling and Earnings Management: A Controlled Experiment". *Journal of Finance* 71 (3): 1251–1294.

French, Kenneth F. 2016. "Data Library". Visited on 02/18/2016. http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Gow, Ian D., Gaizka Ormazabal, and Daniel J. Taylor. 2010. "Correcting for Cross-Sectional and Time-Series Dependence in Accounting Research". *Accounting Review* 85 (2): 483–512.

Huang, Rocco R. 2008. "Evaluating the real effect of bank branching deregulation: Comparing contiguous counties across US state borders". *Journal of Financial Economics* 87 (3): 678–705.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White. 2006. "Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis". *Journal of Finance* 61 (6): 2551–2595.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". *Econometrica* 55 (3): 703–708.

Petersen, Mitchell A. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches". *Review of Financial Studies* 22 (1): 435–480.

Siegel, Jordan, and Prithwiraj Choudhury. 2012. "A reexamination of tunneling and business groups: New data and new methods". *Review of Financial Studies* 25 (6): 1763–1798.

Thompson, Samuel B. 2011. "Simple formulas for standard errors that cluster by both firm and time". *Journal of Financial Economics* 99 (1): 1–10.

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

Table 1: Standard error estimation in DID studies

	Frequency	
Procedure	Absolute	Relative
OLS, White, Newey-West, undisclosed	32	22.1
Data aggregation, bootstrap	9	6.2
Single-clustered, cross-section	76	52.4
Single-clustered, time	13	9.0
Double-clustered, cross-section and time	15	10.3
Sum	145	100

Table 1 reports absolute and relative frequencies on standard error estimation 137 empirical studies that have been published in *The Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies* since 2010. Some studies used several estimation procedures, so multiple mentions occurred. The column „relative frequency" is reported in percent.

Table 2: Estimated coefficients of placebo laws in basic DID specifications

Specification	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Firm level								
Mean	0.0001	0.0004	−0.0002	−0.0002	0.0001	0.0004	−0.0003	−0.0001
Median	0.0026	0.0027	−0.0011	−0.0003	0.0001	0.0028	0.0000	−0.0005
Std. dev.	0.0229	0.0171	0.0149	0.0090	0.0225	0.0174	0.0112	0.0091
Skewness	−0.0117	−0.2482	0.0636	0.0779	0.0161	−0.2258	0.0574	0.1030
Kurtosis	1.6117	2.0326	2.6695	3.0569	1.6193	1.9721	2.7177	3.0152
Panel B: Industry level								
Mean	−0.0004	0.0001	−0.0005	0.0002	0.0000	0.0001	−0.0001	0.0002
Median	−0.0003	0.0001	0.0002	0.0003	−0.0010	0.0001	−0.0009	0.0003
Std. dev.	0.0123	0.0115	0.0127	0.0120	0.0121	0.0115	0.0125	0.0120
Skewness	0.0389	0.0055	−0.0078	0.0417	0.0810	0.0055	0.0902	0.0417
Kurtosis	2.8049	2.7880	2.8636	2.7766	2.7864	2.7880	2.9229	2.7766
Includes er_m?	No	No	No	No	Yes	Yes	Yes	Yes
Industry FE?	No	No	Yes	Yes	No	No	Yes	Yes
Year FE?	No	Yes	No	Yes	No	Yes	No	Yes

Table 2 presents the mean, the median, the standard deviation, the skewness, and the kurtosis of the estimated effects of 500 placebo laws for different specifications of the DID model as indicated by the last three rows. With the exceptions of skewness and kurtosis, all values are reported in decimals p.a.

Table 3: Size and power of parametric tests and our approximate permutation test

Controls	Law setting	SE estimator	Rejection rate		
			Placebo	Effective	Implicit
		Panel A: Firm level			
Industry FE	Random	White	0.17	0.75	2.95
year FE	random	1cl firm	0.33	0.86	2.54
		1cl industry	0.12	0.63	3.28
		1cl month	0.00	0.19	4.75
		1cl ind-month	0.00	0.01	5.97
		2cl firm & month	0.01	0.24	4.60
		2cl ind & month	0.10	0.53	3.59
		Non-parametric	0.05	0.61	3.00
		Panel B: Industry level			
Industry FE	Random	White	0.00	0.06	5.02
year FE	random	**1cl industry**	0.05	0.45	4.04
		1cl month	0.03	0.34	4.44
		2cl ind & month	0.07	0.46	4.09
		Non-parametric	0.05	0.42	3.73

Table 3 reports rejection rates at the 5%-level for simulated laws with a placebo-effect and an additive effect of +2% p.a. for various parametric tests and the non-parametric test we suggest. The implicit effect is the additive effect required for a correct power of a one-sided hypothesis test, i.e., the additive effect required to reject the H_0 of $\lambda \leq 0$ at the 5%-level in 95% of the cases. The DID-regressions are estimated on the level of firms in Panel A and on the level of industries in Panel B, using excess returns er_{it}^{+} as LHS-variable and industry dummies D_i^{ind} as well as year dummies D_t^{yr} as control variables. SE estimators in **bold font** indicate a correct size, i.e., we cannot reject the null hypothesis that the according rejection rate r for placebo laws is $r = 0.05$ at the 1%-level. SE estimators in *italic font* indicates an overcritical size, that is, we cannot reject the null hypothesis that the according rejection rate r for placebo laws is $r \leq 0.05$ at the 1%-level.

Table 4: Estimated coefficients of placebo laws of DID regressions including market models

Specification	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Firm level						
Mean	0.0006	0.0003	0.0002	−0.0002	0.0002	0.0002
Median	0.0007	0.0005	0.0005	−0.0001	0.0000	−0.0004
Std. dev.	0.0121	0.0100	0.0093	0.0073	0.0084	0.0087
Skewness	−0.1051	0.0139	−0.1307	0.0403	0.1403	0.2374
Kurtosis	2.5112	2.7356	2.9096	2.7674	3.0226	2.9409
Panel B: Industry level						
Mean	0.0003	−0.0002	−0.0006	−0.0006	0.0002	0.0004
Median	0.0002	−0.0002	−0.0001	0.0001	0.0000	0.0001
Std. dev.	0.0110	0.0104	0.0103	0.0097	0.0107	0.0107
Skewness	−0.0830	0.1531	−0.1445	−0.1254	−0.0606	0.0195
Kurtosis	3.1639	2.9728	3.0474	2.5447	2.6138	2.6864
Filtered LHS?	Yes	Yes	Yes	Yes	No	No
Factor controls?	No	No	No	No	Yes	Yes
Industry FE?	No	No	Yes	Yes	No	No
Year FE?	No	Yes	No	Yes	No	Yes

Table 4 presents the mean, the median, the standard deviation, the skewness, and the kurtosis of the estimated effects of 500 placebo laws for different specifications of the DID model as indicated by the last four rows. With the exceptions of skewness and kurtosis, all values are reported in decimals p.a.

Table 5: Size and power for filtered returns

Controls	Law setting	SE estimator	Rejection rate		Implicit
			Placebo	Effective	
		Panel A: Firm level			
Industry FE	Random	White	0.23	0.81	2.71
year FE	random	1cl firm	0.30	0.85	2.56
		1cl industry	0.11	0.63	3.31
		1cl month	0.02	0.30	4.03
		1cl ind-month	0.01	0.29	3.96
		2cl firm & month	0.02	0.32	3.98
		2cl ind & month	0.11	0.59	3.37
		Non-parametric	0.06	0.47	3.18
		Panel B: Industry level			
Industry FE	Random	*White*	0.01	0.36	4.05
year FE	random	**1cl industry**	0.06	0.50	3.51
		1cl month	0.02	0.38	4.04
		2cl ind & month	0.08	0.52	3.50
		Non-parametric	0.05	0.48	3.47

Table 5 reports rejection rates at the 5%-level for simulated laws with a placebo-effect and an additive effect of +2% p.a. for various parametric tests and the non-parametric test we suggest. The implicit effect is the additive effect required for a correct power of a one-sided hypothesis test, i.e., the additive effect required to reject the H_0 of $\lambda \leq 0$ at the 5%-level in 95% of the cases. The DID-regressions are estimated on the level of firms in Panel A and on the level of industries in Panel B, using filtered excess returns $\eta_{it}^{+} =$ as LHS-variable and industry dummies D_i^{ind} as well as year dummies D_t^{yr} as control variables. SE estimators in **bold font** indicate a correct size, i.e., we cannot reject the null hypothesis that the according rejection rate r for placebo laws is $r = 0.05$ at the 1%-level. SE estimators in *italic font* indicates an overcritical size, that is, we cannot reject the null hypothesis that the according rejection rate r for placebo laws is $r \leq 0.05$ at the 1%-level.

Table 6: Size and power using factor model estimates as control variables

Controls	Law setting	SE estimator	Rejection rate		Implicit
			Placebo	Effective	
		Panel A: Firm level			
Carhart	Random	White	0.41	0.96	1.94
alpha &	random	1cl firm	0.71	0.99	1.52
risk premium		1cl industry	0.53	0.97	1.87
year FE		**1cl month**	0.05	0.62	3.02
		1cl ind-month	0.05	0.62	2.99
		2cl firm & month	0.06	0.69	2.99
		2cl ind & month	0.29	0.65	2.89
		Non-parametric	0.05	0.61	2.79
		Panel B: Industry level			
Carhart	Random	White	0.10	0.58	3.51
alpha &	random	1cl industry	0.37	0.37	2.68
risk premium		1cl month	0.12	0.12	3.46
year FE		2cl ind & month	0.41	0.78	2.58
		Non-parametric	0.05	0.48	3.54

Table 6 reports rejection rates at the 5%-level for simulated laws with a placebo-effect and an additive effect of +2% p.a. for various parametric tests and the non-parametric test we suggest. The implicit effect is the additive effect required for a correct power of a one-sided hypothesis test, i.e., the additive effect required to reject the H_0 of $\lambda \leq 0$ at the 5%-level in 95% of the cases. The DID-regressions are estimated on the level of firms in Panel A and on the level of industries in Panel B, using excess returns as LHS-variable and alpha and risk premium according to a Carhart (1997) model based on estimated counterfactual returns and industry dummies D_i^{ind} as well as year dummies D_t^{yr} as control variables. SE estimators in **bold font** indicate a correct size, i.e., we cannot reject the null hypothesis that the according rejection rate r for placebo laws is $r = 0.05$ at the 1%-level. SE estimators in *italic font* indicates an overcritical size, that is, we cannot reject the null hypothesis that the according rejection rate r for placebo laws is $r \leq 0.05$ at the 1%-level.

Table 7: Average returns around the Reg SHO announcement and effective dates

	Announcement date			Event date		
	Pilot	Control	Diff.	Pilot	Control	Diff.
NYSE in Diether et al. (2009)						
Average daily return	0.332	0.281	0.051	0.273	0.286	−0.013
p-value 2cl firm & date			0.317			0.660
NYSE, our sample						
Average daily return	0.331	0.276	0.056	0.252	0.254	−0.002
p-value White			0.326			0.970
p-value 1cl firm			0.256			0.965
p-value 1cl date			0.357			0.934
p-value 2cl firm & date			0.296			0.744
p-value permutation test			0.281			0.972
NASDAQ in Diether et al. (2009)						
Average daily return	0.523	0.563	−0.040	0.357	0.383	−0.026
p-value 2cl firm & date			0.503			0.582
NASDAQ, our sample						
Average daily return	0.496	0.567	−0.071	0.356	0.387	−0.031
p-value White			0.493			0.734
p-value 1cl firm			0.425			0.681
p-value 1cl date			0.423			0.643
p-value 2cl firm & date			0.317			0.469
p-value permutation test			0.426			0.711

Table 7 reports the average returns for pilot und control stocks, their differences and the p-values of hypothesis tests of the differences during the announcment window and the event window of SEC regulation SHO. We run DID regressions of daily returns by exchange using the data from July 26, 2004 to May 4, 2005:

$$r_{it} = a + b_1 \cdot D_t^{ann} + b_2 \cdot D_t^{ann} \times D_i^{pilot} + b_3 \cdot D_t^{event} + b_4 \cdot D_t^{event} \times D_i^{pilot}$$

The dummy variable D_t^{ann} (D_t^{event}) equals one if the date is in between July 26, 2004 and July 30, 2004 (April 28, 2005 and May 4, 2005), inclusive, and zero otherwise and D_i^{pilot} is a dummy variable that equals one if a given stock is pilot stock, and zero otherwise. The columns for the announcent date (event date) report average returns for pilot stocks during the announcement (event) window, $a + b_1 + b_2$ ($a + b_3 + b_4$); the same measure for control stocks, $a + b_1$ ($a + b_3$); and the difference between the two, b_2 (b_4). All returns are report in percent per day.