Stanford Economic Review

Stanford Undergraduate Economics Journal

Volume 13

Editors and Staff

Note from the Editors

On behalf of the *Stanford Economic Review* Editorial Board, we are honored to present the thirteenth volume of Stanford University's undergraduate economics journal.

During the 2024–2025 academic year, we have focused on fostering a stronger connection between the undergraduate and graduate economics communities at Stanford. Most notably, in an effort to lower barriers to entry into research, we launched ”Why Econ?”—a speaker series featuring economics PhD students describing the various subfields of economics represented in academic research through informal conversations accessible to undergraduates. Internally, we have streamlined the editorial process, giving our team greater flexibility throughout the revision process.

The thirteenth volume features undergraduate research delving into the various intricacies of monetary policy, the optimal method to split samples into balanced subsamples in empirical studies, and even the efficacy of laws regulating sex offenders post-incarceration. Commentaries we've published this year to date have discussed pertinent domestic and global issues, ranging from the many faucets of energy policy to applying matching theory to fostering safe and inclusive ride-sharing to the ongoing interaction between developing technologies and the climate.

We would like to close by thanking the authors who have contributed an article to the journal or a commentary to our site. It would be impossible for the Stanford Economic Review to function without high-quality submissions and we especially appreciate all the time and effort put into communicating with our editors throughout the revision process. Finally, we would like to extend our gratitude to the Stanford Economics Association (SEA) and the Stanford Department of Economics for their continued support for our publication and mission.

Eric Gao and Tina Li
2024-25 Editors in Chief

Contents

Who's the Leader?
Evaluating International Monetary Policy Spillovers

Henry Weiland and Megan Yeo

Harvard University

***Abstract*—Do foreign central bank policy announcements impact domestic markets? In which direction do these spillovers occur? To answer these questions, we use changes in a country's two-year nominal yield around policy announcements as a proxy for monetary shocks. We aim to determine the causal effect of these announcements on other countries' financial instruments. We observe significant movements in 10-year yield spreads surrounding monetary policy announcements in developed economies, yet not stemming from announcements in emerging-market economies. Results appear most pronounced among geographic neighbors. We also find significant results for the Bank of Japan prior to the Great Financial Crisis, and the Federal Reserve after the Great Financial Crisis. We extend this framework to measure the response of break-even inflation rates, foreign exchange, and equities to policy announcements and macroeconomic releases. Our findings contribute to the literature on international bond market spillovers and central bank independence.**

I. Introduction

Interest rates often move in tandem globally as central banks adjust their monetary policies in response to economic conditions. During the global tightening of $2022-2023$, many central banks adjusted their policy in a significant break from the previous decade of low interest rates. This phenomenon led to concerns from the World Bank that monetary tightening could intensify due to the interconnected nature of the financial system (Guénette et al., 2022). We investigate the World Bank's main concern that the effects of a monetary policy announcement in one country could spill over to economies across the world.

However, the exact dynamics of *how* various countries influence each other within the financial market, specifically through their sovereign central bank policy announcements, remains an open question. Previous research such as Hanson and Stein (2015); Albagli et al. (2019); Lakdawala et al. (2021); Kaminska et al. (2021) has highlighted the impact of the Federal Reserve's policy announcements on long-term interest rates, an effect often attributed to portfolio rebalancing or exchange rate fluctuations. This focus on the United States' outward effect on other countries is likely due to the Federal Reserve's status as the central bank of the world's reserve currency; yet, it is unknown whether other central banks have a similar impact on other economies. (Ca' Zorzi et al., 2023) studies this question in the (limited) framework of the Federal Reserve and the European Central Bank.

To investigate the direction of international spillovers more broadly, we analyze monetary policy announcements from major central banks, including the Federal Reserve, the Bank of England, the European Central Bank, the Bank of Canada, the Bank of Japan, the Reserve Bank of Australia, and the Bank of New Zealand. We also study 9 emerging market central banks, namely the Bank of Mexico, Central Bank of Chile, Central Bank of Colombia, Bank of Israel, South African Reserve Bank, Reserve Bank of India, Bank of Thailand, Bank of Korea, and Central Bank of the Republic of China. Our goal is to understand how foreign central bank announcements influence domestic market changes, including spillovers on fixed income and equity measures. We utilize an event study with two-day periods surrounding these announcements with the change in the two-year nominal bond yield as a proxy for monetary policy shocks (Gürkaynak et al. (2005); Bauer and Swanson (2022, 2023)). We then regress the change in various foreign yields, such as the ten-year nominal yields, break-even inflation rates, foreign exchange, and stock market measures in response to these shocks. This isolated event ensures that changes in a country's yields are primarily attributed to foreign central bank announcements, serving as an effective instrument for analyzing monetary policy (Bauer and Swanson, 2022).

We find substantial evidence of monetary policy spillovers that affect nominal yields across countries, primarily among developed countries. The Fed exhibits outsized spillovers in both developed and emerging market economies, while the Bank of England, the European Central Bank, the Bank of Japan, and the Reserve Bank of Australia exhibit smaller ones. These spillovers are mainly isolated within the developed country sample, with some spillovers extending from developed countries to emerging ones. In the opposite direction, few spillovers are identified. In particular, the Bank of Japan exerts the most influence before the Great Financial Crisis while the Federal Reserve exerts the most influence after the Great Financial Crisis. Initially, these findings collectively point to a network of interdependent relationships among central banks; spillovers hold even after controlling for movements in U.S. yields.

We then extend the analysis to other financial measures, including break-evens, foreign exchange rates, and equities. To identify the spillover channel, we proceed to test whether information releases, such as CPI or GDP announcements, also move markets. These results more closely mirror the

We thank Paul Tucker, Jeremy Stein, Lawrence Summers, Namrata Narain, Alex Wu, and seminar participants in Harvard Economics 1499 and 980z for helpful guidance, feedback, and encouragement. We are responsible for all errors.

nominal yield results, supporting the story that central bank announcements creates an information channel that moves investor expectations about the future path of policy worldwide.

Our paper adds to the existing body of research on international bond market spillovers and the broader context of central bank independence. The paper is structured as follows. Section II provides background information on the international bond market and past spillover effects. Section III details the data set and the methodology. Section IV discusses the baseline regression results. Section V explores various channels. Section VI contains a robustness check. Section VII concludes the paper.

II. Background

International monetary policy spillovers (the influence decisions one country's central bank has on the rest of the world) into the bond market are well-studied. In a simplified model without the risk of spillovers from the bond market, prices would primarily reflect expectations of short-term interest rates according to the structure of the term. However, bond yields also include a term premium, which is the additional yield that investors require to hold longer-term bonds. This component of the yield is susceptible to exogenous shifts beyond the expected policy path (Ghosh and Qureshi, 2017).

These ripples into the international bond market have a well-documented history. A notable example occurred in 1994 when the Bundesbank initiated an easing cycle while the Federal Reserve announced a tightening phase. United States bonds subsequently increased, reflecting the anticipation of higher future short-term rates. However, European yields also rose in tandem with the U.S. market, even with their policy expected to take another direction (Ghosh and Qureshi, 2017). The influence of monetary policy decisions is not unidirectional. For example, during the mid-2010s, negative term premiums in the Eurozone appeared to contribute to subsequent decreases in U.S. term premium measures (Mojon and Pegoraro, 2014).

The twenty-first century heightened this sense of dependency as the Great Financial Crisis marked the beginning of widespread use of large-scale asset purchases (LSAPs), commonly referred to as quantitative easing. Central banks now aimed to influence government bond yields through signaling and buying securities. The increased purchase of government bonds increased the demand for longer-term securities, thereby reducing their term premium. This search for yield, particularly around the QE1 and QE2 announcements, revealed that investors are active agents in reacting to monetary policy announcements or the now-common-place press conference (Alpanda and Kabaca (2020); Bauer and Swanson (2023)).

The aftermath of the COVID-19 pandemic presents an opportunity to revisit the study of international bond market spillovers in a post-Great Financial Crisis context. The pandemic caused central banks to launch another round of aggressive asset purchase. Whereas the quantitative easing announcements were a U.S. measure to tackle a U.S. problem, the same purchases during the COVID era acted as an economic stimulus to other countries (International Monetary Fund. Research Dept., 2021). This phenomenon extends beyond the bond market: For instance, news about employment and economic activity in the U.S. was found to reduce credit spreads on U.S.-dollar-denominated bonds issued by foreign countries (Engler et al., 2023).

A. Literature Review

The literature on the identification of monetary policy shocks in a New-Keynesian framework traces back to Gürkaynak et al. (2005), which found that long-term forward rates exhibit significant movements in response to news regarding the anticipated trajectory of the federal funds rate around Federal Open Market Committee (FOMC) meetings. Thus, it is feasible to use variables that encapsulate information about the expected path of medium-term interest rates as a proxy to assess the effects of monetary policy.

The methodology for investigating the impact of monetary policy relies on the assumption that central bank announcements are exogenous events. This assumption validates the use of monetary policy announcements as instruments for estimating structural VARs or conducting event studies, as is approached in this paper. The exogeneity of these announcements has been challenged due to the predictability of certain monetary policy announcements. However, while structural VARs may inconsistently estimate the effects of monetary policy, standard event studies focused around announcements have yielded more consistent estimations of the monetary policy shock (Bauer and Swanson, 2023).

The technique, taken from the shelf in the early twenty-first century, was most notably used by Hanson and Stein (2015) to analyze the effects of changes in two-year yields around FOMC announcement days on instantaneous nominal, real and inflation forward rates. Their findings reveal that these two-year nominal rates significantly influence distant forward rates, largely through fluctuations in term premia by yield-oriented investors. Although the paper encompasses both the United States and the U.K., it does not explore how U.S. policies impact foreign countries or the converse question.

The causal framework was modified for the international context to measure U.S. spillovers by categorizing countries into developed markets (DEV) and emerging market economies (EME), utilized by Hofmann and Takáts (2015) and Albagli et al. (2019). In particular, yields in both DEV and EME countries tend to rise and steepen in reaction to U.S. monetary tightening, which marks a departure from the pre-GFC period (Gilchrist et al., 2015).

Although numerous studies have explored how monetary policy shocks impact different segments of the yield curve, there is relatively little research on the effects of international monetary policies on the U.S.. International spillovers are believed to influence DEV countries primarily through the risk neutral component (RN) and EME countries through term premia fluctuations (TP), but this may depend on whether the shock was due to unexpected monetary policy actions or revisions in interest rate expectations (Kaminska et al., 2021). However, in the opposite direction, spillovers to DEV countries occur in the term premia and to EME countries in the RN component, has also been found (Lakdawala et al., 2021). Regarding the extent of these spillovers, the few billateral studies indicate a significant impact of U.S. policies on Europe, but this influence is not reciprocated (Ca' Zorzi et al., 2023).

There remains much to be explored with the event study framework, from sentiment analysis of monetary policy statements to understanding new areas of the globe U.S. monetary policy reaches Hubert and Labondance (2021); Gai and Tong (2022). We hope to use the framework to construct an extension of the spillover literature to non-U.S. central banks. This study aims to extend the understanding of spillovers in the international bond market to the present day and provide an in-depth look at pairwise country interactions.

III. Data and Methodology

A. Study Sample

Our study examines the following central banks listed in Table 1 below. This list nearly parallels previous studies Gilchrist et al. (2015); Albagli et al. (2019), minus Indonesia, which we did not have reliable data. The categorization sorts the economies based on criteria provided by the IMF, UN, MSCI World index, and the DJIA. For consistency, we use the same categorization. Notably, European countries are grouped under the European Central Bank.

We manually collected data on central bank monetary policy announcements, projection materials, and minutes from 2000 to 2023 from the respective central bank websites. For some countries, the data coverage starts later due to archival limitations; for example, reliable records for Australia and Israel begin in 2006. Furthermore, we expanded the panel to cover the Federal Reserve from 1975 to 2023 and included historical data from the Deutsche Bundesbank of Germany (BUBA) for a future historical analysis section. We have made this dataset of central bank meeting and release dates available for future use.[1]

Developed Countries (DEV)	**Emerging Market Economies (EME)**
United States (Fed)	Mexico (Banxico)
United Kingdom (BOE)	Chile (CBC)
Europe (ECB)	Colombia (BanRep)
Canada (BOC)	Israel (BOI)
Japan (BOJ)	South Africa (SARB)
Australia (RBA)	India (RBI)
New Zealand (RBNZ)	Thailand (BOT)
	South Korea (BOK)
	Taiwan (CBRC)

TABLE I: Developed and Emerging Market Economies Classification

Since 2016, most central banks have adopted the standard of convening eight times a year, except Australia, New Zealand, South Africa, and Taiwan. Before 2016, it was more common for central banks to hold monthly meetings, as evidenced by the United Kingdom, Japan, Mexico, Chile, Colombia, Israel, India, and South Korea. A norm among countries is to also release a forecast, referred to as projection materials, which includes commentary about the economy and the future path of monetary policy. All central banks, minus Taiwan, release some version of a report. Information can also be signaled through the release of committee minutes, which contain notes on the discussion at each monetary policy meeting. Canada, New Zealand and South Africa are the only countries that do not release minutes. A table of each bank's policy can be found in **??**.

It is important to note that the frequency of the projection materials and the release dates of the minutes vary significantly between central banks and even within individual banks over time. For example, the Bank of England previously held a monetary policy meeting, released quarterly projection materials one week later, and issued the minutes two weeks after the initial meeting date. In 2015, this schedule was consolidated, so all three were released on the same day. Similar variations in release timing are also observed with the Fed minutes, Chile's projection materials, and Israel's minutes.

B. Measuring Monetary Policy Shocks

Analagous to previous event-study methodologies, we create two-day event study windows around central bank policy announcements to measure monetary policy shocks. We measure the change in the nominal yield of two years, denominated in the home currency, from $t-1$ to $t+1$ as our primary predictor variable. This choice reflects a standard in the literature that focuses on medium-term interest rate expectations rather than short rates. [2] The rationale, according to Hanson and Stein (2015), is as follows: actual changes in the policy rate are often infrequent and anticipated by the market. Hence, movement in short-term yields in a narrow window around monetary policy dates are a better reflection of changes to the expected path of the policy rate over subsequent quarters. In line with previous studies Albagli et al. (2019); Bauer and Swanson (2022), we select a two-day window to accommodate timezone differences that can create a disjunct between the timing of monetary policy shocks and when they are realized in different markets. An ideal experiment would involve minute-by-minute data showing changes in the yield, which would further isolate the monetary policy shock. Although we do not incorporate these data, we refer to previous work (Hanson and Stein, 2015) that shows that the two-day windows mirror the results from shorter ones.

We incorporate at least one full market response day to monetary policy announcements to address the impact of timezone differences. For example, consider the FOMC meeting on March 16, 2022. The Fed typically announces its interest rate decision at 2 PM, followed by a press conference. Considering

[1] The dates data is available at this repository: https://github.com/henryrweiland/Central-Bank-Meeting-Dates

[2] The following countries do not service a two-year bond: ECB, Israel, Mexico, and South Africa. In these cases, we used the change in the one-year yield in our baseline results.

the time-zone differences, markets in Japan, Australia, and New Zealand would react to this U.S. news on the morning of March 17, 2022. Consequently, for these markets, we measure the response to the Fed's announcement from the close of the market day on March 16 to the close on March 17. This methodology raises the concern that the effect on markets in far-away time zones might appear quantitatively smaller because of a shorter shock duration.

Importantly, we consider both directions of monetary policy announcements for pairing of countries (i, j). That is, we will measure the spillovers of country i on j as well as j on i. Our regression methodology is succinctly visualized as a matrix in Figure 1. [3]

Central Bank	**U.S.**	**Europe**	**Canada**	**Japan**
U.S.	*U.S. on U.S.*	U.S. on Europe	U.S. on Canada	U.S. on Japan
Europe	Europe on U.S.	*Europe on Europe*	Europe on Canada	Europe on Japan
Canada	Canada on U.S.	Canada on Europe	*Canada on Canada*	Canada on Japan
Japan	Japan on U.S.	Japan on Europe	Japan on Canada	*Japan on Japan*

Fig. I: Regression Methodology Matrix

C. Spillover Identification

To assess the various market responses to monetary policy announcements, we use a variety of financial measures, including long-term nominal bond yields, break-even inflation rates, foreign exchange and the domestic stock market. Our approach aims to capture the diverse movements of long-term instruments (Gürkaynak et al., 2020). All measures are based on the national currency of each country.

In our baseline regression, we use the change in long-term nominal yields,[4] which includes the ten-year and, where available, thirty-year yields.[5] It is crucial to note that decomposing the specific drivers of yield movements can be challenging, as changes in yields consist of shocks to expected inflation, real rates, and term premia (Cieslak and Schrimpf, 2019). The various financial instruments attempt to orthogonalize the shock to each of these components.[6]

We specify a year and month fixed effects OLS panel to identify the causal effect of country's j's change in the two-year yield around a policy announcement on country i's financial instrument:

$$\Delta y_{i,t} = \alpha_{month} + \alpha_{year} + \beta \text{MP}^{2y}_{j,t} + \gamma \text{MP}^{2y}_{i,t} + \epsilon_{j,t} \qquad (1)$$

where $MP_{i,t}$ denotes the change in the two-year yield of country i, $MP_{j,t}$ denotes the change in the two-year yield of country j, δy_i denotes the change in yield of the financial instrument for country i, α_{month} and α_{year} are monthly and yearly fixed effects.

To illustrate, consider measuring the spillover effects of Bank of Thailand monetary policy announcements on U.S. financial markets. The outcome variable, $\Delta y_{i,t}$, represents the change in a U.S. financial instrument (e.g., 10-year yield, term premium) observed within a two-day window around the announcement of the Thai monetary policy. To account for concurrent domestic factors, we control for $\text{MP}^{2y}_{i,t}$, the change in the U.S. two-year yield around the Thai meeeting. The primary variable of interest, $\text{MP}^{2y}_{j,t}$, captures the change in the Thai two-year yield during the announcement window, representing the shock of the monetary policy. Robust standard errors are clustered at the country-pair level to account for correlations across multiple announcements within the same pair. This specification builds on prior research Hanson and Stein (2015); Albagli et al. (2019) by generalizing the analysis to include spillovers from non-U.S. central banks.

We also specify a similar regression in Equation 2 to include a post-GFC interaction term to see if spillovers have changed over time:

$$\begin{aligned} \Delta y_{i,t} = \alpha_{month} + \alpha_{year} + \beta_1 \text{MP}^{2y}_{j,t} + \beta_2 \text{Post-GFC}_t \times \text{MP}^{2y}_{j,t} \\ + \gamma_1 \text{MP}^{2y}_{j,t} + \gamma_2 \text{Post-GFC}_t \times \text{MP}^{2y}_{j,t} + \epsilon_{j,t} \end{aligned} \qquad (2)$$

To address identification concerns, we implement several modifications to our analysis. First, we test only four specific matrices: developed countries, the Americas, Europe, and the Pacific region. The rationale behind focusing on these four is that they are more likely to induce spillovers, either among large central banks or neighboring countries. For the regional matrices, isolating countries within the same region allows us to measure the full change across two-day announcement windows without the complications introduced by differing time zones. This approach could increase statistical power, particularly for smaller central banks.[7]

Second, we re-specify Equations (1) and (2) to allow i to represent a full panel of countries rather than just a single (i, j) pairing. We let i be all countries as well as those in the developed and emerging markets, as originally defined in Table 1. For example, instead of measuring the effect of Bank of Thailand monetary policy announcements solely on the U.S. market, we extend i to include the financial instruments of all countries, developed economies, or emerging markets. This modification could mitigate the influence of potentially spurious movements within pairs.

Third, we alter Equations (1) and (2) to include the change in the U.S. yield in every regression. Given that the Federal Reserve is the largest central bank and controls the monetary policy behind the world's reserve currency, it is plausible that

[3] Figure 1 can be read as a matrix. Let i be on j. For example, in $(2, 1)$, means the predictor (independent) variable is the change in the U.S. nominal 2-year yield around a Fed announcement, and the outcome (dependent) variable is the change in the ECB nominal 10-year yield in the same window. Reading down a column can be interpreted as the spillover of country j onto each country i.

[4] Nominal yield data was obtained from Global Financial Data through a subscription provided by Harvard Business School Baker Library.

[5] The following countries issue a 10-year bond but not a 30-year: Chile, Colombia, New Zealand, and Thailand.

[6] Break-even inflation rates, foreign exchange, and the domestic stock market measures were taken from the Bloomberg Terminal accessed at Harvard Baker Library.

[7] We ran the full matrix, including every (i, j) pair, but many of the results were insignificant.

global monetary policy movements are influenced directly or indirectly by the U.S. at all times. Although this assumption is strong, incorporating the U.S. yield change provides a check against this possibility.

We implement the first two modifications in our results section and address the third in more detail in our robustness checks.

IV. Baseline Results

A. Nominal Yields

We first regress changes in the nominal yield windows of the ten and thirty year yields as the dependent variable using the regression specification in Section III.3 among the developed country sample. Changes in the two-year yields were used as predictors for all countries, except for ECB meeting dates, where the German two-year bond was used due to limited ECB data availability. Since the nominal yield incorporates information on the risk neutral, term premium, and expected inflation, the measure is useful for identifying the existence of a spillover, if any. As outlined in Section III.3, we begin by building regional matrices, measuring changes after the Great Financial Crisis, and then the entire country panel.

1) Individual-Level Country Pairs: The first regression matrix in Figure 2 illustrates the spillovers among the largest central banks identified in Table 1. Reading vertically down a column represents the impact of the column country's monetary policy announcement on the yield of the row country during the announcement window. Robust standard errors are reported in parentheses.[8]

Central Bank	Fed	BOE	ECB	BOC	BOJ	RBA	RBNZ
Fed	10y: 0.83*** (0.08) 30y: 0.47*** (0.08)	10y: −0.01 (0.04) 30y: −0.01 (0.05)	10y: 0.10** (0.04) 30y: 0.11* (0.05)	10y: 0.00 (0.06) 30y: 0.02 (0.06)	10y: 0.08 (0.06) 30y: 0.07 (0.06)	10y: 0.12*** (0.04) 30y: 0.10** (0.05)	10y: −0.02 (0.04) 30y: −0.01 (0.05)
BOE	10y: 0.20*** (0.04) 30y: 0.16*** (0.04)	10y: 0.72*** (0.06) 30y: 0.54*** (0.08)	10y: 0.12* (0.05) 30y: 0.11 (0.07)	10y: 0.00 (0.03) 30y: 0.10* (0.05)	10y: −0.01 (0.05) 30y: −0.02 (0.04)	10y: 0.05 (0.06) 30y: 0.02 (0.07)	10y: 0.05 (0.04) 30y: 0.01 (0.05)
ECB	10y: 0.21*** (0.07) 30y: 0.13** (0.06)	10y: 0.30*** (0.07) 30y: 0.30* (0.09)	10y: 0.31*** (0.09) 30y: 0.20*** (0.07)	10y: 0.15* (0.08) 30y: 0.13* (0.07)	10y: 0.05 (0.05) 30y: 0.11* (0.05)	10y: 0.10 (0.06) 30y: 0.14** (0.08)	10y: 0.02 (0.04) 30y: 0.03 (0.05)
BOC	10y: 0.21** (0.10) 30y: 0.08 (0.08)	10y: 0.02 (0.03) 30y: 0.01 (0.03)	10y: 0.07** (0.04) 30y: 0.07*** (0.03)	10y: 0.61*** (0.05) 30y: 0.34*** (0.06)	10y: 0.07** (0.03) 30y: 0.04 (0.04)	10y: 0.06** (0.03) 30y: 0.04 (0.03)	10y: −0.03 (0.02) 30y: −0.04 (0.03)
BOJ	10y: 0.03 (0.03) 30y: 0.01 (0.03)	10y: 0.05** (0.02) 30y : 0.00 (0.02)	10y: 0.00 (0.00) 30y: 0.00 (0.00)	10y: 0.00 (0.00) 30y: −0.01 (0.05)	10y: 0.06 (0.07) 30y: 0.05 (0.07)	10y: 0.06* (0.03) 30y: 0.03 (0.04)	10y: 0.06* (0.03) 30y: 0.05 (0.03)
RBA	10y: 0.07 (0.05) 30y: 0.06 (0.11)	10y: 0.145** (0.06) 30y: 0.31* (0.17)	10y: 0.05* (0.03) 30y: 0.12 (0.13)	10y: −0.02 (0.05) 30y: 0.02 (0.10)	10y: 0.11** (0.05) 30y: 0.44 0.50	10y: 0.69*** (0.05) 30y: 0.46*** (0.15)	10y: 0.01 (0.04) 30y: 0.18 (0.13)
RBNZ	10y: 0.15** (0.07)	10y: 0.04 (0.03)	10y: 0.00 (0.00)	10y: 0.07** (0.03)	10y: 0.18*** (0.05)	10y: 0.04 (0.07)	10y: 0.50*** (0.05)

Fig. II: Developed Countries Matrix

We observe vertically in columns 1, 2, and 3 that the U.S., the U.S., and Europe generate the largest spillovers among developed countries. In particular, these spillovers are not always bidirectional. For example, a 100bp rise in U.S. Treasury yields following a Fed announcement leads to a 20bp increase in English Gilts, but the U.K.'s announcements have no measurable impact on Treasuries. A similar pattern is observed with Australia's announcements, where Treasuries respond, but Australian bonds do not react to U.S. announcements on the same day. There are also statistically significant relationships between distant central banks. For example, New Zealand bonds move in response to the U.S. and Canada's announcements, despite being separated by a calendar day.

There also appears to be a "neighbor" effect among monetary policy. The U.K. and Europe exhibit strong mutual spillovers, while Japan, Australia, and New Zealand, located in similar time zones, show interconnected monetary policy reactions in the southeast part of the matrix. However, these neighbor relationships are often imbalanced. For example, the U.S. exerts influence on Canada, but Canada announcements have no significant impact on the U.S.. Similarly, Japan induces spillovers of 11 and 18bp on Australia and New Zealand, respectively, but only experiences a movement of 6bp on their announcements.[9]

We can extend this matrix framework to a region-by-region analysis for the Americas, Europe, and Asia. Since all of the central banks are located roughly within the same time zone, each of the countries receives a full two-day "treatment" of the monetary policy announcement, unlike splicing across time zones in the developed countries matrix. The table of results can be found below in Figures 3, 4, and 5 below. Columns indicate the central bank which monetary policy announcement dates are being tested (i.e. country j), and rows indicate the central bank which long yields are included in the regression (i.e. country i).

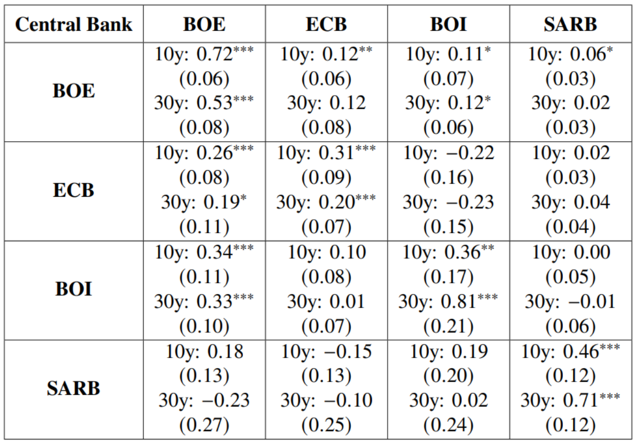
Central Bank	Fed	BOC	Banxico	CBC	BanRep
Fed	10y: 0.82*** (0.08) 30y: 0.47*** (0.08)	10y: 0.00 (0.06) 30y: 0.02 (0.06)	10y: 0.02 (0.04) 30y: 0.03 (0.05)	10y: 0.03 (0.05) 30y: 0.05 (0.05)	10y: 0.03 (0.04) 30y: 0.05 (0.04)
BOC	10y: 0.21** (0.10) 30y: 0.08 (0.08)	10y: 0.61*** (0.05) 30y: 0.34*** (0.06)	10y: 0.01 (0.03) 30y: −0.01 (0.04)	10y: −0.02 (0.04) 30y: −0.01 (0.04)	10y: 0.07*** (0.02) 30y: 0.07*** (0.02)
Banxico	10y: 0.36 (0.23) 30y: 0.47* (0.27)	10y: 0.13 (0.17) 30y: 0.42*** (0.11)	10y: 0.58*** (0.13) 30y: 0.42** (0.20)	10y: −0.09 (0.13) 30y: 0.06 (0.12)	10y: 0.27*** 30y: 0.08 (0.17)
CBC	10y: 0.21* (0.11)	10y: −0.25 (0.25)	10y: 0.16 (0.15)	10y: 0.37*** (0.10)	10y: −0.05 (0.08)
BanRep	10y: 0.06 (0.23)	10y: −0.09 (0.24)	10y: 0.17 (0.14)	10y: 0.01 (0.13)	10y: 0.41*** (0.09)

Fig. III: Americas Matrix

Across all matrices, changes in the nominal yield of two years exhibited the strongest effect on changes in the ten-year yield, with the influence generally diminishing for bonds of longer maturities. This trend is likely influenced by the term structure of interest rates, as changes in the two-year yields are more incorporated into the ten-year yields than into longer maturities. However, we observed a substantial impact of these changes in medium-term yields on even 30-year bonds. Consistent with expectations, monetary policy shocks are also linked to substantial shifts in 10-year yields within the same market, likely due to the structure of the yield

[8] A 10% p-value is indicated by $*$, a 5% p-value by $**$, and a 1% p-value by $***$.

[9] Japan is also the only diagonal entry where its own monetary policy announcement do not affect its longer term yields.



Central Bank	BOE	ECB	BOI	SARB
BOE	10y: 0.72*** (0.06) 30y: 0.53*** (0.08)	10y: 0.12** (0.06) 30y: 0.12 (0.08)	10y: 0.11* (0.07) 30y: 0.12* (0.06)	10y: 0.06* (0.03) 30y: 0.02 (0.03)
ECB	10y: 0.26*** (0.08) 30y: 0.19* (0.11)	10y: 0.31*** (0.09) 30y: 0.20*** (0.07)	10y: −0.22 (0.16) 30y: −0.23 (0.15)	10y: 0.02 (0.03) 30y: 0.04 (0.04)
BOI	10y: 0.34*** (0.11) 30y: 0.33*** (0.10)	10y: 0.10 (0.08) 30y: 0.01 (0.07)	10y: 0.36** (0.17) 30y: 0.81*** (0.21)	10y: 0.00 (0.05) 30y: −0.01 (0.06)
SARB	10y: 0.18 (0.13) 30y: −0.23 (0.27)	10y: −0.15 (0.13) 30y: −0.10 (0.25)	10y: 0.19 (0.20) 30y: 0.02 (0.24)	10y: 0.46*** (0.12) 30y: 0.71*** (0.12)

Fig. IV: Europe Matrix

curve. For instance, a 100 bp increase in the two-year Treasury yield causes with an 82 bp rise in the ten-year Treasury yield around FOMC announcements, representing the most significant monetary policy shock observed in the U.S. market.

Regional matrices reveal that spillovers are more pronounced among central banks with similarly sized balance sheets, typically occurring within the group rather than across groups. For example, in the Americas matrix, U.S. Treasury yields do not respond to other central bank announcements, but U.S. monetary policy announcements significantly impact the Canadian 10-year, Mexican 30-year and Chilean 10-year bonds; these are the largest spillovers within the group. Regional spillovers, such as Colombian central bank announcements that affect the Canadian market, are notably smaller, about one sixth the size of the spillovers caused by BOC announcements on the same market. Much the same holds in the European matrix, but with the U.K. and Europe instead inducing substantial spillovers onto each other.

The Asia matrix, being the largest, does not reveal a consistent pattern across central bank announcements. For example, BOJ announcements influence Australia and New Zealand but have limited spillover effects beyond these countries. In general, it seems that the larger central banks are more responsive to each other's movements, while the smaller central banks have a much smaller impact on the global market.

2) Changes After the Great Financial Crisis: Building on the second regression specification detailed in Section III.3, we incorporate an interaction term to evaluate the changes in the effect of monetary policy announcements on the international bond market following the Great Financial Crisis (GFC). The GFC marked the era of LSAPs, leading us to anticipate that these monetary policy announcements might influence signaling or portfolio re-balancing channels. The coefficient for the interaction term is listed as "Post-GFC" in the table. Robust standard errors are reported in parentheses.

Examining the panel before and after the 2008 financial crisis highlights significant changes in the financial landscape. Notably, English and Australian bond yields show increased sensitivity to Fed announcements post-crisis, suggesting that the financial market became more U.S.-centric in this period. However, this heightened responsiveness does not appear to extend broadly across other markets. Preliminarily, these findings support the hypothesis that the flight to the dollar following the Great Recession heightened investor sensitivity to U.S. financial announcements.

In contrast, investors were significantly more responsive to Japan's announcements before the Great Recession than afterward. Interaction terms for Japanese spillovers on the U.S., Canadian, and Australian markets are negative post-crisis, whereas the pre-crisis coefficients were large and significant. This shift, from responsiveness to Japan to a stronger focus on the U.S., suggests that investors reallocated their attention following the Great Financial Crisis, potentially leading to a permanent change in their reaction to macroeconomic news.

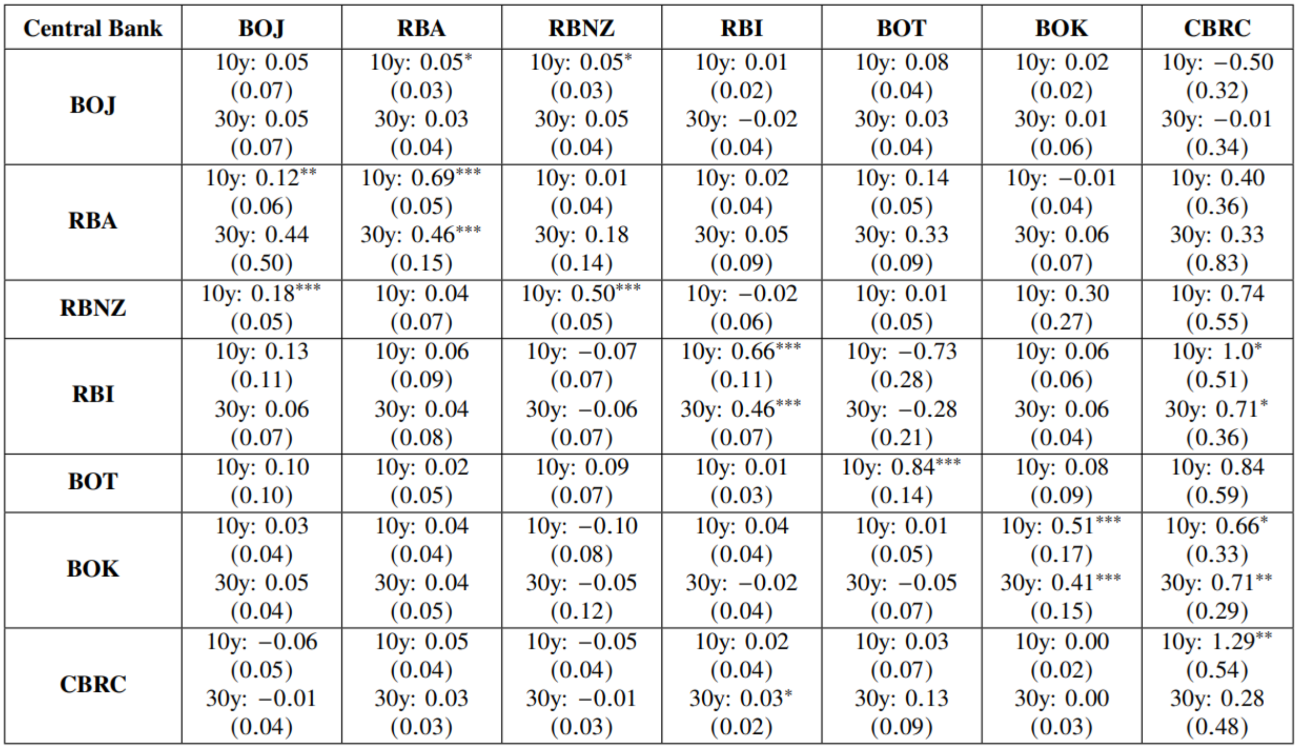
3) Country Panel: The final panel, which includes nominal yields of 10 and 30 years, extends the pairwise matrix framework to encompass all countries, all developed countries (DEV) or all emerging market countries (EME) in the fixed effects regression described in Section III.3. This approach offers the advantage of significantly increasing the sample size for each central bank announcement while potentially reducing spurious correlations between yields observed in the regional matrices. [10]

Country	10 Year			30 Year		
	All	Dev	EME	All	Dev	EME
Fed	0.20***	0.18***	0.21***	0.13***	0.12***	0.21***
BOE	0.19***	0.23***	0.10***	0.13***	0.15***	0.08***
ECB	0.19***	0.06***	0.07**	0.16***	0.07***	0.06***
BOC	0.12***	0.16***	0.06	0.13***	0.14***	0.15
BOJ	0.07***	0.09***	0.06***	0.04***	0.07**	0.07**
RBA	0.12***	0.15***	0.09**	0.10***	0.12***	0.09***
RBNZ	0.04**	0.05**	0.00	0.02**	0.02	−0.01
Banxico	0.07***	0.07***	0.09***	0.05***	0.06***	0.04
CBC	0.02	0.03***	−0.01	0.02*	0.04**	−0.03*
BanRep	0.07*	0.06**	0.08	0.08	0.05*	0.14
BOI	0.09	0.12	0.01	0.00	0.01	0.13
SARB	0.01	0.01	0.06**	0.01	0.01	0.02
RBI	0.00	0.01	−0.01	0.00	0.00	0.01
BOT	0.09***	0.08***	0.14	0.06*	0.09**	−0.06
BOK	0.05	0.05	0.04	0.01	0.01	0.05
CBRC	0.25	0.27	0.25	0.21	0.18	0.21

TABLE II: Country Nominal Yield Panel

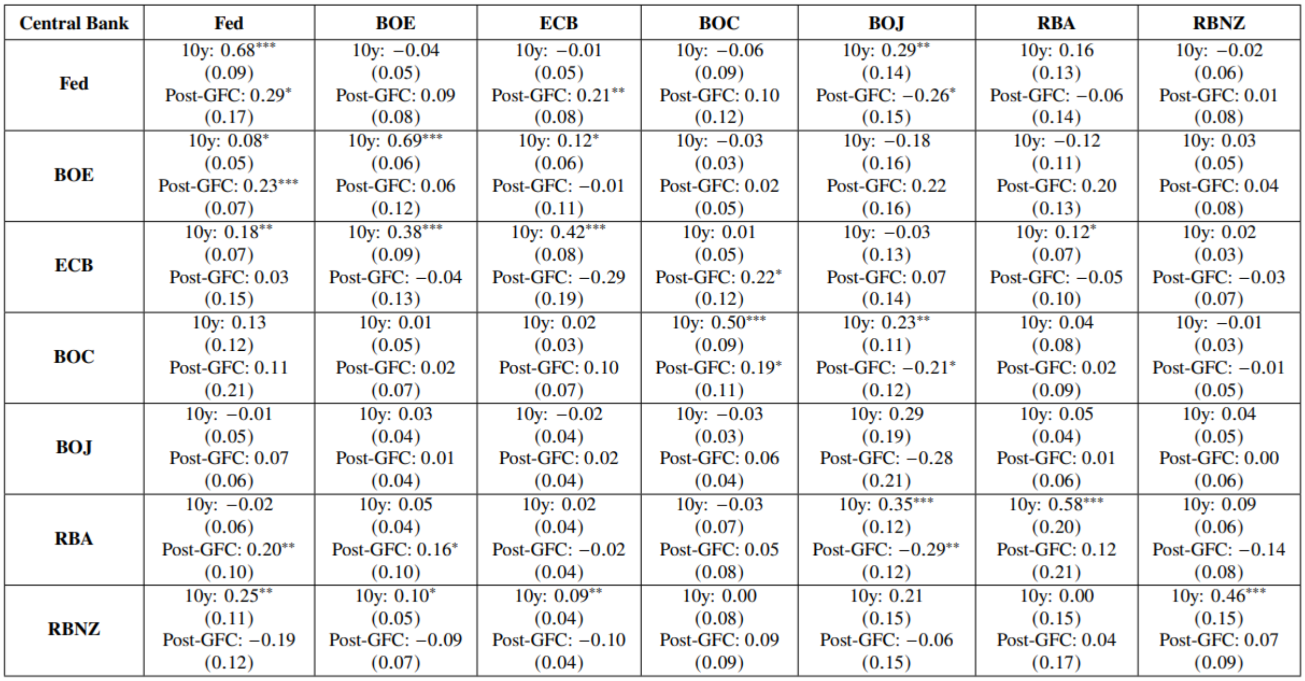
As observed in the regional matrices above, DEV central banks exert a significantly larger degree of spillovers onto the international market. The coefficients for the seven central banks of DEV are all significant at the level 1%, with spillovers that affect both the DEV and EME samples in 10-year and 30-year yields. For smaller DEV central banks, such as Canada or New Zealand, the coefficients are insignificant

[10] A 10% p-value is indicated by $*$, a 5% p-value by $**$, and a 1% p-value by $***$.



Central Bank	BOJ	RBA	RBNZ	RBI	BOT	BOK	CBRC
BOJ	10y: 0.05 (0.07) 30y: 0.05 (0.07)	10y: 0.05* (0.03) 30y: 0.03 (0.04)	10y: 0.05* (0.03) 30y: 0.05 (0.04)	10y: 0.01 (0.02) 30y: −0.02 (0.04)	10y: 0.08 (0.04) 30y: 0.03 (0.04)	10y: 0.02 (0.02) 30y: 0.01 (0.06)	10y: −0.50 (0.32) 30y: −0.01 (0.34)
RBA	10y: 0.12** (0.06) 30y: 0.44 (0.50)	10y: 0.69*** (0.05) 30y: 0.46*** (0.15)	10y: 0.01 (0.04) 30y: 0.18 (0.14)	10y: 0.02 (0.04) 30y: 0.05 (0.09)	10y: 0.14 (0.05) 30y: 0.33 (0.09)	10y: −0.01 (0.04) 30y: 0.06 (0.07)	10y: 0.40 (0.36) 30y: 0.33 (0.83)
RBNZ	10y: 0.18*** (0.05)	10y: 0.04 (0.07)	10y: 0.50*** (0.05)	10y: −0.02 (0.06)	10y: 0.01 (0.05)	10y: 0.30 (0.27)	10y: 0.74 (0.55)
RBI	10y: 0.13 (0.11) 30y: 0.06 (0.07)	10y: 0.06 (0.09) 30y: 0.04 (0.08)	10y: −0.07 (0.07) 30y: −0.06 (0.07)	10y: 0.66*** (0.11) 30y: 0.46*** (0.07)	10y: −0.73 (0.28) 30y: −0.28 (0.21)	10y: 0.06 (0.06) 30y: 0.06 (0.04)	10y: 1.0* (0.51) 30y: 0.71* (0.36)
BOT	10y: 0.10 (0.10)	10y: 0.02 (0.05)	10y: 0.09 (0.07)	10y: 0.01 (0.03)	10y: 0.84*** (0.14)	10y: 0.08 (0.09)	10y: 0.84 (0.59)
BOK	10y: 0.03 (0.04) 30y: 0.05 (0.04)	10y: 0.04 (0.04) 30y: 0.04 (0.05)	10y: −0.10 (0.08) 30y: −0.05 (0.12)	10y: 0.04 (0.04) 30y: −0.02 (0.04)	10y: 0.01 (0.05) 30y: −0.05 (0.07)	10y: 0.51*** (0.17) 30y: 0.41*** (0.15)	10y: 0.66* (0.33) 30y: 0.71** (0.29)
CBRC	10y: −0.06 (0.05) 30y: −0.01 (0.04)	10y: 0.05 (0.04) 30y: 0.03 (0.03)	10y: −0.05 (0.04) 30y: −0.01 (0.03)	10y: 0.02 (0.04) 30y: 0.03* (0.02)	10y: 0.03 (0.07) 30y: 0.13 (0.09)	10y: 0.00 (0.02) 30y: 0.00 (0.03)	10y: 1.29** (0.54) 30y: 0.28 (0.48)

Fig. V: Asia Matrix



Central Bank	Fed	BOE	ECB	BOC	BOJ	RBA	RBNZ
Fed	10y: 0.68*** (0.09) Post-GFC: 0.29* (0.17)	10y: −0.04 (0.05) Post-GFC: 0.09 (0.08)	10y: −0.01 (0.05) Post-GFC: 0.21** (0.08)	10y: −0.06 (0.09) Post-GFC: 0.10 (0.12)	10y: 0.29** (0.14) Post-GFC: −0.26* (0.15)	10y: 0.16 (0.13) Post-GFC: −0.06 (0.14)	10y: −0.02 (0.06) Post-GFC: 0.01 (0.08)
BOE	10y: 0.08* (0.05) Post-GFC: 0.23*** (0.07)	10y: 0.69*** (0.06) Post-GFC: 0.06 (0.12)	10y: 0.12* (0.06) Post-GFC: −0.01 (0.11)	10y: −0.03 (0.03) Post-GFC: 0.02 (0.05)	10y: −0.18 (0.16) Post-GFC: 0.22 (0.16)	10y: −0.12 (0.11) Post-GFC: 0.20 (0.13)	10y: 0.03 (0.05) Post-GFC: 0.04 (0.08)
ECB	10y: 0.18** (0.07) Post-GFC: 0.03 (0.15)	10y: 0.38*** (0.09) Post-GFC: −0.04 (0.13)	10y: 0.42*** (0.08) Post-GFC: −0.29 (0.19)	10y: 0.01 (0.05) Post-GFC: 0.22* (0.12)	10y: −0.03 (0.13) Post-GFC: 0.07 (0.14)	10y: 0.12* (0.07) Post-GFC: −0.05 (0.10)	10y: 0.02 (0.03) Post-GFC: −0.03 (0.07)
BOC	10y: 0.13 (0.12) Post-GFC: 0.11 (0.21)	10y: 0.01 (0.05) Post-GFC: 0.02 (0.07)	10y: 0.02 (0.03) Post-GFC: 0.10 (0.07)	10y: 0.50*** (0.09) Post-GFC: 0.19* (0.11)	10y: 0.23** (0.11) Post-GFC: −0.21* (0.12)	10y: 0.04 (0.08) Post-GFC: 0.02 (0.09)	10y: −0.01 (0.03) Post-GFC: −0.01 (0.05)
BOJ	10y: −0.01 (0.05) Post-GFC: 0.07 (0.06)	10y: 0.03 (0.04) Post-GFC: 0.01 (0.04)	10y: −0.02 (0.04) Post-GFC: 0.02 (0.04)	10y: −0.03 (0.03) Post-GFC: 0.06 (0.04)	10y: 0.29 (0.19) Post-GFC: −0.28 (0.21)	10y: 0.05 (0.04) Post-GFC: 0.01 (0.06)	10y: 0.04 (0.05) Post-GFC: 0.00 (0.06)
RBA	10y: −0.02 (0.06) Post-GFC: 0.20** (0.10)	10y: 0.05 (0.04) Post-GFC: 0.16* (0.10)	10y: 0.02 (0.04) Post-GFC: −0.02 (0.04)	10y: −0.03 (0.07) Post-GFC: 0.05 (0.08)	10y: 0.35*** (0.12) Post-GFC: −0.29** (0.12)	10y: 0.58*** (0.20) Post-GFC: 0.12 (0.21)	10y: 0.09 (0.06) Post-GFC: −0.14 (0.08)
RBNZ	10y: 0.25** (0.11) Post-GFC: −0.19 (0.12)	10y: 0.10* (0.05) Post-GFC: −0.09 (0.07)	10y: 0.09** (0.04) Post-GFC: −0.10 (0.04)	10y: 0.00 (0.08) Post-GFC: 0.09 (0.09)	10y: 0.21 (0.15) Post-GFC: −0.06 (0.15)	10y: 0.00 (0.15) Post-GFC: 0.04 (0.17)	10y: 0.46*** (0.15) Post-GFC: 0.07 (0.09)

Fig. VI: Developed Countries Matrix, Post-GFC Indicator

or comparable to those of Mexico or Thailand. It is initially surprising that these central bank announcements can influence markets, but their coefficients are relatively small as they are less than half the size of those associated with U.S. announcements. Further robustness checks, incorporating the Fed's 2-year yield in every regression, are detailed in Section

VI.

V. Monetary Policy Transmission Channels

In this section, we attempt to investigate the channels through which spillovers of monetary policy occur. To do so, we test a few instruments that provide insight into channels of monetary policy transmission: breakevens, foreign exchange rates, and equities. Finally, we run tests on the effect of key macroeconomic information releases in the major economies that have had the greatest spillover effects in other economies (the U.S., U.K. and Europe).

A. Future Expectations

We first assess whether monetary policy announcements influence inflation expectations over the next ten years. To do so, we again use the regression specification in Section III.3, adopting ten-year break-even inflation rates as the dependent variable. It is essential to note that these rates are not direct indicators of expected inflation alone, as they also include other market factors, such as inflation risk and liquidity premiums. To accurately capture the market's inflation expectation, we either sourced these break-even inflation rates directly or derived them by subtracting the yield of ten-year inflation-linked bonds from the ten-year nominal bond yield.

Central Bank	**Fed**	**BOE**	**BOC**	**RBA**	**RBNZ**
Fed	-0.048 (0.08)	0.0026 (0.056)	-0.043 (0.043)	0.138^{**} (0.048)	0.043 (0.044)
BOE	0.11^{*} (0.05)	0.62^{***} (0.053)	0.057 (0.030)	-0.0094 (0.060)	-0.011 (0.039)
BOC	0.66 (0.21)	0.073 (0.055)	0.047^{***} (0.099)	-0.19 (0.20)	-0.025 (0.029)
RBA	-0.33^{***} (0.08)	0.0088 (0.090)	-0.067 (0.060)	-0.14 (0.11)	-0.028 (0.050)
RBNZ	0.124 (0.07)	0.15^{**} (0.050)	0.070 (0.055)	-0.043 (0.12)	0.047^{***} (0.054)

Fig. VII: Developed Economies Break-evens Matrix

Break-even regressions involving the U.S., the U.K., Canada, Australia and New Zealand are included in Figure 7. The analysis excludes Europe and Japan due to data limitations, as break-even or inflation-linked bonds could not be found for these economies. Results generally do not suggest significant movements in breakeven inflation expectations due to changes in monetary policy - we note that for the U.S. and Canada, for instance, domestic monetary policy does not cause a significant change in their own breakeven. We note some outliers: the U.S. does influence the U.K. and Australia's breakevens, and, most importantly, Australian central bank announcements demonstrate some impact on the U.S..

Future studies could aim to calculate their break-even rates and incorporate them into a more comprehensive dataset. These current findings highlight the existence of monetary policy spillovers, extending even to inflation expectations. Generally, business cycles in developed economies are interlinked; therefore, increasing two-year yields around monetary policy announcements could reinforce the currently heightened inflation expectations. However, most of the data sample, especially before 2020, involves periods of low inflation and anchored expectations. Future research could understand why inflation expectations shown in breakevens in some countries respond to specific monetary policy announcements, such as those from the U.S. and Australia, but not to others.

B. Foreign Exchange Rates

We investigate changes in exchange rates in response to monetary policy. Albagli et al. (2019) claims that EMEs are more likely to intervene in their currency markets and therefore have a much smaller magnitude of response to central bank policy announcements than developed countries. They argue that intervention in currency increases capital inflows, affecting bond term premia. Here, we expand the FX sample to include the effects of monetary policy announcements by the U.K. and Europe, given that these are the central banks seen to have the most spillovers into other economies besides the Fed. Here, we use the percentage change in the exchange rate as a dependent variable. The results are reported in Table 3:

Central Bank	**Foreign Exchange Rates**		
	All	**Dev**	**EME**
BOE	-1.402^{***}	-1.384^{***}	-1.414^{***}
ECB	0.701^{***}	0.543	0.806^{***}

TABLE III: BOE and ECB Spillovers on Foreign Exchange Rates

A negative coefficient suggests an appreciation in the currency (EUR or GBP), while a positive coefficient suggests a depreciation against other currencies. Interestingly, first note that while the signs of the coefficients for the U.K. appear to cohere with economic theory, this is not the case for Europe. Indeed, (Gürkaynak et al., 2021) find that for the Fed and ECB, exchange rates depreciate rather than appreciate during unexpected policy tightening, consistent with our results.

The results for the BOE are consistent with Albagli et al. (2019)'s finding, as the GBP appreciates more against the currencies of EME economies than developed economies, supporting the hypothesis of some currency intervention for emerging economies. For the EUR, it depreciates more against emerging market economies than against developed economies. Although the direction of foreign exchange rate movement is uncertain, it is clear that the currencies are emerging market economies that fluctuate more in response to monetary policy shocks than developed economies, further supporting the existence of the exchange rate channel for emerging market economies.

C. Information Releases

In this section, we draw an extension to understand how much the information channel could play a role in determining spillovers from monetary policy announcements. According to Hanson and Stein (2015) and Albagli et al. (2019), if responses are equally or more strong on days with macro news than on monetary policy dates, it could suggest the role of the information channel, since the information on the day is more to do with the central bank's reaction than to macro news itself. The coefficients on the regressions of the ten-year on the two-year may be seen as elasticities of long rates given either types of release (monetary policy announcement or macro news).

We thus test the effect of macro news from the U.S., U.K. and Eurozone on other economies in our samples, making sure to exclude overlapping dates of macro news with counterparties.

We obtain data from Bloomberg's Economic News calendar for each economy. From 2004 to 2024, Bloomberg's sample of dates of inflation and GDP data releases yields 1,279 dates for the U.S., 652 for the U.K. and 952 for the Eurozone. Due to data availability, only a subsample of economies are considered. [11] For instance, inflation news in the U.S. includes CPI and PCE releases, and GDP includes GDP releases. We acknowledge that Bloomberg's coverage is likely very comprehensive, and not all releases cataloged could be equally important, nevertheless, they represent an information release that investors pay attention to.

The results are shown in Table 4. We find strong effects of macroeconomic releases from the U.S., U.K., and Europe on long rates in other developed economies. In particular, Fed information releases influence most economies. Long rates in emerging economies are less responsive to information releases from the U.S., U.K. and Europe, with the exception of Thailand.

Country	10-Year Yields		
	Fed	BOE	ECB
Fed	0.808***	0.062***	0.074
BOE	0.087***	0.774***	0.143***
ECB	0.151***	0.324**	0.265***
RBNZ	0.094	0.087***	0.073**
CBC	0.061	0.079	0.153
BanRep	0.099	0.166	0.414
BOI	0.224***	0.196	0.206
BOT	0.084***	0.121**	0.128***
CBRC	0.064***	0.009	0.031

TABLE IV: Co-Movements between Two-Year Yields and Ten-Year Yields on Macro News Dates

In general, comparing the coefficients for developed countries to their counterparts in the developed countries matrix on central bank monetary policy dates to deduce the elasticity of long rates on these two types of dates, we make a few notable observations. For the U.S., coefficients on the response on macro news release dates are lower than on FOMC announcement dates, confirming the hypothesis that the information channel is not as important since investors are more responsive to FOMC announcement dates. For the U.K., results are more mixed - for comovements of the two-year yield with its own 10-year yield, the coefficient is higher for information dates as compared to policy announcement dates, while the coefficient on comovements with long rates in other developed economies are lower. The information channel may thus be more important domestically than for foreign investors. Similarly, for Europe, the information channel appears less significant in the domestic context, as well as its spillovers to England. Overall, it appears that the information channel may be more important for spillovers from the U.K. and Europe than for the U.S..

D. Equities

Finally, we investigate movements in daily equity returns. This gives us insight into how monetary policy influences the returns of riskier and more volatile assets. In this regression, our independent variable remains the change in the two-year yield (or one or three-year yield for some economies), while our dependent variable is the percentage change in the price of a representative stock market index. The results are reported in Table 5. As much as possible, we obtain representative indices of large-cap stocks, such as the S&P 500 in the U.S. or the EuroStoxxLarge. Where data was unavailable, the best FTSE index option was used (e.g. FTSE South Africa).

The results for equities, shown in Table 5, appear mixed. In the developed economies section, the central banks of the U.K., Japan, and New Zealand demonstrate the most significant impact, as a rise in the two-year yield correlates with a decrease in equities in other economies. Surprisingly, the Fed's tightening leads to a small negative effect in emerging market economies, but is less significant on developed economies. For emerging market economies, surprisingly, we observe that the two-year yields of India, South Korea, and Taiwan demonstrate the most significant correlation with a rise in equity returns.

Based on the evidence from our event study, we are unable to deduce a clear relation between monetary policy in one economy and equity returns in other economies. Indeed, an important point to note is that the correlation between bond yields and equity returns in themselves may not be well-established: research suggests that correlations were previously negative but switched sign in 2021 (Lombardi and Sushko, 2023).

VI. Robustness Check

One concern with the event study methodology is the challenge of isolating monetary policy shocks across international borders. For example, while we examined the information release channel in Section V.3, it is plausible that movements in international yields on a given day could be driven *entirely* by the Federal Reserve, rather than by

[11] Data will be expanded to the full sample at the earliest opportunity.

Central Bank	Equities		
	All	Dev	EME
U.S. (S&P 500)	−0.312	0.279	−0.859**
BOE (FTSE 100)	0.989**	1.674***	0.373
ECB (EuroStoxx Large)	0.135***	0.181**	0.100
BOC (FTSE Large Cap Canada)	−0.109	0.025	−0.199
BOJ (Nikkei 225)	−8.917***	−8.972***	−8.927***
RBA (ASX 50)	−0.009**	−0.011**	−0.008
RBNZ (FTSE NZ Large Cap)	2.332***	3.028***	1.783***
Banxico (Mexico IPC Large Cap)	−0.472	0.126	−0.993***
CBC (IGPA)	−0.318	−0.286	−0.346
BanRep (COLCAP)	−1.729***	−0.873	−2.042**
RBI (SENSEX)	1.034***	0.885***	1.164***
BOI (TA 125)	0.725	1.499*	0.086
SARB (FTSE South Africa)	0.444	1.217***	−0.320
BOT (SET)	−0.112	−0.846**	0.516
BOK (KOSPI)	1.094***	1.396***	0.835***
CBRC (FTSE Taiwan)	17.245***	19.398***	15.423***

TABLE V: Central Bank Monetary Policy Effects on Equity Returns

other factors. This could explain why Thai monetary policy announcements appear to impact global yields: despite being a smaller central bank, such movements may instead reflect unaccounted for actions by the Fed. We therefore modify Equation 1 to always control for movements in the U.S. 2-Year Treasury yield during every foreign central bank monetary policy announcement.

$$\Delta y_{i,t} = \alpha_{month} + \alpha_{year} + \beta \text{MP}^{2y}_{j,t} + \gamma \text{MP}^{2y}_{i,t} + \delta \text{MP}^{2y}_{\text{Fed},t} + \epsilon_{j,t} \tag{3}$$

We repeat the panel presented in Section IV.1.3 but add the Fed control. The results are presented in Table 3. The coefficients for the U.K., Europe and Australia remain significant; however, their magnitudes decrease with the inclusion of the control variable. In contrast, the coefficients for Canada and New Zealand are now insignificant. Among emerging market countries, Mexico, Chile and Thailand exhibit small but significant relationships, although the results are inconsistent between the 10-year and 30-year yields. This suggests that even in a scenario where all movements in global yields are attributed solely to the Fed, there is still evidence that other central banks can influence global yields on their days of monetary policy.

VII. Conclusion

This paper examines how central bank monetary policy announcements in developed and emerging market economies influence financial instruments in international bond markets. Our method involves analyzing event study windows centered on these central bank monetary policy announcements. We proxy for the monetary policy shock using a regression with the change in a country's two-year nominal bond yield as the independent variable and various financial indicators such as the ten-year, break-even inflation rates, foreign exchange, and equities as outcomes. Our findings indicate significant spillover effects, particularly among developed economies in the international bond market. Furthermore, we observe a weak "neighbor effect," suggesting that central banks that are geographically close have stronger mutual spillover effects upon each other. These results hold even after controlling for U.S. movements in all regressions.

Country	10-Year Yields			30-Year Yields		
	All	Dev	EME	All	Dev	EME
Fed	0.20***	0.18***	0.21***	0.13***	0.12***	0.21***
BOE	0.07***	0.08**	0.05**	0.06**	0.05	0.06*
ECB	0.09***	0.04**	0.03	0.08**	0.05**	0.01
BOC	0.00	0.01	−0.02	0.05	0.05	0.08
BOJ	0.10***	0.10***	0.04**	0.06*	0.06*	0.07*
RBA	0.09***	0.10***	0.07*	0.06***	0.07***	0.06***
RBNZ	0.00	0.01	−0.04	0.00	0.00	−0.02
Banxico	0.04**	0.03***	0.06*	0.03*	0.04**	0.01
CBC	0.00	0.02***	−0.03	0.01	0.03*	−0.04**
BanRep	0.08	0.08	0.05	0.09	0.06	0.03
SARB	0.00	0.00	0.04**	0.01	0.00	0.03
RBI	0.01	0.02	−0.02	0.01	0.01	0.00
BOT	0.07	0.05**	0.13	0.05	0.08	−0.07
BOK	0.00	0.00	0.03	−0.01	−0.01	0.01
CBRC	0.23	0.26	0.22	0.14	0.11	0.13

TABLE VI: Country Nominal Yield Panel, Fed Control

We also attempt to understand the channels of monetary policy transmission by testing various financial instruments. We find that break-even inflation expectations may not be significantly affected by monetary policy spillovers, with few exceptions, such as between Australia and the U.S.. We find further evidence supporting the Albagli et al. (2019) hypothesis of an exchange rate channel for economies in the extended sample, as well as a significant response of investors to information releases. However, for the U.S., an information channel of monetary policy spillovers is less significant, while it may play a larger role for the U.K. and Europe. The transmission of monetary policy to riskier assets, such as equities, remains mixed.

References

Albagli, E., Ceballos, L., Claro, S., and Romero, D. (2019). Channels of us monetary policy spillovers to international bond markets.

Alpanda, S. and Kabaca, S. (2020). International spillovers of large-scale asset purchases.

Bauer, M. D. and Swanson, E. T. (2022). A reassessment of monetary policy surprises and high-frequency identification. Working Paper 29939, National Bureau of Economic Research.

Bauer, M. D. and Swanson, E. T. (2023). An alternative explanation for the "fed information effect". *American Economic Review*, 113(3):664–700.

Ca' Zorzi, M., Dedola, L., Georgiadis, G., Jarociński, M., Stracca, L., and Strasser, G. (2023). Making waves: Monetary policy and its asymmetric transmission in a globalized world.

Cieslak, A. and Schrimpf, A. (2019). Non-monetary news in central bank communication. *Journal of International Economics*, 118:293–315.

Engler, P., Piazza, R., and Sher, G. (2023). Spillovers to emerging markets from us economic news and monetary policy. Technical report, International Monetary Fund.

Gai, P. and Tong, E. (2022). Information spillovers of us monetary policy.

Ghosh, A. R. and Qureshi, M. S. (2017). From great depression to great recession: The elusive quest for international policy cooperation.

Gilchrist, S., López-Salido, D., and Zakrajšek, E. (2015). Monetary policy and real borrowing costs at the zero lower bound.

Gürkaynak, R. S., Kisacikoğlu, B., and Wright, J. H. (2020). Missing events in event studies: Identifying the effects of partially measured news surprises. *American Economic Review*, 110(12):3871–3912.

Guénette, J. D., Kose, M. A., and Sugawara, N. (2022). Is a global recession imminent? Equitable Growth, Finance, and Institutions Policy Note No. 4, World Bank, Washington, DC.

Gürkaynak, R. S., Kara, A. H., Kısacıkoğlu, B., and Lee, S. S. (2021). Monetary policy surprises and exchange rate behavior. *Journal of International Economics*, 130:103443.

Gürkaynak, R. S., Sack, B., and Swanson, E. (2005). The sensitivity of long-term interest rates to economic news.

Hanson, S. G. and Stein, J. C. (2015). Monetary policy and long-term real rates.

Hofmann, B. and Takáts, E. (2015). International monetary spillovers.

Hubert, P. and Labondance, F. (2021). The signaling effects of central bank tone.

International Monetary Fund. Research Dept. (2021). Chapter 4 shifting gears: Monetary policy spillovers during the recovery from covid-19. Retrieved Dec 16, 2024.

Kaminska, I., Mumtaz, H., and Šustek, R. (2021). Monetary policy surprises and their transmission through term premia and expected interest rates.

Lakdawala, A., Moreland, T., and Schaffer, M. (2021). The international spillover effects of us monetary policy uncertainty.

Lombardi and Sushko (2023). The correlation of equity and bond returns. *BIS Quarterly Review*.

Mojon, B. and Pegoraro, F. (2014). Decoupling euro area and US yield curves.

Impacts of Pandemic Instruction Mode on High School Students' Education Outcomes: Evidence from U.S. States

Erica Ewing Zhou
Northwestern University

Abstract—**Previous research has established that the transition from face-to-face instruction to distance education during the COVID-19 pandemic adversely impacts the academic performance of students in grades 3 through 8. To date, however, little is known if similar effects of instruction mode changes apply to students in higher grade levels. This study investigates this gap by using high school test proficiency and dropout rates data from six U.S. states. Contrary to the existing literature, this study reveals that high school students on average did not experience considerable learning losses during the pandemic, with no compelling evidence of widening achievement gaps among marginalized groups. Additionally, the effect of instruction mode varied considerably across diverse education and social contexts, with some states demonstrating notable benefits from virtual and hybrid learning. Changes in dropout rates exhibit greater variability and less economic and statistical significance, suggesting intricate dropout dynamics influenced by various pandemic-related disruptions beyond mere instruction mode changes.**

I. Introduction

On March 11, 2020, the World Health Organization declared COVID-19 a pandemic after over 118,000 reported cases and 4,291 deaths across 114 countries. To mitigate the spread of the coronavirus, the United States enacted various social restrictions, including workplace lock-downs, school closures, and stay-at-home orders. Uncertainties surrounding the pandemic's duration and severity posed challenges for education leaders in formulating policies that would adequately support students and staff. During the 2020-21 school year, schools nationwide adopted a spectrum of learning models, ranging from completely virtual schooling to full-time in-person instruction, as well as a hybrid approach combining both modalities. In-person schooling both within and across states was more common in political conservative areas and communities with higher COVID-19 infection rates (Jack et al., 2023).

According to survey-based studies, student experience with virtual learning differed. While some perceived flexible class schedules and recorded content as beneficial, others emphasized the importance of face-to-face instruction for academic and social well-being (Photopoulos et al., 2022). Quantitative research assessing student performance in online environments broadly suggested learning losses due to school closures. They also reported larger losses among students of color and socioeconomically disadvantaged students, which exacerbated preexisting achievement gaps. These conclusions were primarily drawn from direct comparisons of pre- and post-pandemic test scores. In this case, other pandemic-related factors, such as increased stress and anxiety, also contributed to the observed learning losses. Moreover, declines in test scores from the previous year might be following a trend that was present even before the pandemic. Therefore, without a proper identification strategy, it is challenging to conclude if it is the change of instruction mode that harmed student achievements.

The answer to this question holds profound implications for education policymakers tasked with designing effective strategies to support student learning even in a post-pandemic era. This study contributes to this discourse by leveraging previous research and utilizing the COVID-19 School Data Hub, a public repository documenting school-level instruction modes during the 2020-21 school year. Building upon seminal works by Jack et al. (2023), which estimated larger learning losses among grade 3-8 students in districts with less in-person instruction, this study extends its focus to high school students across six U.S. states: Arizona, Colorado, Georgia, Illinois, Indiana, and Wisconsin. This selection of states is primarily based on data availability and quality, which are explained in detail in Section III. High school students across states generally have fewer standardized testing requirements than their younger counterparts, making comprehensive data collection more challenging. These six states provided relatively robust and accessible datasets for analysis. In addition to evaluating test proficiency rates, akin to prior research, this study explores the effect of instruction mode on dropout rates, an education metric of special relevance to the high school population. Potential disparities in outcomes across racial and income groups are also examined through attribute interactions. Moreover, alongside the aggregated data, regression analyses encompassing state-mode interactions are utilized to discern whether the estimated effects remain consistent across diverse educational and social contexts. In synthesizing these efforts, this study provides nuanced insights into the impact of instruction mode on high school students' education outcomes, guiding policymakers and prompting continued research in fostering an effective and inclusive learning environment.

II. Literature Review

Since spring 2020, researchers worldwide have conducted comparative analyses of student performance pre- and post-pandemic, largely without direct reference to instruction mode. The majority highlighted adverse effects of COVID-19 related

This research was assisted by a grant administered by Northwestern University's Office of Undergraduate Research. Special thanks to Prof. Arvind Krishna, Prof. Charles Manski, and Prof. Joel L. Horowitz for their invaluable guidance and support.

disruptions on academic outcomes. For instance, Dorn et al. (2021) revealed that students testing in spring 2021 lagged behind by 10 points in math and 9 points in reading compared to their counterparts from previous years. Similarly, Maldonado and De Witte (2020) observed significant learning losses among grade 4-6 students in Belgium following school closures. Conversely, some studies indicated positive outcomes associated with distance education. Clark et al. (2021) demonstrated that online lessons delivered by high-quality teachers produced better exam results. More evidence suggested a beneficial influence of online education on math test scores among high school students in China (Zhang et al., 2021).

Studies on the relationship between pandemic instruction mode and test scores are pioneered by Jack et al. (2023), who established a public repository termed the COVID-19 School Data Hub (CSDH) for their analyses. The CSDH sourced from education agencies to provide school-level or district-level learning mode information during the 2020-21 school year. Utilizing this repository, along with other data from eleven U.S. states, they estimated standard difference-in-difference regressions and concluded that a shift from fully virtual to fully in-person is predicted to mitigate learning losses by 13.4% in math and 8.3% in ELA for grade 3-8 students. Furthermore, districts with larger populations of Black students were shown to benefit more from in-person instruction. To date, however, whether similar effects of instruction mode apply to higher-grade students in the United States remains largely unknown. Studies conducted abroad indicated varying impacts of distance education across age groups. Specifically, Tomasik et al. (2014) used extensive data from Switzerland to reveal that primary school students were more adversely affected by virtual learning than secondary school students. Given these potential disparities, education leaders require additional insights to design appropriate schooling policies for students beyond grade 3-8.

In addition to test proficiency rates, this study analyzes the relationship between instruction mode and dropout rates. High school dropout is a pivotal educational concern due to its profound implications for youth, including heightened unemployment, crime, and substance abuse (Belfield, 2021). A wealth of literature delves into the retrospective factors contributing to student attrition, predominantly emphasizing the lack of interest in classes and academic motivation (Bridgeland et al., 2006). While many anticipated a decline in graduation rates following school closures, Harris and Chen (2022), drawing on data from twenty-five U.S. states, suggested otherwise. However, they also acknowledged the influence of multiple factors, notably the temporary relaxation of graduation standards. The bulk of research examining the relationship between instruction mode and dropout rates has focused on higher education settings. For example, Patterson and McFadden (2009) illustrated that attrition rates were significantly higher in online master's programs compared to their campus-based counterparts. Nevertheless, the reasons behind adult dropout may systematically differ from those of high school students, who have distinct educational objectives (Barbour and Reeves, 2009). Given the ongoing surge in K-12 enrollment in online courses, investigating whether high school dropout rates are impacted by instruction mode changes remains a pertinent endeavor.

III. Data

The analyses of impacts of pandemic instruction mode on high school students' education outcomes involve three groups of data: (1) school-level instruction mode (% Virtual, % Hybrid) during the 2020-21 school year; (2) school-level test proficiency rates and dropout rates, which measure the percentage of students who score at or above the state's standard in math or ELA and who drop out of high school respectively; and (3) school-level demographic characteristics (% Black, % Hispanic, % White, % Economically Disadvantaged), student enrollment, and the number of tested students. To reduce bias caused by unbalanced sample sizes, only schools with all required data from the 2016-17 to the 2018-19 school year and the 2020-21 school year are considered. Data for the 2019-20 school year is not available due to nationwide testing cancellation and reporting hurdles following the COVID-19 outbreak in March 2020. Below are detailed descriptions of each data source:

A. Pandemic Instruction Mode Data

Pandemic instruction mode data are obtained from the COVID-19 School Data Hub. This public repository classifies a time period in the 2020-21 school year as either 1) "in-person" (if all or most students received instruction in person, 5 days a week); 2) "hybrid" (if all or most students received a blend of in-person and virtual instruction); or 3) "virtual" (if all or most students received instruction virtually, 5 days a week). Only states that provide school-level instruction mode data at monthly, bi-weekly, or weekly intervals are considered in the analyses. For every school on record, the respective shares of virtual and hybrid learning are calculated as the total number of days students spent in each instruction mode divided by the total number of days for the entire school year.

B. Assessment Data

To measure changes in math and ELA proficiency rates, state-level standardized assessment data are utilized. Unlike grade 3-8, high school students are not consistently required to participate in annual state-level assessments. Even in states where such tests are administrated annually, the data comparability, availability, and completeness are often compromised, posing significant challenges for large-scale analyses. Moreover, during the pandemic, many states either altered assessment content or relaxed testing requirements. For instance, California truncated assessment duration from 4 hours to 2 hours, accompanied by substantial modifications on testing content. In spring 2021, New York City permitted testing exemption for students meeting specific educational criteria, resulting in an extremely low participation rate. These problems could introduce severe bias and impede comparisons. Therefore, both quantitative aspects (e.g., test participation rates) and qualitative aspects (e.g., test content and scope) of state-level assessments are evaluated to ensure that the states included in this study experienced minimal changes and had relatively comparable test scores before and during the pandemic.

C. Dropout Data

In a parallel manner to assessment data, dropout data are sourced from individual states' public education repositories. A significant concern in dropout rate research revolves around the inconsistency in defining ”dropout.” For instance, Illinois altered the dropout calculation metric since the 2018-19 school year, resulting in a huge spike compared to preceding years. To safeguard the accuracy and applicability of regression results, this study exclusively incorporates states that employ comparable and consistent calculation metrics across all included years: Arizona, Colorado, Georgia, and Wisconsin. Specifically, dropout rates refer to the proportion of students who drop out of high school during a twelve-month reporting period. It is calculated as the number of dropouts divided by the number of enrolled students during the respective school year.

D. Student Demographics Data

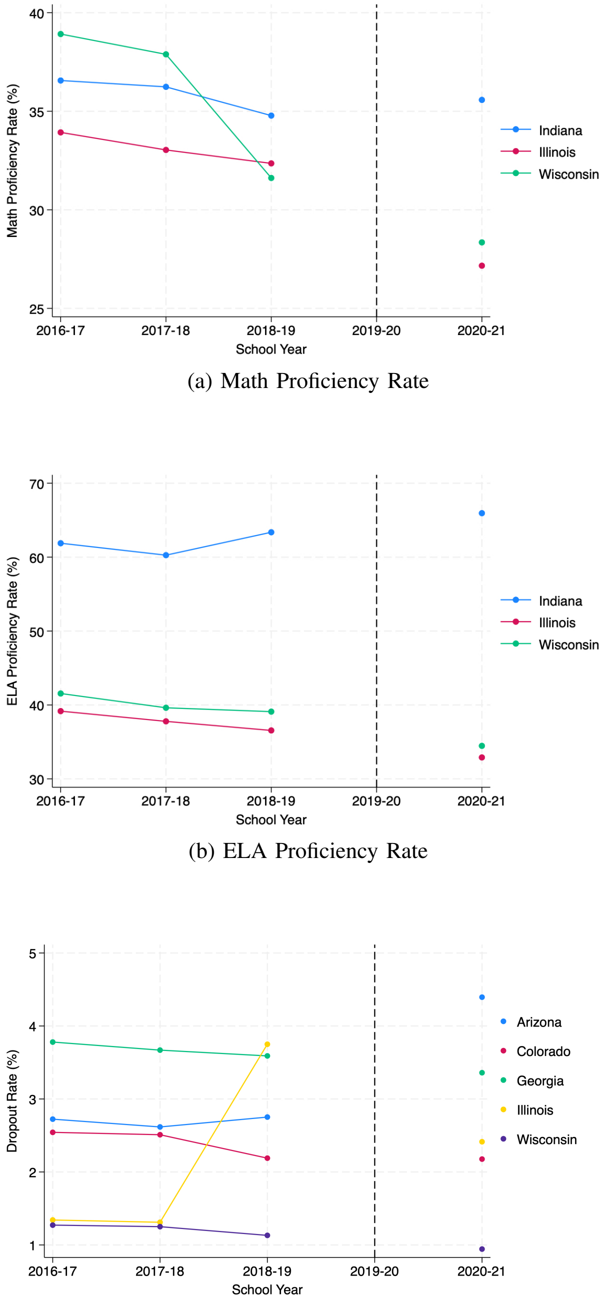
In addition to the primary data sources, demographic characteristics data are obtained from states' education repositories and utilized as controls in the regression analyses. These data encompass school-level information regarding the proportion of enrolled students by race and ethnicity and socioeconomic status, as indicated by the eligibility for Free or Reduced Price Meals (FRPM). The final variables employed comprise percentage of Black, Hispanic, White, and economically disadvantaged (ED) students.

Finally, by merging pandemic instruction mode data, education outcome data, and demographic characteristics data, a school-year level panel data file spanning from the 2016-17 to 2018-19 school year and the 2020-21 school year is constructed. The sample contains a total of six states (Arizona, Colorado, Georgia, Illinois, Indiana, and Wisconsin). Among them, three provide suitable data for math and ELA proficiency rates analyses (Illinois, Indiana, and Wisconsin) and four offer appropriate data for dropout rates analyses (Arizona, Colorado, Georgia, and Wisconsin). Table B.1 in the Appendix shows summary statistics for each of the datasets used in the final analyses. The code to construct all data files is available in the data folder of this GitHub repository.

E. Descriptive Analysis

Figures 1 illustrates changes in test proficiency and dropout rates across different states. Regarding math proficiency rate, Illinois and Wisconsin experienced notable decreases of 5.94% and 5.19%, respectively, from the baseline (averaged across all pre-pandemic years) and 7.79% and 3.27% from the previous year during the 2020-21 school year. Given Wisconsin's significant decline in math proficiency rate in the 2018-19 school year, it remains uncertain whether the subsequent decrease

Fig. 1: Test Proficiency and Dropout Rate Change by State



(a) Math Proficiency Rate

(b) ELA Proficiency Rate

(c) Dropout Rate

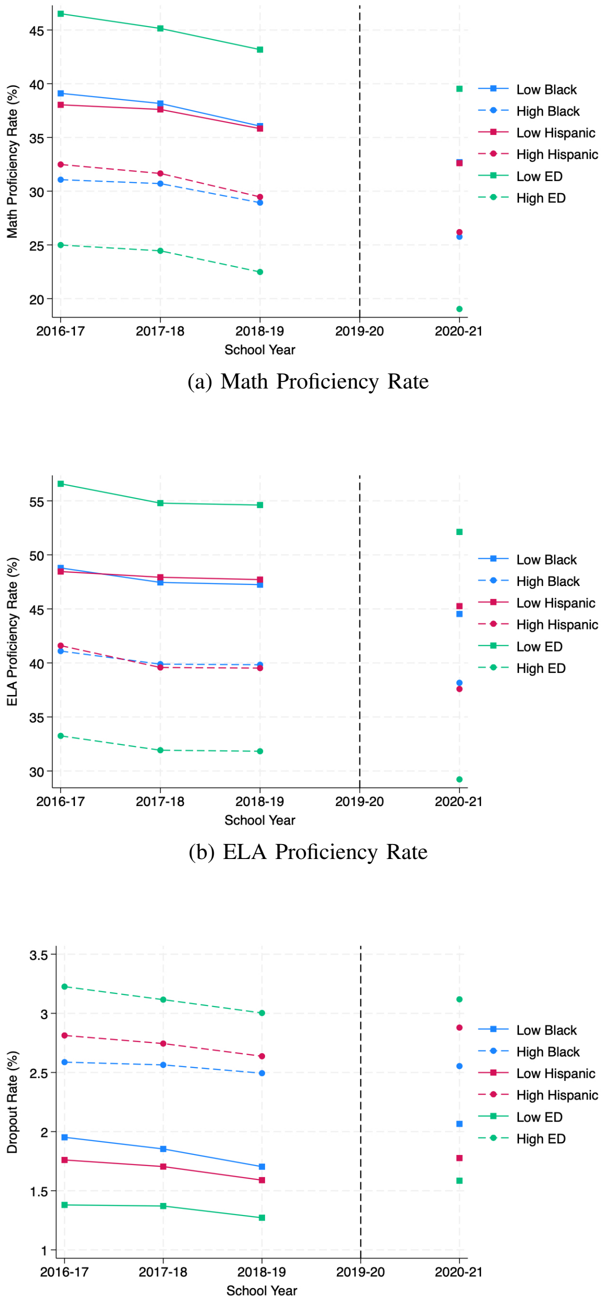
Note: This figure shows the average test proficiency and dropout rates from the 2016-17 school year to the 2020-21 school year, considering individual state samples. The data is weighted by the number of tested and enrolled students, respectively. Test data and dropout data for the 2019-20 school year are unavailable due to nation-wide testing cancellation and reporting hurdles following the COVID-19 outbreak in March 2020.

was primarily driven by COVID-19 or were indicative of pre-existing trends. Conversely, Indiana demonstrated only a marginal decrease of 0.28% from the baseline but a 0.8% increase from the previous year. In terms of ELA, proficiency rate changes varied, with Illinois and Wisconsin showing consistent modest declines while Indiana maintained a trend of gradual improvement. Collectively, plots on test proficiency rate changes do not indicate significant performance decreases among the high school population during the pandemic, contrary to existing findings concerning younger students. Similarly, dropout rate fluctuations varied significantly across states. Only Arizona and Illinois observe statistically significant changes in average dropout rates during the pandemic. In the 2020-21 school year, dropout rate in Arizona increased by 1.69% from the baseline and 1.55% from the previous year, which holds economic significance given the typically small magnitudes of dropout rates. Illinois witnessed a drastic surge in dropout rate in the 2018-19 school year; however, upon closer examination, this increase appeared to stem from modifications in Illinois' dropout calculation metric. Due to its inconsistency in the calculation of dropout, Illinois was excluded from subsequent dropout analyses. Dropout rates calculation in the remaining four states remained comparable across all included years.

Figures 2 depicts changes in test proficiency and dropout rates across demographic attributes, aggregating outcomes from all included state samples. The data underscores significant achievement gaps that predated the pandemic, with schools having a high proportion of Black, Hispanic, or ED students showing lower test proficiency rates and higher dropout rates, on average. However, contrary to widespread reports during the pandemic that document larger learning losses among marginalized groups, the data does not present evidence of widening achievement gaps attributable to COVID-19. Rather, schools with a high proportion of Black, Hispanic, or ED students appear to have experienced smaller increase in dropout rates compared to their counterparts.

Figures 3 details changes in test proficiency and dropout rates by instruction mode. In contrast to the main descriptive findings of Jack et al. (2023), schools with higher proportions of virtual or hybrid learning do not appear to experience larger decrease in test performance compared to schools with more in-person instruction. Another notable observation is the disparity in students' educational outcomes prior to the pandemic based on instruction modes during the pandemic. These plots suggest that access to virtual learning during the pandemic may correlate with factors influencing test proficiency and dropout rates. Table B.2 in the Appendix illustrates the pairwise correlations between demographic characteristics and virtual learning during the 2020-21 school year. It indicates a higher prevalence of virtual learning in schools with larger Black, Hispanic, or ED student populations in both the test and dropout sample data. Essentially, the plots, along with the correlation metric, imply that schools with a higher proportions of historically underserved students were more likely to offer virtual instruction. This insight aligns with previous research

Fig. 2: Test Proficiency and Dropout Rate Change by Student Demographics



(a) Math Proficiency Rate

(b) ELA Proficiency Rate

(c) Dropout Rate

Note: This figure shows the average test proficiency and dropout rates from the 2016-17 to the 2020-21 school year, considering individual state samples. The data is weighted by the number of tested and enrolled students, respectively. Comparisons are presented by the share of students who are Black, Hispanic, or Economically Disadvantaged (ED). The "Low" and "High" categories correspond to schools with below and above weighted median by state, respectively.

and strengthens the rationale for incorporating demographic characteristics in subsequent regression analyses to prevent omitted variable bias.

Lastly, it is crucial to acknowledge that these changes alone cannot be used to draw significant conclusions about the effect of instruction mode on education outcomes. An identification strategy is essential to disentangle other factors and biases, such as unobserved heterogeneity. Additionally, confidence intervals of weighted point estimates in each figure are available in the descriptive analysis folder of this GitHub repository. Attention to statistical significance reinforces the need for caution in drawing insights solely from descriptive graphs.

IV. Methodology

To examine the relationship between instruction mode and education outcomes, a weighted least square regression model with time and entity fixed effects is utilized. Existing literature suggests that schools with larger Black, Hispanic, or ED student populations may respond differently to factors affecting education, which include learning mode changes. As briefly mentioned in Section I, a primary contribution of this study is to discern whether the estimated impact remains consistent across diverse educational and social contexts. To address this, attribute interactions are included in the model specification to account for heterogeneous treatment effects. The general structure of regression analyses incorporating interaction terms is outlined below:

$$\begin{aligned} Y_{s,t} = \alpha X_{s,t} + \beta V_{s,t} + \delta V_{s,t} \cdot INT_{s,t} \\ + \eta_s + \tau_t + \varepsilon_{s,t} \end{aligned} \tag{1}$$

with specific interaction:

$$\begin{aligned} V_{s,t} \cdot INT_{s,t} = V_{s,t} * \text{State} + V_{s,t} * \text{Black}_{s,t} \\ + V_{s,t} * \text{Hispanic}_{s,t} + V_{s,t} * \text{ED}_{s,t} \end{aligned} \tag{2}$$

Here, $Y_{s,t}$ represents either the math proficiency rate, ELA proficiency rate, or dropout rate in school s and year t, while $X_{s,t}$ denotes the demographic characteristics in school s and year t. $V_{s,t}$ signifies the respective share of instruction mode. As detailed in Section III.1, the share of virtual and hybrid learning is computed separately for each school, making both % Virtual and % Hybrid primary explanatory variables in the regression analyses. η_s and τ_t represent the entity and time fixed effects, respectively.

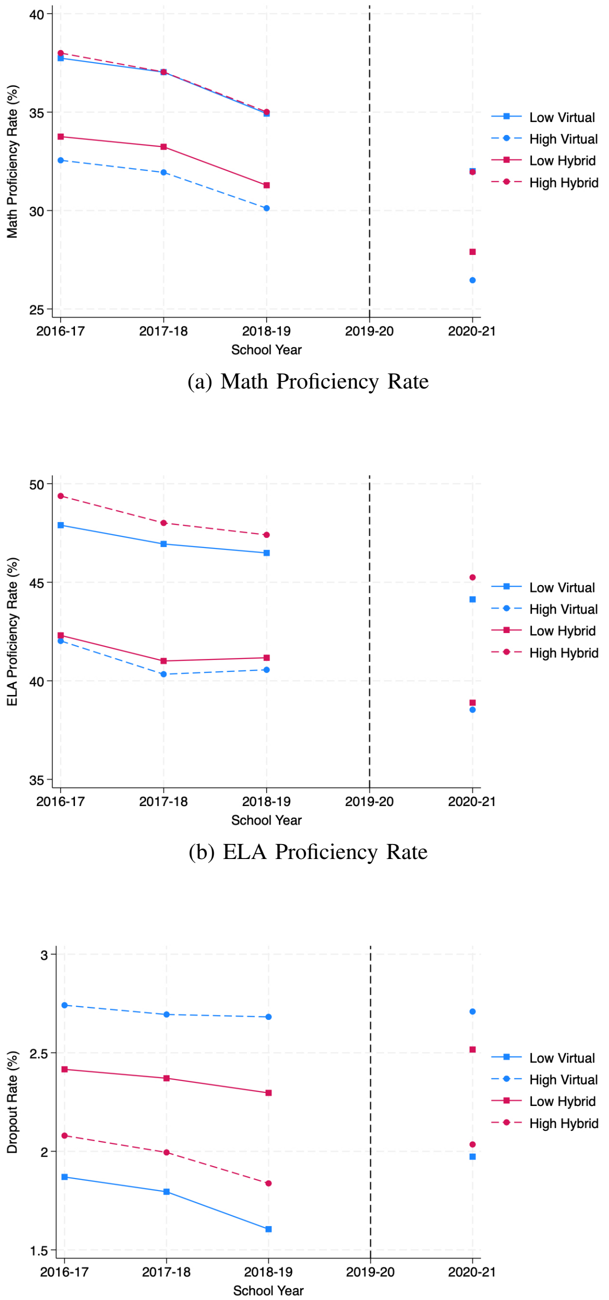
While β indicates the predicted change in $Y_{s,t}$ for the baseline when virtual or hybrid learning increases from 0% to 100%, δ gauges how the estimated impact changes with the interacted population. A significant δ across interactions suggests a substantial heterogeneous effect of virtual or hybrid learning on education outcomes.

V. Model Diagnostics

A. Model Selection and Assumption

The use of panel regression model with two-way fixed effects in this study is justified by domain knowledge. It is reasonable to assume that both time-specific (yet school-invariant)

Fig. 3: Test Proficiency and Dropout Rate Change by Instruction Mode



(a) Math Proficiency Rate

(b) ELA Proficiency Rate

(c) Dropout Rate

Note: This figure shows the average test proficiency and dropout rates from the 2016-17 to the 2020-21 school year, considering individual state samples. The data is weighted by the number of tested and enrolled students, respectively. Comparisons are presented by the share of virtual and hybrid learning during the 2020-21 school year. The ”Low” and ”High” categories correspond to schools with below and above weighted median by state, respectively.

and school-specific (yet time-invariant) characteristics of the sample would influence education outcomes. For instance, test proficiency and dropout rates may witness greater fluctuations during the 2020-21 school year compared to pre-pandemic periods. Similarly, unobserved time-invariant characteristics at the school level, such as school culture, resources, and leadership quality, may systematically influence educational outcomes. Since it is impractical to obtain and control for all unobserved factors, two-way fixed effects models are particularly suitable for this study due to their ability to control for both time-invariant and school-invariant unobservables. In contrast, random effects models assume that the unobserved factors are uncorrelated with the observed independent variables, which may not hold true in practice. Additionally, the use of fixed effects allows for the identification of within-school changes over time, providing more precise estimates of the influence. Furthermore, two-way fixed effects models are robust to omitted variable bias, which is significant to this study as it helps ensure that the estimated effects of the variables of interest are not confounded by omitted variables that could otherwise bias the results. Therefore, by accounting for both time and entity fixed effects, this study ensures more reliable and interpretable results in examining the relationship between instruction mode and educational outcomes.

In addition to intuitive insights, rigorous statistical analyses are employed to validate the appropriateness of fixed effects models in addressing the research question. Analyses on panel data, combining longitudinal and cross-sectional characteristics, typically involve the evaluation of three models: (1) Pooled Ordinary Least Squares (OLS) regression with no panel effects; (2) Panel OLS with fixed effects; and (3) Panel OLS with random effects. Both the Breusch-Pagan test and partial F-test on panel effects consistently favor fixed effects models over pooled OLS alternatives, indicating their superiority in capturing nuanced temporal and spatial variations. Subsequently, the Durbin-Wu-Hausman test confirms the suitability of fixed effects models over random effects models for this study. Following model selection, various diagnostic tests are conducted to assess key assumptions of fixed effects models, including the absences of serial correlation and perfect multicollinearity and homoscedasticity in residuals. The Durbin-Watson test and Variance Inflation Factor (VIF) calculation confirm the absences of serial correlation and perfect multicollinearity within the sample. However, the presence of heteroscedasticity, identified by the Breusch-Pagan test, necessitates the use of robust standard errors in model fitting to ensure the validity of statistical inferences. Results of diagnostic tests performed in the modeling process are presented in Table B.3 in the Appendix.

B. Influential Point Analysis

Preliminary regression analyses were conducted on the acquired dataset. To ensure the validity and generalizability of results, statistical techniques were employed to identify influential observations, which are data points that significantly impact the estimated coefficients. Since DFBETAS, a common influence measure, gauges how much a particular regression coefficient changes when an individual observation is removed from the dataset, this study computed DFBETAS values pertaining parameters of interest for each observation and compared them with the critical value, calculated as 2 divided by the square root of sample sizes. Upon analysis, 193 out of 5724 observations were flagged as influential for regressions on math, while 157 out of 5724 observations were identified as influential for ELA. Additionally, 123 out of 5592 observations were deemed influential for dropout. Notably, over 95% of these influential observations occurred in the 2020-21 school year. An important consideration arose regarding whether to remove influential observations in a specific year or across all years. Since % Virtual and % Hybrid were consistently set to zero from 2016-17 to 2018-19, simply removing the influential observation in the 2020-21 school year would not be prudent, as the remaining pre-COVID years lack meaningful information on the effect of instruction mode on educational outcomes. Therefore, instead of removing individual observations, influential entities (schools with influential observations in any given year) were targeted for removal to maintain data integrity. It is noteworthy that the removal of influential entities was not undertaken blindly; rather, a thorough data inspection was conducted to ensure the validity of the decision-making process. Specifically, entities exhibiting significant influences or obvious data abnormalities were removed. These abnormalities typically manifest in three areas: (1) education outcome variables; (2) demographic characteristics variables; and (3) the number of students tested or enrolled. Table I below presents specific examples illustrating each case pertaining to regression analyses on math proficiency rates.

TABLE I: Examples of Removed Influential Entities

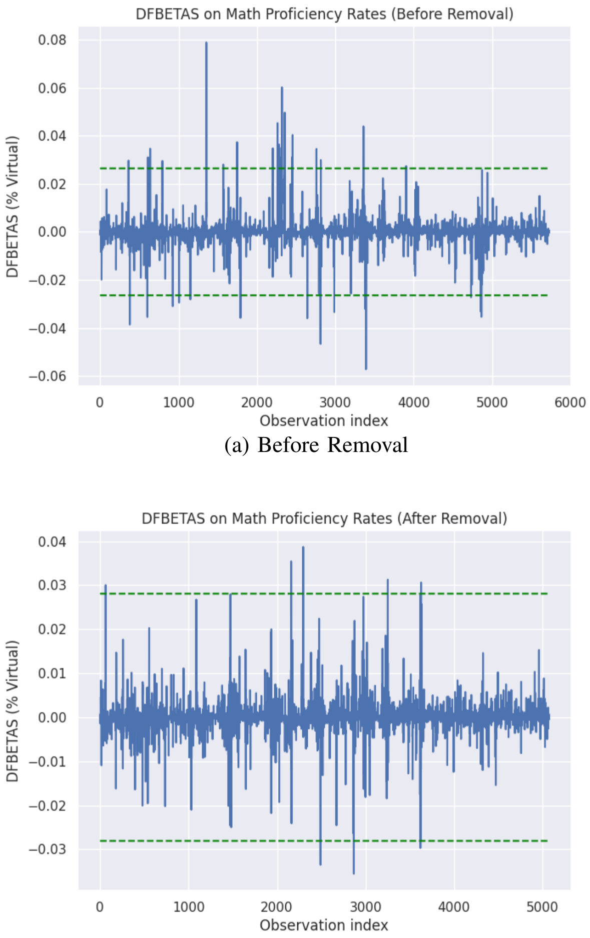
Panel A: Abnormal Demographics Variations						
Year	# Tested	% Pass	% Black	% Hispanic	% White	% ED
17	416	8.3	78.7	19.5	0.5	9.6
18	377	6.9	77.2	21.5	0.0	53.6
19	437	7.8	73.6	24.4	0.2	64.1
21	314	6.7	67.4	29.4	0.0	87.7
Panel B: Abnormal Outcome Variations						
Year	# Tested	% Pass	% Black	% Hispanic	% White	% ED
17	122	33.6	8.4	43.3	42.6	76.99
18	106	13.2	10.9	39.8	42.7	64.79
19	114	16.7	11.1	37.1	45.5	68.55
21	112	9.8	12.2	40.1	40.7	68.66
Panel C: Significantly Imbalanced Sample Size						
Year	# Tested	% Pass	% Black	% Hispanic	% White	% ED
17	352	6.8	47.1	44.3	2.6	24.6
18	472	7.8	47.2	47.8	2.1	17.9
19	426	3.1	43.1	53.1	0.8	25.3
21	282	3.9	39.8	57.1	0.9	31.9

Note: Panels A-C present reasons for removal, and each sub-table denotes a specific school identified with the data issue outlined in each panel. Notably, instruction mode variables (% Virtual and % Hybrid) are not considered in the influential points removal process, as they reasonably vary between 0% and 100% in the 2020-21 school year and remain at 0% during pre-pandemic periods.

Finally, 163, 114, and 173 schools (influential entities) were removed for math, ELA, and dropout regression analyses respectively, leaving the final datasets with 1268, 1317, and 1268 schools (or 5072, 5268, and 5072 observations) respectively.

After identifying and removing influential entities with data quality issues, the computation of DFBETAS pertaining each parameter of interest was repeated. Figure 4 below illustrates an example of DFBETAS comparison before and after the removal of influential points pertaining the % Virtual parameter in math regressions. Comprehensive comparisons of DFBETAS for all regression analyses are included in Figure A.1-3 in the Appendix. The comparison reveals a considerable improvement in the stability and reliability of the results post-removal of influential points.

Fig. 4: Examples of DFBETAS Comparisons on Math Proficiency Rates



(a) Before Removal

(b) After Removal

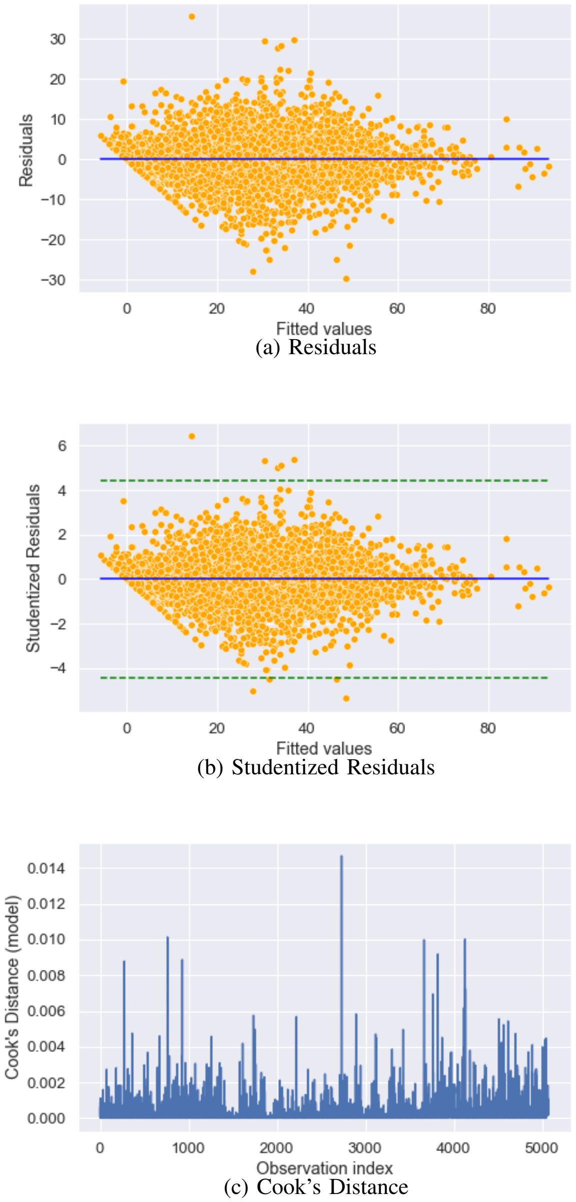
Note: This figure presents a comparison of DFBETAS values before and after the removal of influential points. The blue bars indicate the DFBETAS value for all included observations concerning the % Virtual parameter on math proficiency rates. The green dashed line depicts the critical value used to identify influential observations, calculated as $\frac{2}{\sqrt{n}}$

While this technique is not commonly utilized in social science research involving panel data, this study demonstrates its value in enhancing the robustness and generalizability of regression results. By systematically identifying and removing influential entities, especially those with data quality issues, this study ensures that the estimated coefficients of the final models are more accurate and representative of the total population. Moreover, this study addresses a fundamental challenge in influential analysis involving panel data: the decision between removing influential observations of a specific year versus influential entities across all years. By elucidating a clear framework on this matter, this study offers valuable insights for relevant research endeavors, particularly those investigating the short-term impacts of temporary treatment effects, such as Jack et al. (2023) research on grade 3-8 student populations. Furthermore, the analysis sheds light on the inherent randomness and noise present in granular school-level education data, especially during times of disruption such as the COVID-19 pandemic. In such contexts, the ability to discern and remove influential data points that could potentially bias results becomes imperative.

C. *Diagnostic Visualizations*

The residual, studentized residual, and Cook's distance plots of the final regression models for both ELA proficiency and dropout rates are depicted in Figure A.4-5 in the Appendix. Below, Figure 5 shows the results for math proficiency rate as an illustrative example. Upon examination of these figures, it is observed that the residuals are distributed relatively randomly around zero. Although a few outliers are detected from the studentized residual plots, these observations, along with their corresponding entities, do not exhibit data quality issues and are deemed non-influential due to their low leverages. Moreover, the Cook's distance metric for each observation indicates that data points included in the final data sample do not exert significant influence on the regression models, as their values remain well below the critical threshold. These diagnostic visualizations further provide assurance regarding the robustness and reliability of the estimated coefficients.

Fig. 5: Diagnostic Visualizations for Math Proficiency Rates



(a) Residuals

(b) Studentized Residuals

(c) Cook's Distance

Note: This figure displays residuals against fitted values, studentized residuals against fitted values, and Cook's distance for each observation in regression analyses for math proficiency rates. All plots are generated using the final dataset after removing influential points. In the studentized residual plot, the green dashed line represents critical values for outliers, calculated at the 95% confidence interval. The blue bars in the Cook's distance plot denote Cook's distance of each observation included in the final model. The cutoff for highly influential points was calculated as 0.99, which is significantly larger than the Cook's distance of any observation.

VI. Results

The results of model (1) for math proficiency rates are summarized in Table II. Notably, the interaction between % ED and % Virtual is excluded from the final model due its high variance inflation factor (VIF), indicating significant multicollinearity. Further investigation reveals a strong correlation between the % ED and % Virtual interaction and the % Black and % Virtual interaction. Consequently, excluding the % ED and % Virtual interaction ensures that the marginal effects of virtual learning on the proportion of Black students are accurately reflected. Following the removal, VIFs are recalculated for parameters of interest to validate the model specifications. This adjustment is applied to subsequent regressions. Detailed statistical analyses of this process are provided in Table B.3 in the Appendix.

Table II shows significant disparate effects across states. Illinois and Wisconsin both exhibit pronounced negative impacts of virtual and hybrid learning on math proficiency rates. Specifically, the baseline virtual coefficient and the sum of coefficients, including state-mode interaction terms, suggest that a complete transition from fully in-person to fully virtual learning is expected to decrease math proficiency rates by 6.84% and 9.53% for Illinois and Wisconsin, respectively. Comparable negative effects are observed for hybrid learning within these two states. In contrast, virtual learning demonstrates no significant impact in Indiana. Furthermore, a full transition to hybrid learning is estimated to increase math proficiency rates by 5.12% in Indiana within a 90% confidence interval.

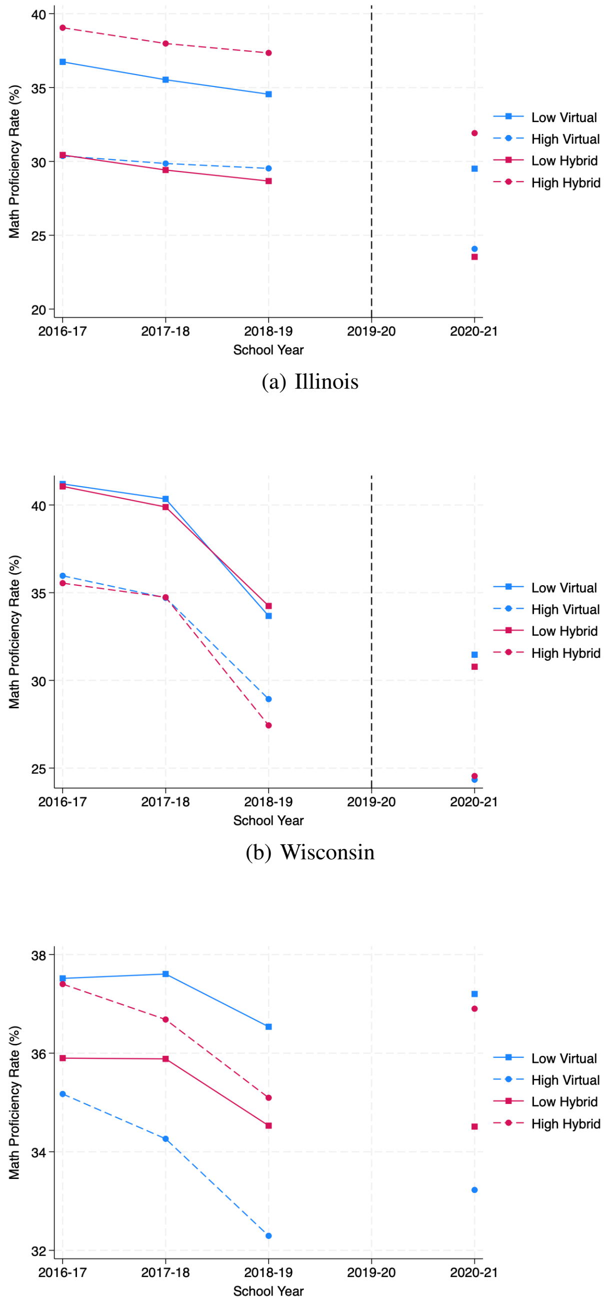
These findings are visually supported by Figure 6, where the top two plots represent Illinois and Wisconsin, respectively, while the bottom plot depicts Indiana. During pre-pandemic periods, from the 2016-17 school year to the 2018-19 school year, schools with low or high proportions of virtual or hybrid learning generally exhibited parallel trends. However, in the 2020-21 school year, schools with higher proportions of virtual or hybrid learning in Illinois and Wisconsin experienced a larger decrease in math proficiency rates compared to their counterparts. This phenomenon is particularly notable for virtual learning. Conversely, in Indiana, the situation was reversed. Schools with higher proportions of virtual or hybrid learning witnessed a greater increase in math proficiency rates during the 2020-21 school year. Specifically, while schools with low hybrid learning maintained similar levels of math proficiency rates between the 2018-19 and the 2020-21 school years, schools with high hybrid learning increased from 35.3% in 2018-19 to around 37% in 2020-21, reflecting the positive hybrid learning effect shown in the regression table.

TABLE II: Learning Mode and Changes in Math Proficiency Rates

	Math Proficiency Rates (%)	
Influential Points Analysis	**Before Removal**	**After Removal**
Virtual %	-4.98***	-6.84***
	(-7.70, -2.26)	(-9.18, -4.49)
Hybrid %	-6.85***	-6.21***
	(-8.69, -5.01)	(-7.83, -4.59)
IN × Virtual %	9.21***	9.08***
	(6.48, 11.93)	(6.39, 11.77)
IN × Hybrid %	10.11***	11.34***
	(6.28, 13.93)	(6.63, 16.04)
WI × Virtual %	-2.45	-2.69*
	(-5.67, 0.77)	(-5.75, 0.36)
WI × Hybrid %	-0.97	-1.48
	(-3.17, 1.24)	(-3.56, 0.59)
Black % × Virtual %	0.03*	0.05***
	(0.00, 0.07)	(0.02, 0.08)
Black % × Hybrid %	0.00	0.01
	(-0.05, 0.05)	(-0.03, 0.05)
Hispanic % × Virtual %	-0.04	0.03
	(-0.09, 0.01)	(-0.01, 0.06)
Hispanic % × Hybrid %	0.05**	0.01
	(0.01, 0.09)	(-0.03, 0.05)
ED % × Hybrid %	0.06***	0.06***
	(0.02, 0.11)	(0.02, 0.10)
Model Summary		
Observations	5724	5072
Degree of Freedom	1448	1285
R-squared	0.941	0.945
F-statistics	170.9	185.0
Sum of Coefficients		
Virtual % + IN × Virtual %	4.22**	2.24
	(0.67, 7.79)	(-1.21, 5.69)
Virtual % + WI × Virtual %	-7.43***	-9.53***
	(-10.94, -3.91)	(-12.72, -6.34)
Hybrid % + IN × Hybrid %	3.25	5.12*
	(-0.93, 7.44)	(-0.03, 10.28)
Hybrid % + WI × Hybrid %	-7.82***	-7.70***
	(-10.46, -5.18)	(-10.24, -5.16)

*Note: This table illustrates the relationship between instruction mode and math proficiency rates before (left) and after (right) removing influential points for Illinois, Indiana, and Wisconsin. The results of model (1) are presented in the first sub-table. The second sub-table presents the sum of coefficients for state and instruction mode interactions to evaluate the effect of instruction mode on individual state samples. All regressions are weighted by the number of tested students and include both year and school fixed effects. 95% confidence intervals, calculated with robust standard errors, are reported in parentheses. Coefficients are denoted with ***, **, or * for P-values less than 0.01, 0.05, and 0.1, respectively.*

Fig. 6: Math Proficiency Rate Changes by State and by Instruction Mode



(a) Illinois

(b) Wisconsin

(c) Indiana

Note: This figure shows the average math proficiency rates from the 2016-17 to the 2020-21 school year for Illinois, Wisconsin, and Indiana, respectively. The data is weighted by the number of tested students. Comparisons are presented by the share of virtual and hybrid learning during the 2020-21 school year. The "Low" and "High" categories correspond to schools with below and above weighted median by state, respectively.

The results for ELA proficiency rates are detailed in Table III. Overall, the impact of instructional mode on test proficiency rates is more pronounced in ELA compared to math, both in terms of magnitude and statistical significance. Additionally, divergent effects across states persist, mirroring those observed for math proficiency rates. Virtual and hybrid learning are estimated to decrease ELA proficiency rates in Illinois and Wisconsin while increasing ELA proficiency rates in Indiana. The positive effect of virtual learning in Indiana is particularly notable, where a transition from 0% to 100% virtual is projected to increase ELA proficiency rates by 12.45%. Despite Indiana initially having higher ELA proficiency rates on average compared to Illinois and Wisconsin, this estimated increase maintains economic significance. These findings are further supported by Figure 7, which visually illustrates the trends. For instance, it is evident that schools with higher proportions of virtual learning, on average, experienced a larger decline in ELA proficiency rates than those with lower proportions in Wisconsin. Conversely, schools with higher proportions of virtual learning in Indiana saw a significantly larger increase in ELA proficiency rates during the 2020-21 school year.

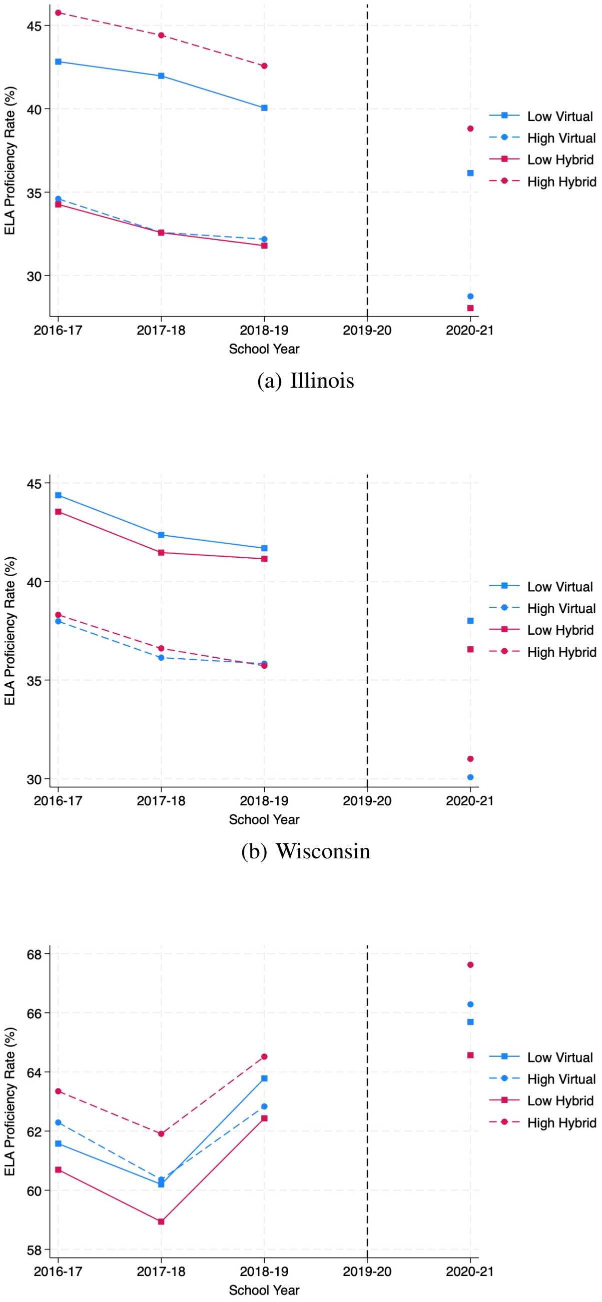
Regarding the marginal effects of instructional mode on historically underserved student populations, evidence suggests that schools with larger proportions of Black or Hispanic students may benefit more from hybrid learning compared to those with lower proportions. This is indicated by the significantly positive coefficients of the interactions between the percentages of Black or Hispanic students and the percentage of hybrid learning.

TABLE III: Learning Mode and Changes in ELA Proficiency Rates

	ELA Proficiency Rates (%)	
Influential Points Analysis	Before Removal	After Removal
Virtual %	-4.46***	-4.16***
	(-7.05, -1.87)	(-6.67, -1.66)
Hybrid %	-7.98***	-8.00***
	(-9.60, -6.36)	(-9.65, -6.36)
IN × Virtual %	16.33***	16.62***
	(12.35, 18.99)	(13.26, 19.98)
IN × Hybrid %	11.33***	12.34***
	(8.46, 14.19)	(8.50, 16.18)
WI × Virtual %	-5.22***	-3.70***
	(-8.34, -2.10)	(-6.26, -1.13)
WI × Hybrid %	3.58***	0.05
	(-1.55, 2.46)	(-1.97, 2.06)
Black % × Virtual %	-0.01	-0.02
	(-0.04, 0.03)	(-0.05, 0.01)
Black % × Hybrid %	0.04	0.09***
	(-0.01, 0.08)	(0.04, 0.13)
Hispanic % × Virtual %	-0.03*	-0.04*
	(-0.07, 0.01)	(-0.07, 0.00)
Hispanic % × Hybrid %	0.06***	0.09***
	(0.03, 0.10)	(0.06, 0.13)
ED % × Hybrid %	0.03	0.00
	(-0.01, 0.07)	(-0.04, 0.04)
Model Summary		
Observations	5724	5268
Degree of Freedom	1448	1334
R-squared	0.959	0.960
F-statistics	262.0	262.5
Sum of Coefficients		
Virtual % + IN × Virtual %	11.86**	12.45**
	(8.72, 15.03)	(8.08, 16.11)
Virtual % + WI × Virtual %	-9.68***	-7.86***
	(-13.23, -6.13)	(-10.83, -4.89)
Hybrid % + IN × Hybrid %	3.35**	4.34**
	(0.39, 6.30)	(0.10, 8.56)
Hybrid % + WI × Hybrid %	-4.40***	-7.95***
	(-7.45, -1.36)	(-10.24, -5.50)

*Note: This table illustrates the relationship between instruction mode and ELA proficiency rates before (left) and after (right) removing influential points for Illinois, Indiana, and Wisconsin. The results of model (1) are presented in the first sub-table. The second sub-table presents the sum of coefficients for state and instruction mode interactions to evaluate the effect of instruction mode on individual state samples. All regressions are weighted by the number of students tested and include both year and school fixed effects. 95% confidence intervals, calculated with robust standard errors, are reported in parentheses. Coefficients are denoted with ***, **, or * for P-values less than 0.01, 0.05, and 0.1, respectively.*

Fig. 7: ELA Proficiency Rate Changes by State and by Instruction Mode



(a) Illinois

(b) Wisconsin

(c) Indiana

Note: This figure shows the average ELA proficiency rates from the 2016-17 to the 2020-21 school year for Illinois, Wisconsin, and Indiana, respectively. The data is weighted by the number of tested students. Comparisons are presented by the share of virtual and hybrid learning during the 2020-21 school year. The "Low" and "High" categories correspond to schools with below and above weighted median by state, respectively.

Regarding dropout rates, the impact of instructional mode varies considerably across states in terms of both direction and significance. Among the four state samples analyzed, only Wisconsin demonstrates negative effects of hybrid learning on dropout rates, suggesting a projected decrease of 0.48% in dropout rates with a full transition to hybrid learning. This effect remains robust both before and after the removal of influential data points. In contrast, Arizona, the baseline state, shows significantly positive effects for both virtual and hybrid learning, indicating that distance education, in general, is predicted to increase high school dropout rates within this state. Similar conclusions can be drawn for the Georgia sample, where a full transition to virtual learning is estimated to increase dropout rates by 0.93%.

Figure 8 illustrates changes in dropout rates by instruction mode for Arizona, Georgia, and Wisconsin, respectively. While the regression results align with the trends observed in the Arizona plot, where schools with high virtual instruction experienced a considerably larger increase in dropout rates, the estimated coefficients from the regression results are not fully reflected in the Georgia or Wisconsin plots. In Georgia, descriptive evidence suggests that schools with more virtual learning exhibited a decrease in dropout rates, while those with less virtual learning showed an increase, potentially presenting conflicting results compared to the regression findings. However, it is crucial to note that these descriptive visualizations offer only superficial insights and do not consider other factors influencing dropout rates. On the contrary, the model specification, which incorporates school-level demographic information and two-way fixed effects, provides a more comprehensive approach to isolating the effect of instruction mode on education outcomes.

Considering the heterogeneous effects on marginalized communities, distance education is expected to have a more pronounced negative effect on dropout rates in schools with higher proportions of Black students compared to those with lower proportions. Taken together with insights gained from ELA proficiency rates, virtual and hybrid learning are estimated to narrow pre-existing achievement gaps, rather than widen them. The positive effects of hybrid learning on ELA proficiency rates, coupled with the decrease in dropout rates associated with remote instructions, may indicate a trend towards improved educational equity.

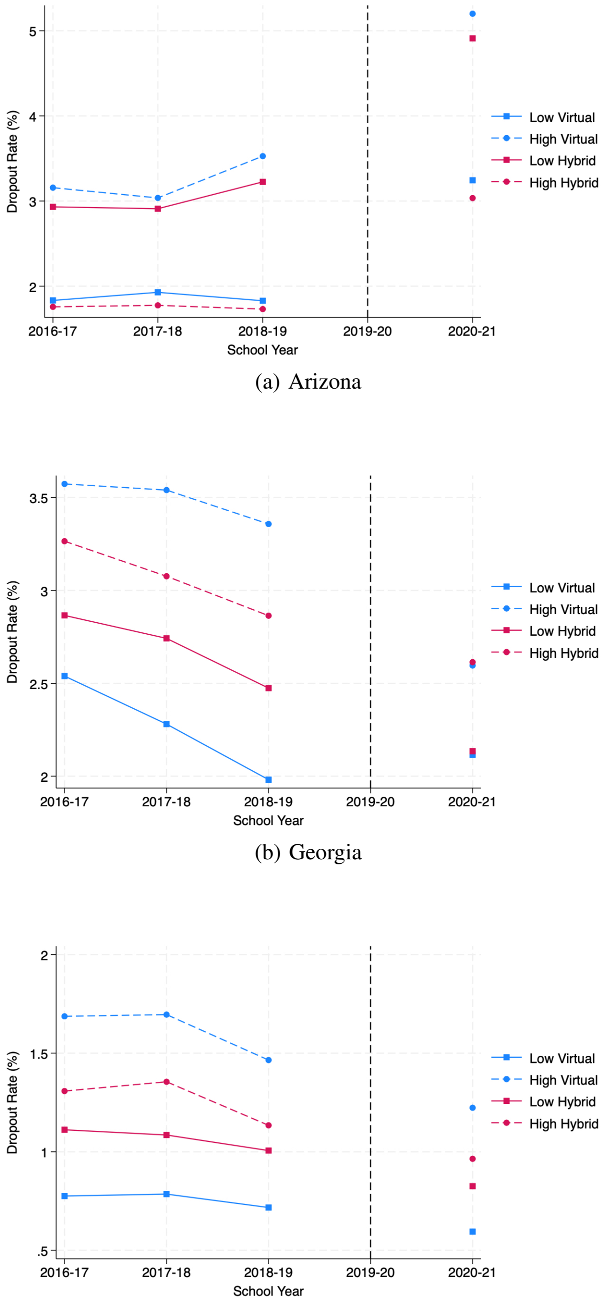
TABLE IV: Learning Mode and Changes in Dropout Rates

	Dropout Rates (%)	
Influential Points Analysis	Before Removal	After Removal
Virtual %	3.40^{***}	3.20^{***}
	(2.22, 4.58)	(2.59, 3.79)
Hybrid %	0.97^{***}	1.01^{***}
	(0.38, 1.56)	(0.68, 1.35)
CO × Virtual %	-3.28^{***}	-3.05^{***}
	(-4.06, -2.50)	(-3.56, -2.54)
CO × Hybrid %	-1.23^{***}	-1.00^{***}
	(-1.80, -0.67)	(-1.30, -0.69)
GA × Virtual %	-2.03^{***}	-2.26^{***}
	(-3.20, -0.85)	(-2.93, -1.59)
GA × Hybrid %	-0.60^{*}	-0.73^{***}
	(-1.24, 0.05)	(-1.25, -0.20)
WI × Virtual %	-3.00^{***}	-2.99^{***}
	(-3.90, -2.10)	(-3.54, -2.45)
WI × Hybrid %	-1.55^{***}	-1.49^{***}
	(-2.16, -0.93)	(-1.89, -1.10)
Black % × Virtual %	-0.03^{***}	-0.03^{***}
	(-0.04, -0.02)	(-0.04, -0.02)
Black % × Hybrid %	-0.02^{*}	-0.02^{**}
	(-0.03, 0.00)	(-0.03, -0.00)
Hispanic % × Virtual %	-0.01	-0.00
	(-0.03, 0.00)	(-0.01, 0.00)
Hispanic % × Hybrid %	0.00	0.00
	(-0.02, 0.02)	(-0.02, 0.01)
ED % × Hybrid %	0.01	0.01^{***}
	(-0.00, 0.03)	(0.00, 0.02)
Model Summary		
Observations	5592	5128
Degree of Freedom	1417	1301
R-squared	0.947	0.933
F-statistics	52.74	67.09

	Sum of Coefficients	
Influential Points Analysis	Before Removal	After Removal
Virtual % + CO × Virtual %	0.12	0.14
	(-0.61, 0.85)	(-0.37, 0.65)
Virtual % + GA × Virtual %	1.37^{***}	0.93^{***}
	(0.77, 1.98)	(0.41, 1.45)
Virtual % + WI × Virtual %	0.40	0.20
	(-0.24, 1.03)	(-0.22, 0.61)
Hybrid % + CO × Hybrid %	-0.27	0.02
	(-0.83, 0.30)	(-0.33, 0.36)
Hybrid % + GA × Hybrid %	0.37	0.29
	(-0.29, 1.03)	(-0.19, 0.76)
Hybrid % + WI × Hybrid %	-0.58^{*}	-0.48^{***}
	(-1.16, 0.01)	(-0.83, -0.14)

*Note: This table illustrates the relationship between instruction mode and Dropout rates before (left) and after (right) removing influential points for Arizona, Colorado, Georgia, and Wisconsin, respectively. The results of model (1) are presented in the first sub-table. The second sub-table presents the sum of coefficients for state and instruction mode interactions to evaluate the effect of instruction mode on individual state samples. All regressions are weighted by the number of enrolled students and include both year and school fixed effects. 95% confidence intervals, calculated with robust standard errors, are reported in parentheses. Coefficients are denoted with ***, **, or * for P-values less than 0.01, 0.05, and 0.1, respectively.*

Fig. 8: Dropout Rate Changes by State and by Instruction Mode



(a) Arizona

(b) Georgia

(c) Wisconsin

Note: This figure shows the average dropout rates from the 2016-17 to the 2020-21 school year for Arizona, Georgia, and Wisconsin, respectively. The data is weighted by the number of enrolled students. Comparisons are presented by the share of virtual and hybrid learning during the 2020-21 school year. The "Low" and "High" categories correspond to schools with below and above weighted median by state, respectively

VII. Discussion and Conclusion

Several limitations must be addressed in these analyses. One primary concern is the potential endogeneity in instruction mode, which could compromise the validity of the results. Schools predominantly adopting virtual learning may have been influenced by various other pandemic-related factors. For instance, a higher prevalence of community COVID-19 infections might lead to less in-person instruction. Meanwhile, if COVID-19 adversely affects test proficiency rates, the parameters of instruction mode could inadvertently absorb these impacts, resulting in potentially overestimated negative effects. Similarly, social lock-downs, such as the closure of local businesses or after-school activities, could impact student learning. Disentangling instruction mode changes from these broader pandemic-related changes is challenging due to their inherent collinearity. Moreover, the effectiveness of virtual or hybrid learning on student achievements is contingent upon school-level access to technological resources. While demographic information offers insights, access to technology may be influenced by unobserved characteristics not captured by either time or entity fixed effects. Another limitation lies in the bias resulting from varying test participation rates within schools across years. Non-participants may systematically differ from participants, and although schools in the sampled states generally exhibit high and consistent participation rates, a 100% rate across all years is rare, potentially introducing biases. The level of comparability among test participants before and during the pandemic may also influence descriptive results. Additional biases could arise from data reporting practices, which, for privacy reasons, may suppress schools with extremely low performances or participation rates; in aggregate, this could impact the representativeness of the analyses. It is also imperative to point out that test proficiency and dropout rates serve as only two measures of student learning during the 2020-21 school year. Whether these metrics adequately capture student achievements is debatable. For instance, declines in dropout rates may not necessarily indicate increased student engagement, especially in the context of distance education, but could be explained by the pandemic's adverse effects on the labor market.

Despite the limitations, this study contributes considerably to the existing education research in several ways. First, unlike previous findings on grade 3-8 students, it reveals that high school students did not experience considerable learning loss during the pandemic and no compelling evidence of widening achievement gaps were present among marginalized groups. Second, conclusions of prior studies on younger students were drawn from aggregated data without considering potential disparities across states. Utilizing attribute interactions, this study identifies significant disparate effects of instruction mode on education outcomes, with some states demonstrating notable benefits from virtual and hybrid learning. Therefore, the widespread perception that distance education exacerbated declines in academic performances during the pandemic may not be generalized to all states. Third, influential analysis is not commonly utilized in education research involving panel data. However, this study highlights its value in establishing reliable and generalizable results, ensuring that the perceived regression results are not driven primarily by several influential data points. The influential observations examination and removal process in this study also provides a statistical evaluation framework that guides research endeavors investigating short-term treatment effects.

Moving forward, future research should aim to understand the underlying reasons for differences across states. In addition to gathering more school-level data on COVID-19 infection rates, access to technology, and local COVID-19 policies, researchers could delve deeper into the socioeconomic and demographic factors influencing the effectiveness of instruction modes. For instance, compared to Illinois and Wisconsin, Indiana adopted a more balanced approach combining in-person and remote learning during the pandemic and emphasized maintaining student engagement through innovative teaching method during hybrid learning. Approaches as such could contribute to higher education performance under disruptions. Exploring variables such as household income, parental education level, and urban-rural divide can also provide insights into the observed disparities. Moreover, comparative analyses across different education systems can guide best practices for implementing distance learning modalities. By examining variations in policy responses, technological infrastructures, and pedagogical approaches across states, researchers can identify transferable lessons and strategies for mitigating the adverse effects of disruptions like the COVID-19 pandemic on education systems nationwide. Additionally, qualitative research methods, such as surveys and interviews with students and teachers, can provide rich insights into the experiential aspects of distance education. Understanding stakeholders' perceptions, experiences, and challenges during remote instruction can inform the design of more effective and equitable educational strategies.

In conclusion, future research endeavors should adopt a multidimensional approach, integrating quantitative and qualitative methods and state-to-state comparisons to comprehensively understand the complex interplay between instruction modes and education outcomes. By addressing these gaps, researchers can contribute to the development of evidence-based policies and practices that promote resilience, equity, and innovation in the education systems.

References

Barbour, M. and Reeves, T. (2009). The reality of virtual schools: A review of the literature. *Computers & Education*.

Belfield, C. (2021). Learning loss due to school closures during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*.

Bridgeland, J., Dilulio, J., and Morison, K. (2006). The silent epidemic: Perspectives of high school dropouts.

Clark, A. E., Nong, H., Zhu, H., and Zhu, R. (2021). Compensating for academic loss: Online learning and student performance during the covid-19 pandemic. *Economics of Education Review*.

Dorn, E., Hancock, B., Sarakatsannis, J., and Viruleg, E. (2021). Covid-19 and education: The lingering effects of unfinished learning.

Harris, D. and Chen, F. (2022). How has the pandemic affected high school graduation and college entry? *Brookings Institution*.

Jack, R., Halloran, C., Okun, J., and Oster, E. (2023). Pandemic schooling mode and student test scores: Evidence from us school districts. *American Economic Review: Insights*, 5(2):173–190.

Maldonado, J. E. and De Witte, K. (2020). The effect of school closures on standardised student test outcomes. Technical report, KU Leuven.

Patterson, B. and McFadden, C. (2009). Attrition in online and campus degree programs. *Online Journal of Distance Learning Administration*.

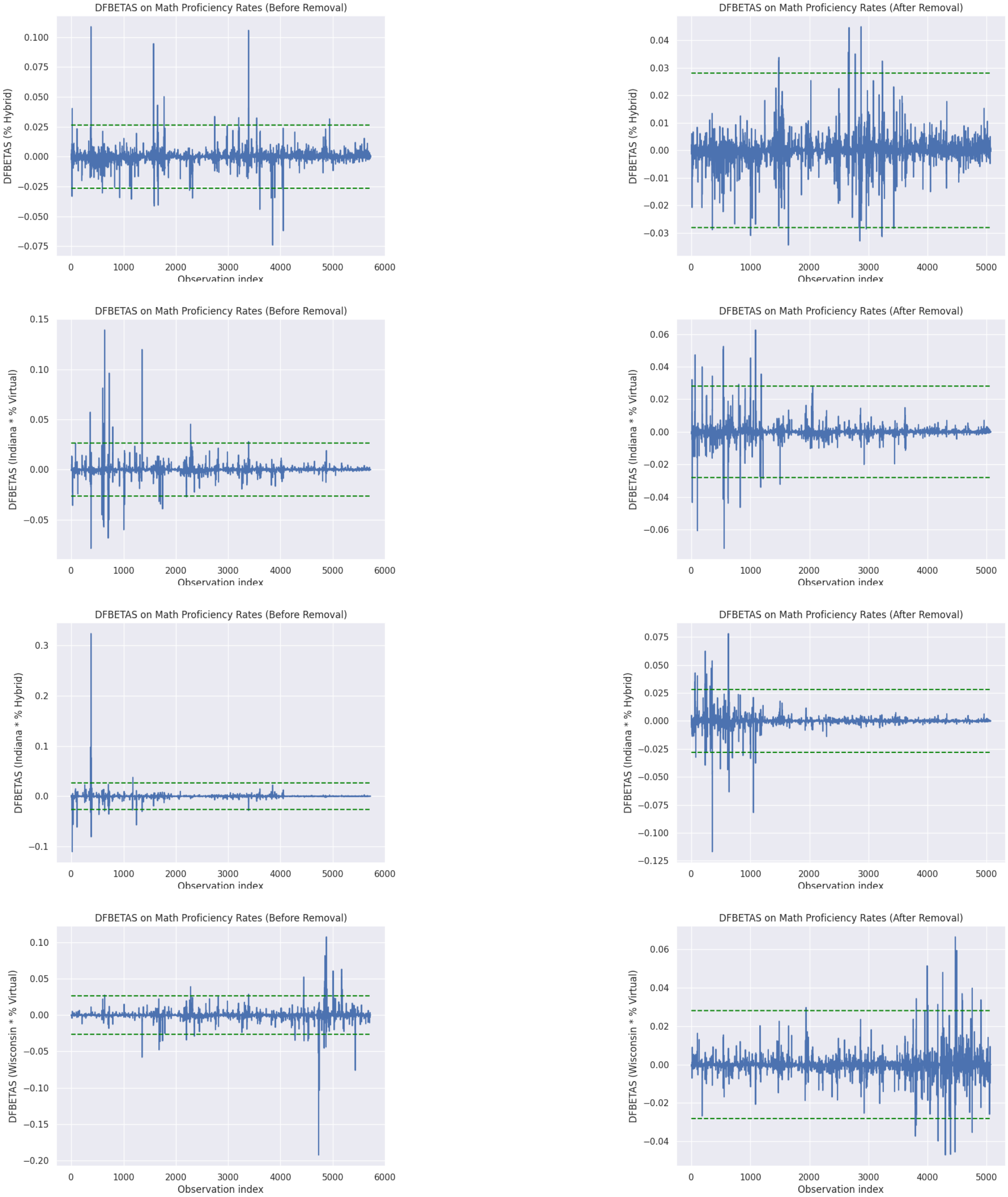
Photopoulos, P., Tsonos, C., Stavrakas, I., and Triantis, D. (2022). Remote and in-person learning: Utility versus social experience. *International Journal of Environmental Research and Public Health*.

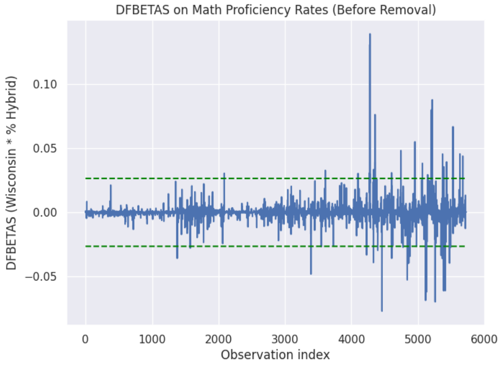
Tomasik, M. J., Helbling, L. A., and Moser, U. (2014). The economic burden of high school dropouts and school suspensions in florida. Technical report, University of California eScholarship.

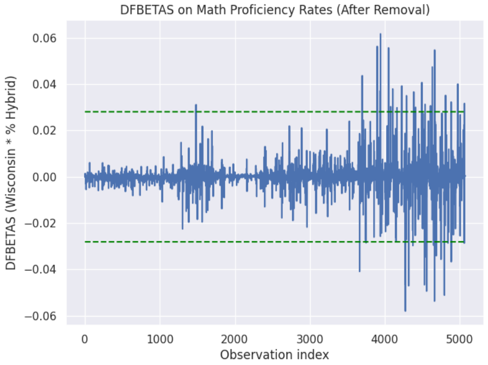
Zhang, Y., Zhao, G., and Zhou, B. (2021). Does learning longer improve student achievement? evidence from online education of graduating students in a high school during covid-19 period. *Economics of Education Review*.

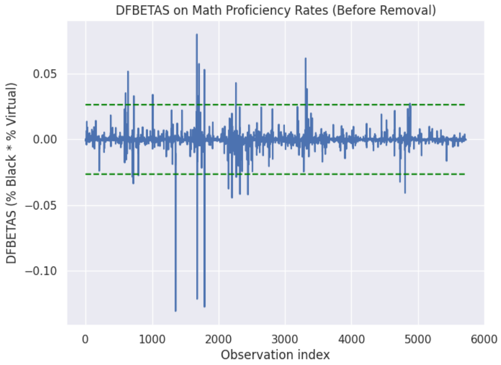
Appendix A: Figures

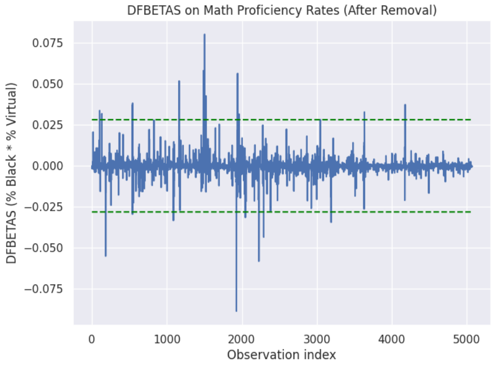
Fig. A.1: DFBETAS Plots Comparison for Math Proficiency Rates

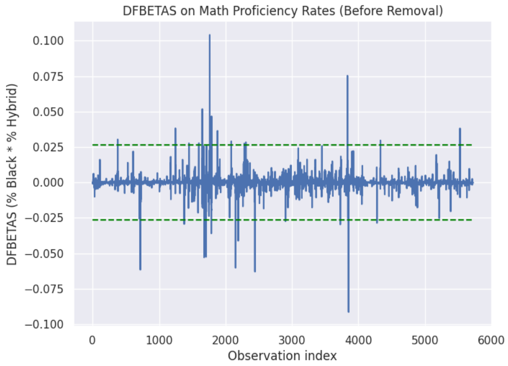


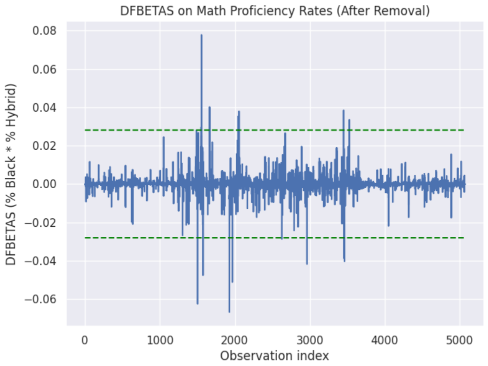

DFBETAS on Math Proficiency Rates (Before Removal)
DFBETAS (Wisconsin * % Hybrid)
0.10
0.05
0.00
−0.05
0
1000
2000
3000
4000
5000
6000
Observation index

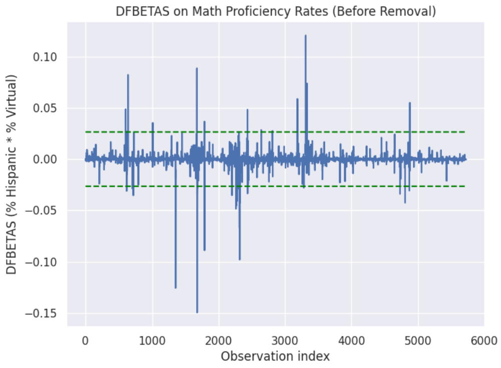

DFBETAS on Math Proficiency Rates (After Removal)
DFBETAS (Wisconsin * % Hybrid)
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0
1000
2000
3000
4000
5000
Observation index

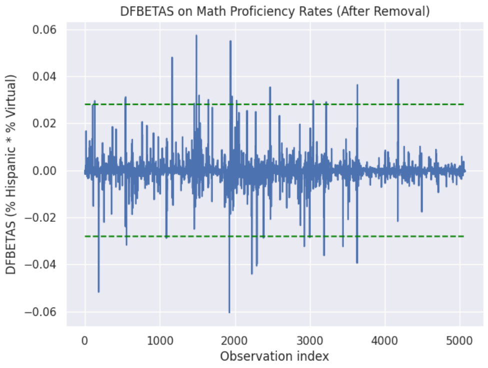

DFBETAS on Math Proficiency Rates (Before Removal)
DFBETAS (% Black * % Virtual)
0.05
0.00
−0.05
−0.10
0
1000
2000
3000
4000
5000
6000
Observation index

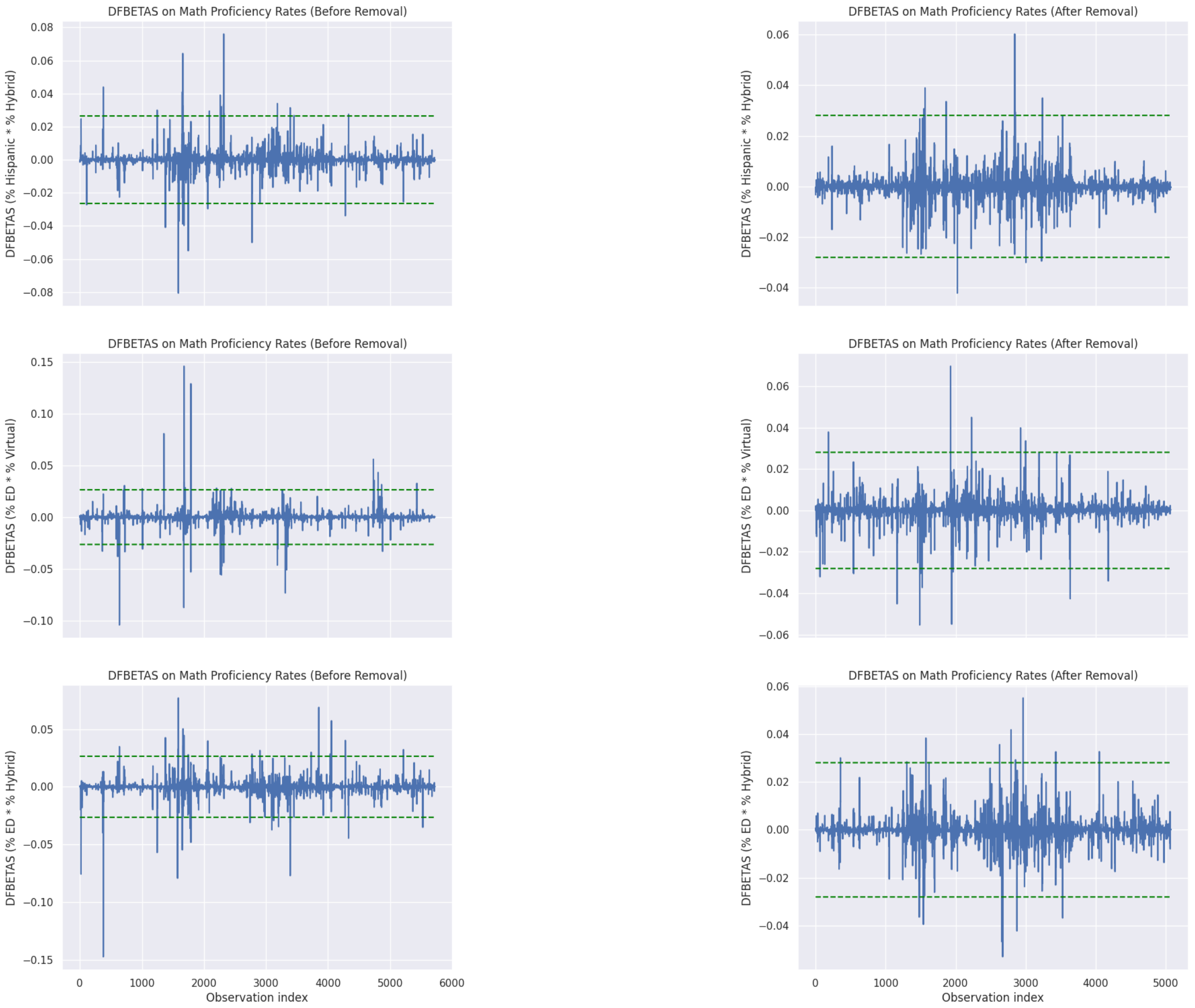

DFBETAS on Math Proficiency Rates (After Removal)
DFBETAS (% Black * % Virtual)
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
0
1000
2000
3000
4000
5000
Observation index


DFBETAS on Math Proficiency Rates (Before Removal)
DFBETAS (% Black * % Hybrid)
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
0
1000
2000
3000
4000
5000
6000
Observation index


DFBETAS on Math Proficiency Rates (After Removal)
DFBETAS (% Black * % Hybrid)
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0
1000
2000
3000
4000
5000
Observation index


DFBETAS on Math Proficiency Rates (Before Removal)
DFBETAS (% Hispanic * % Virtual)
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
1000
2000
3000
4000
5000
6000
Observation index


DFBETAS on Math Proficiency Rates (After Removal)
DFBETAS (% Hispanic * % Virtual)
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0
1000
2000
3000
4000
5000
Observation index

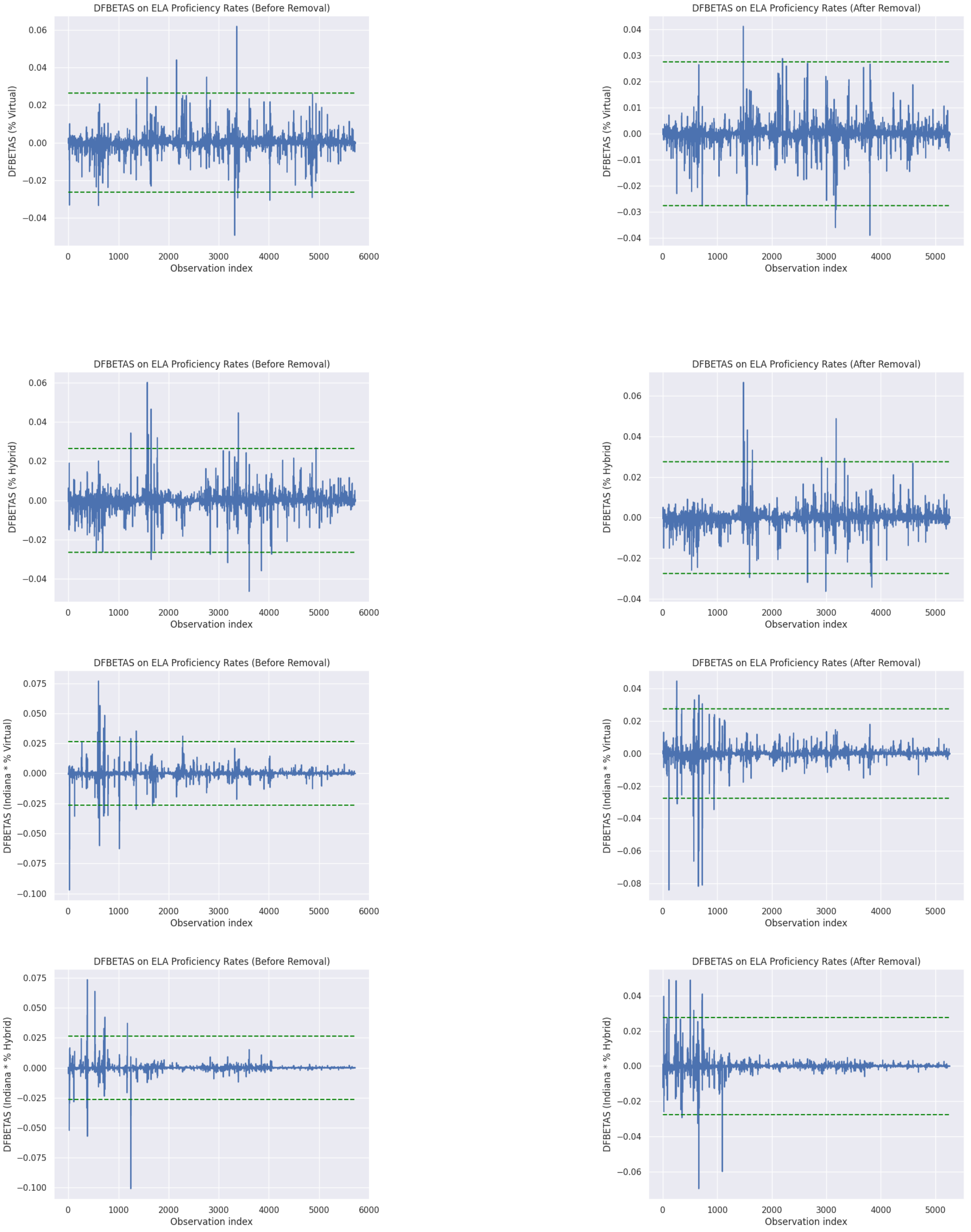


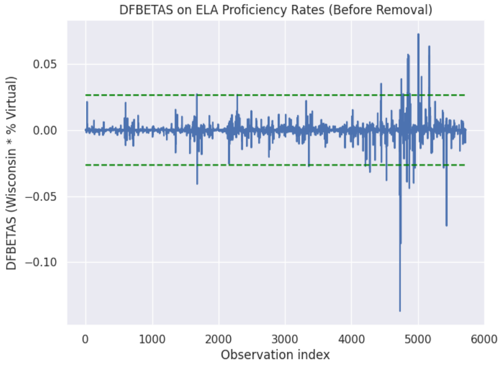
(a) Before Removal

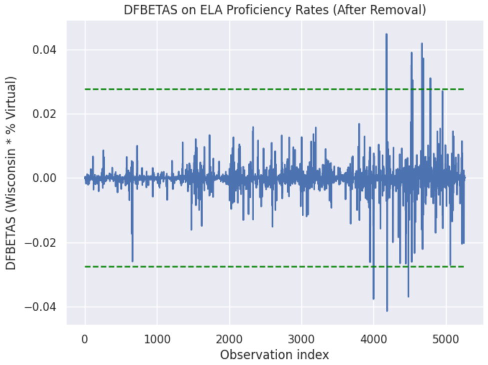
(b) After Removal

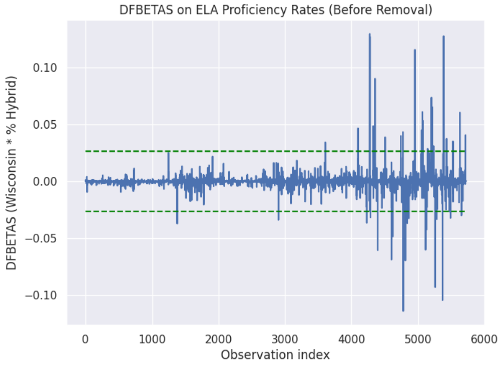
Note: This figure presents a comparison of DFBETAS values before and after the removal of influential points. The blue bars indicate the DFBETAS value for all included observations concerning the parameters of interests on math proficiency rates. The green dashed line depicts the critical value used to identify influential observations, calculated as $2/\sqrt{n}$.

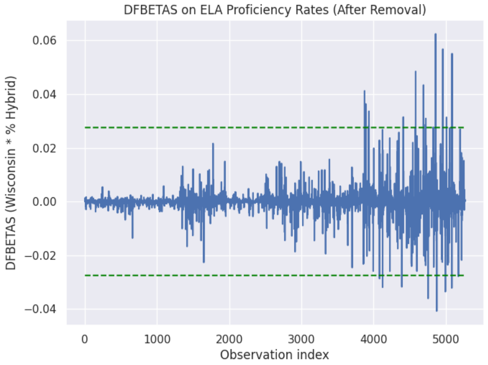
Fig. A.2: DFBETAS Plots Comparison for ELA Proficiency Rates

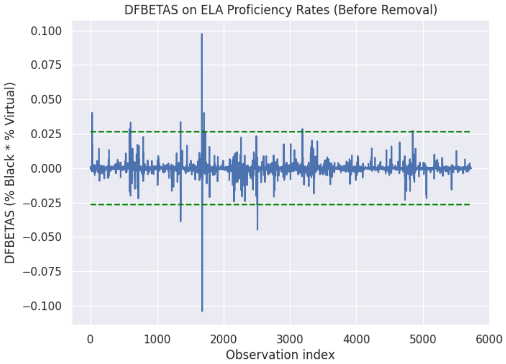


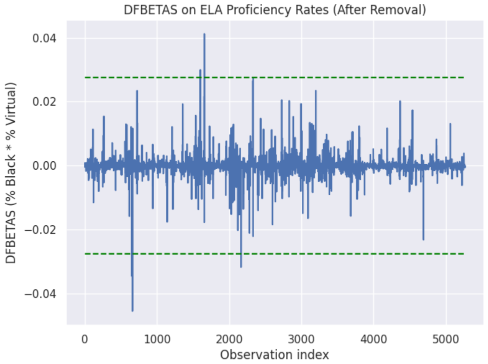

DFBETAS on ELA Proficiency Rates (Before Removal)

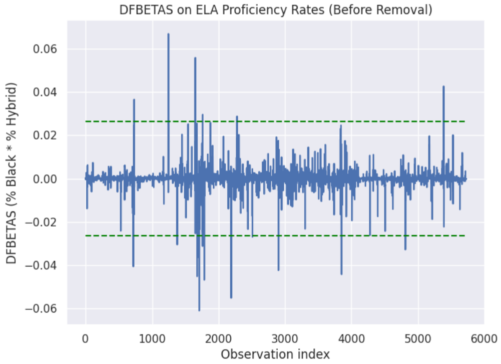

DFBETAS on ELA Proficiency Rates (After Removal)

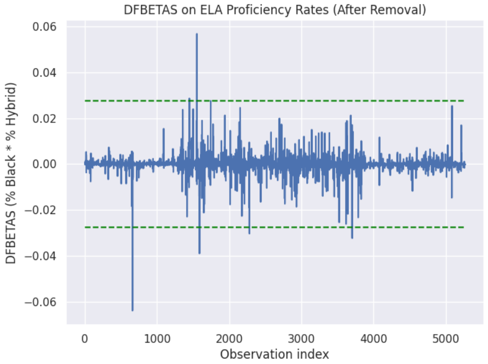

DFBETAS on ELA Proficiency Rates (Before Removal)

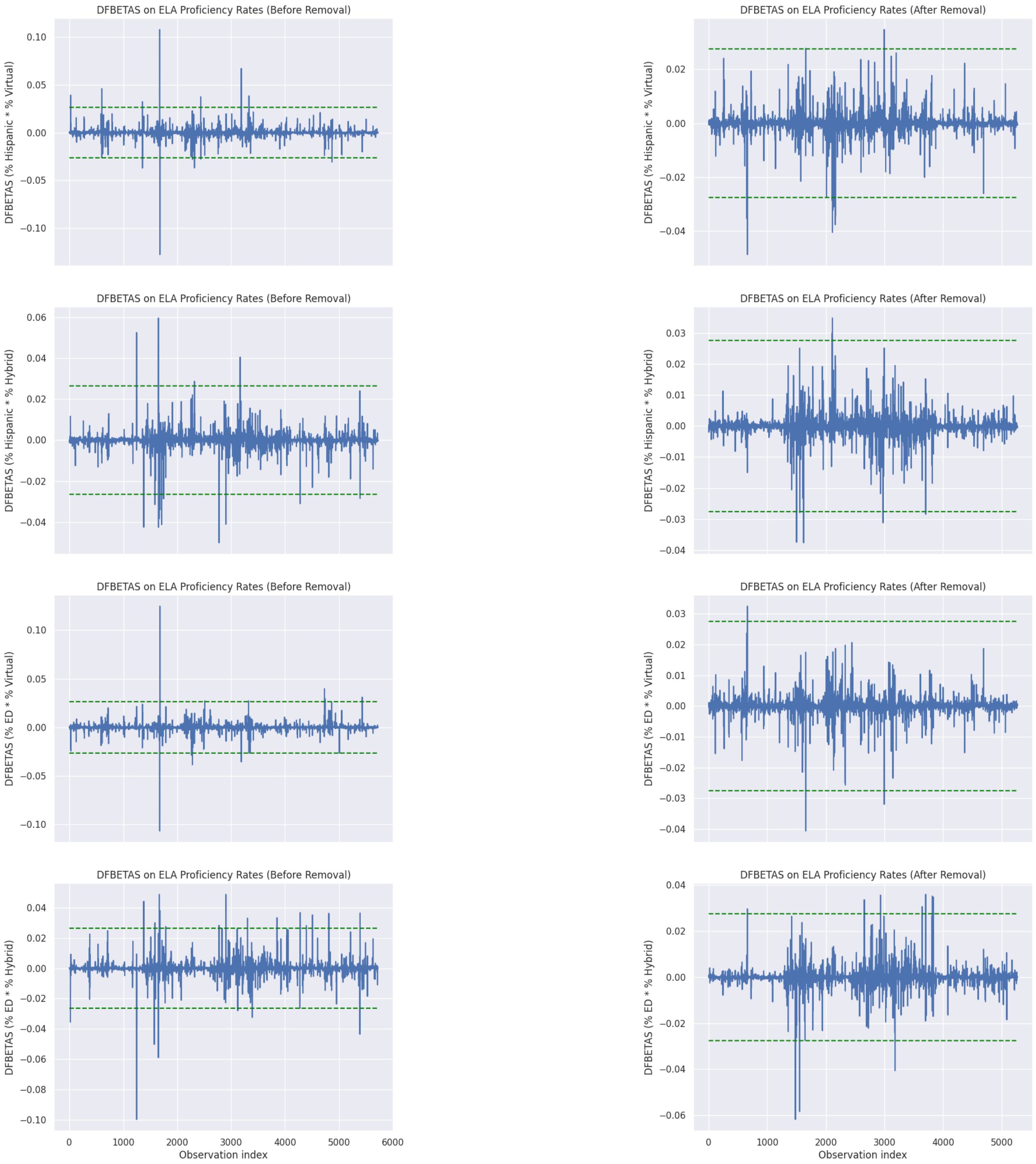

DFBETAS on ELA Proficiency Rates (After Removal)


DFBETAS on ELA Proficiency Rates (Before Removal)


DFBETAS on ELA Proficiency Rates (After Removal)


DFBETAS on ELA Proficiency Rates (Before Removal)


DFBETAS on ELA Proficiency Rates (After Removal)

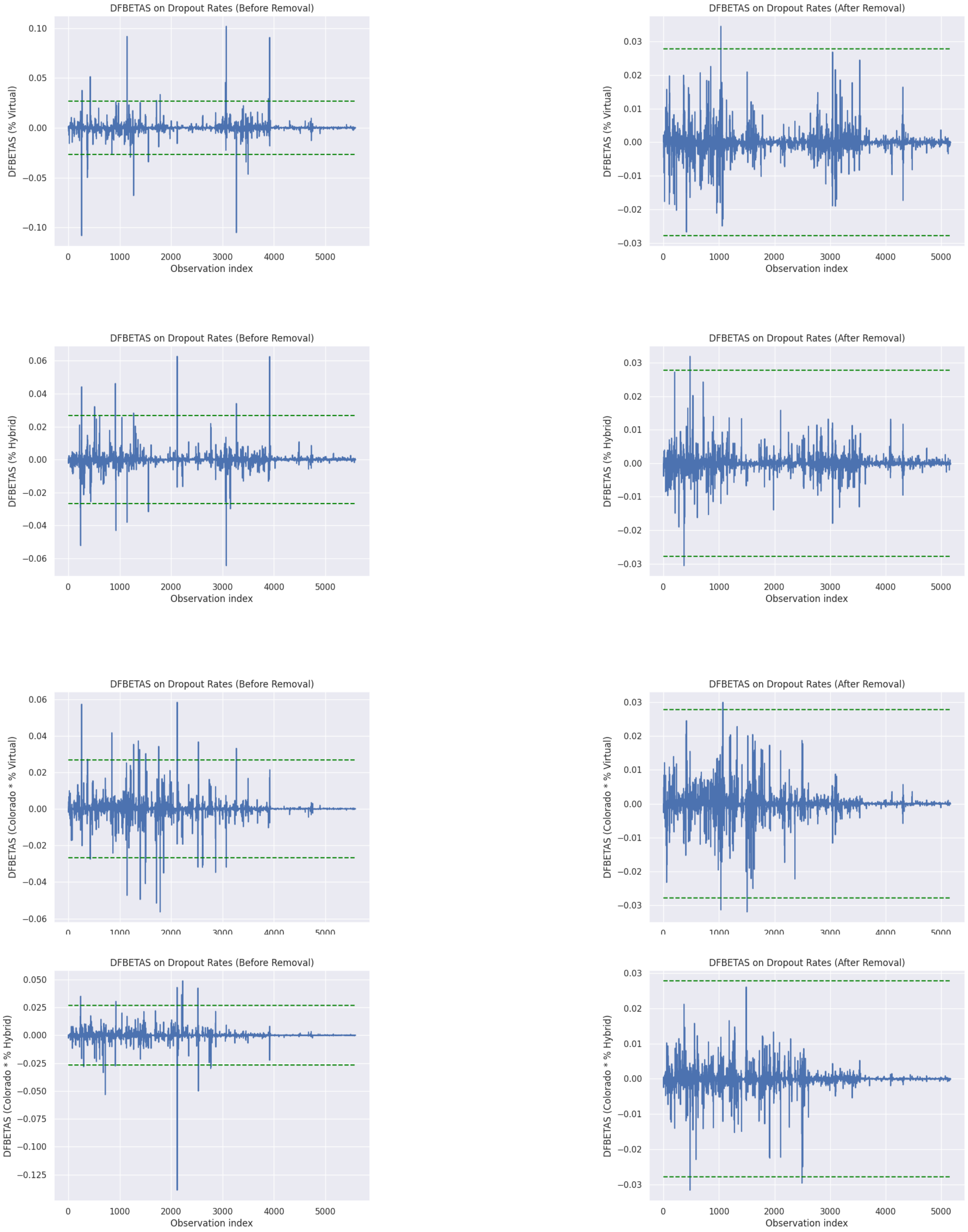


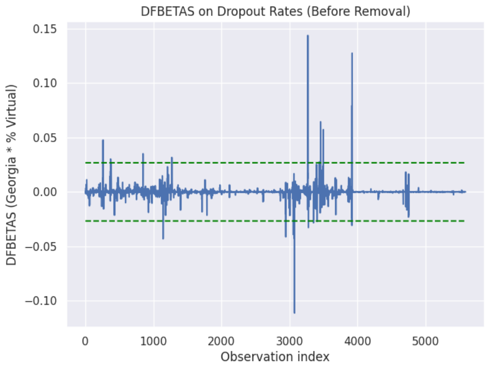
(a) Before Removal

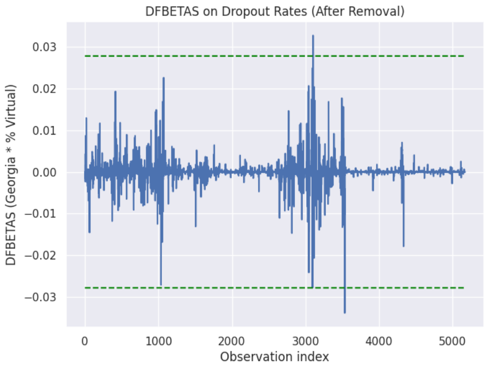
(b) After Removal

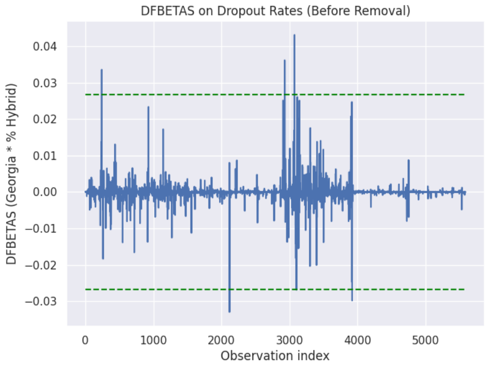
Note: This figure presents a comparison of DFBETAS values before and after the removal of influential points. The blue bars indicate the DFBETAS value for all included observations concerning the parameters of interests on ELA proficiency rates. The green dashed line depicts the critical value used to identify influential observations, calculated as $2/\sqrt{n}$.

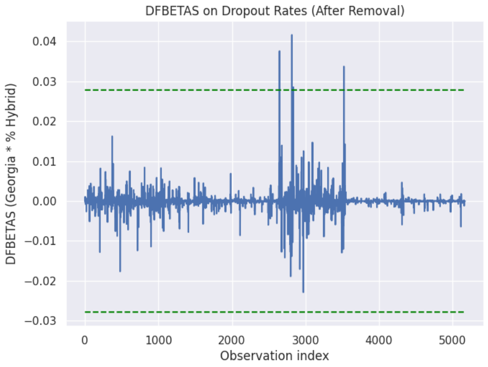
Fig. A.3: DFBETAS Plots Comparison for Dropout Rates

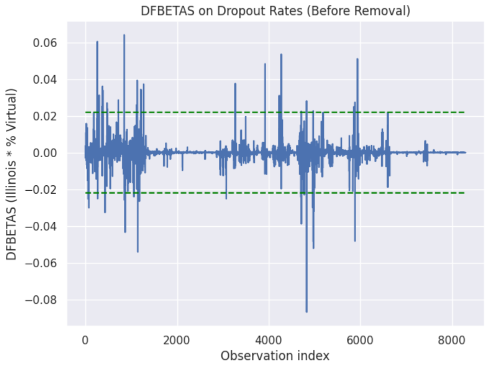


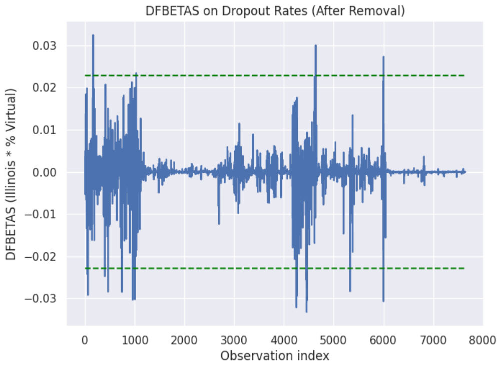

DFBETAS on Dropout Rates (Before Removal)
DFBETAS (Georgia * % Virtual)
0.15
0.10
0.05
0.00
−0.05
−0.10
0
1000
2000
3000
4000
5000
Observation index

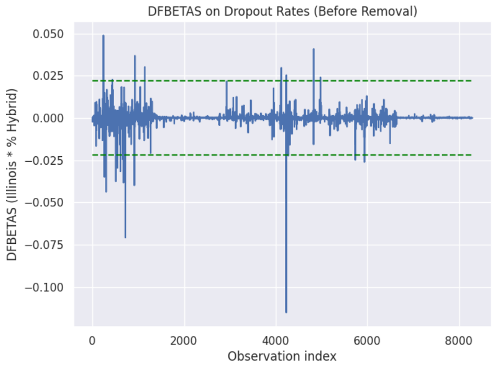

DFBETAS on Dropout Rates (After Removal)
DFBETAS (Georgia * % Virtual)
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
Observation index

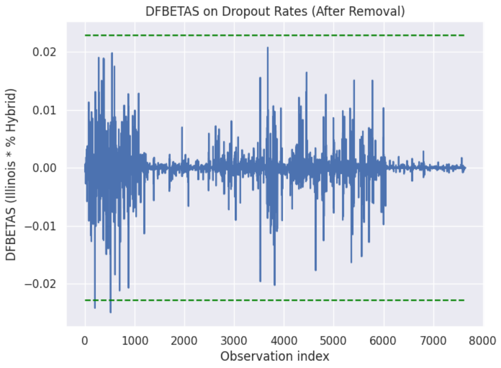

DFBETAS on Dropout Rates (Before Removal)
DFBETAS (Georgia * % Hybrid)
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
Observation index

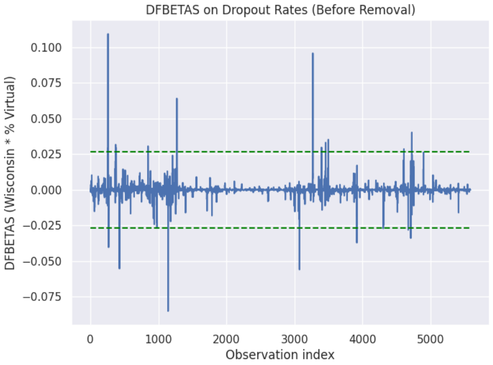

DFBETAS on Dropout Rates (After Removal)
DFBETAS (Georgia * % Hybrid)
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
Observation index

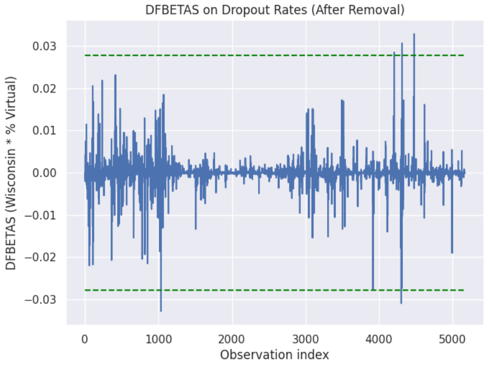

DFBETAS on Dropout Rates (Before Removal)
DFBETAS (Illinois * % Virtual)
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
0
2000
4000
6000
8000
Observation index

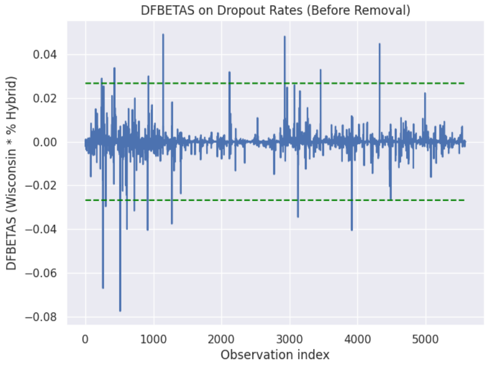

DFBETAS on Dropout Rates (After Removal)
DFBETAS (Illinois * % Virtual)
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
6000
7000
8000
Observation index

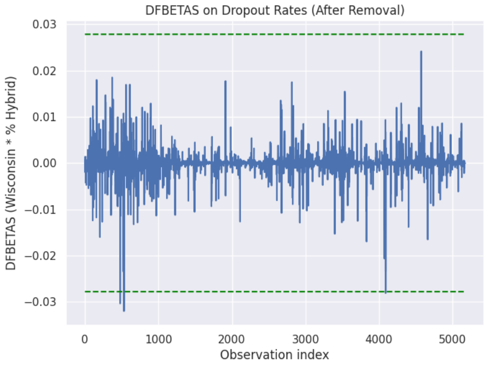

DFBETAS on Dropout Rates (Before Removal)
DFBETAS (Illinois * % Hybrid)
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
0
2000
4000
6000
8000
Observation index

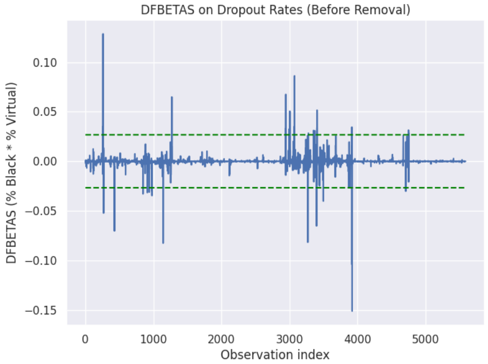

DFBETAS on Dropout Rates (After Removal)
DFBETAS (Illinois * % Hybrid)
0.02
0.01
0.00
−0.01
−0.02
0
1000
2000
3000
4000
5000
6000
7000
8000
Observation index

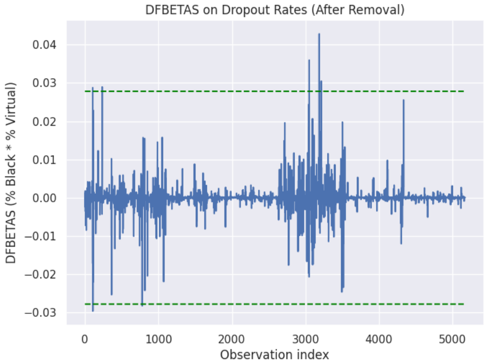

DFBETAS on Dropout Rates (Before Removal)
DFBETAS (Wisconsin * % Virtual)
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
0
1000
2000
3000
4000
5000
Observation index

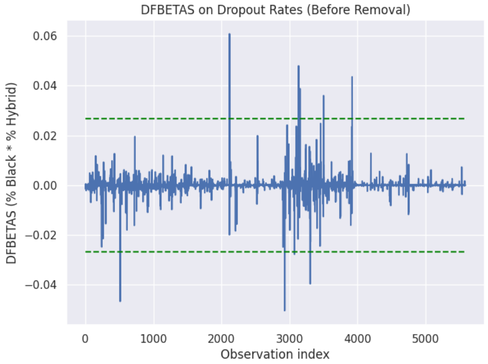

DFBETAS on Dropout Rates (After Removal)
DFBETAS (Wisconsin * % Virtual)
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
Observation index

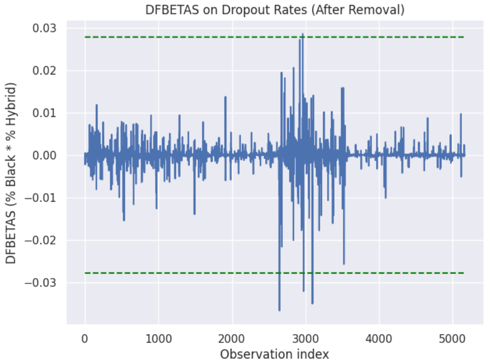

DFBETAS on Dropout Rates (Before Removal)
DFBETAS (Wisconsin * % Hybrid)
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
0
1000
2000
3000
4000
5000
Observation index

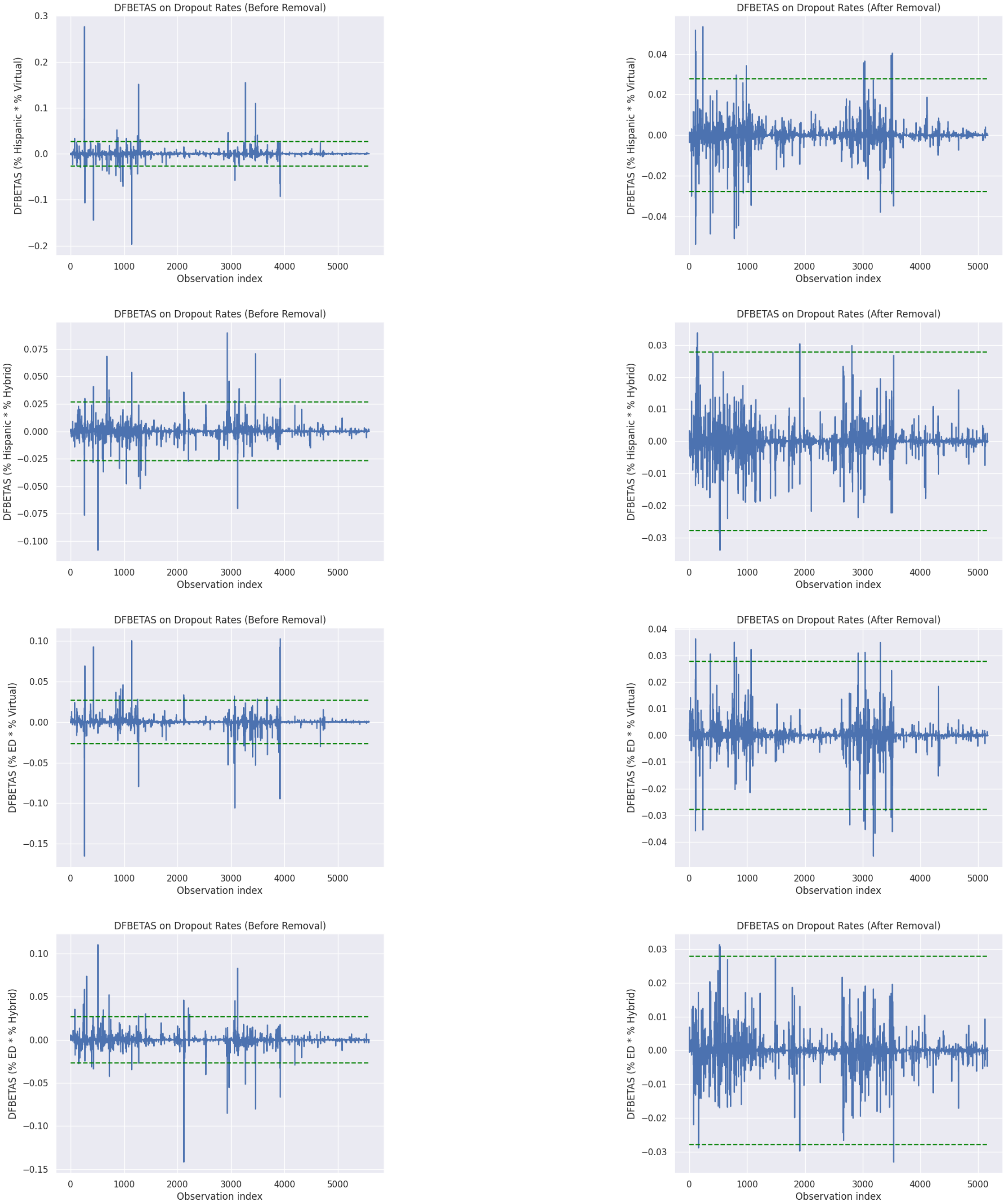

DFBETAS on Dropout Rates (After Removal)
DFBETAS (Wisconsin * % Hybrid)
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
Observation index


DFBETAS on Dropout Rates (Before Removal)
DFBETAS (% Black * % Virtual)
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
1000
2000
3000
4000
5000
Observation index


DFBETAS on Dropout Rates (After Removal)
DFBETAS (% Black * % Virtual)
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
Observation index


DFBETAS on Dropout Rates (Before Removal)
DFBETAS (% Black * % Hybrid)
0.06
0.04
0.02
0.00
−0.02
−0.04
0
1000
2000
3000
4000
5000
Observation index


DFBETAS on Dropout Rates (After Removal)
DFBETAS (% Black * % Hybrid)
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
1000
2000
3000
4000
5000
Observation index

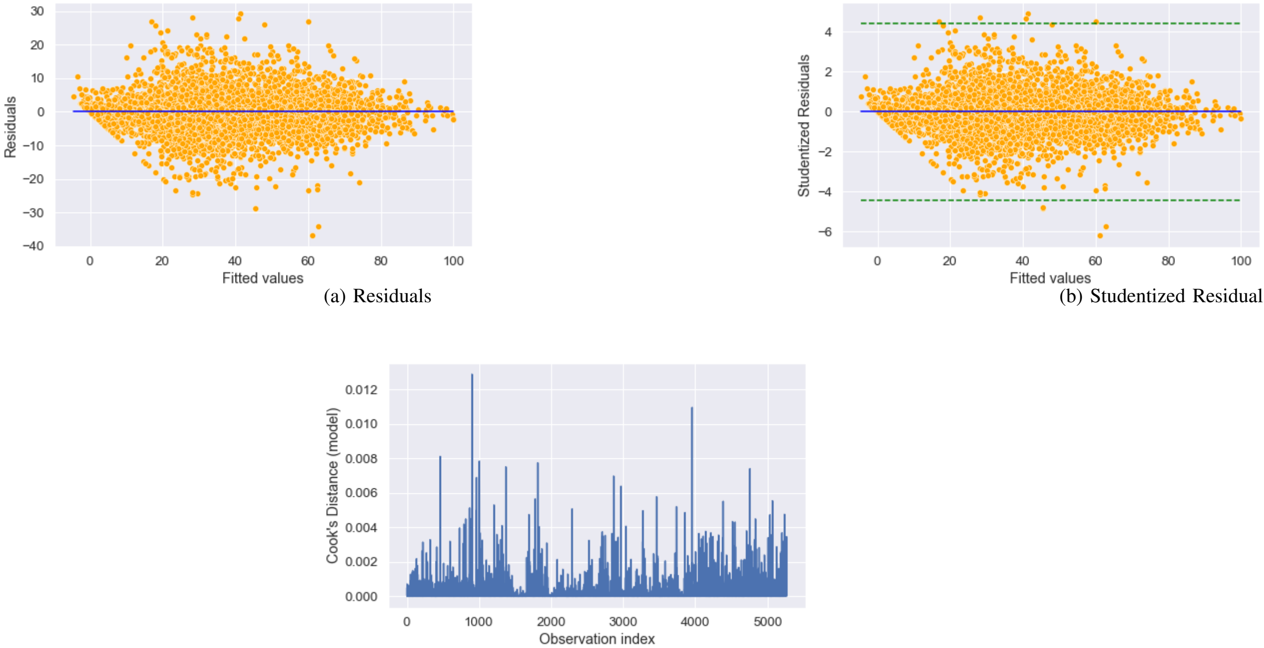


(a) Before Removal

(b) After Removal

Note: This figure presents a comparison of DFBETAS values before and after the removal of influential points. The blue bars indicate the DFBETAS value for all included observations concerning the parameters of interests on dropout rates. The green dashed line depicts the critical value used to identify influential observations, calculated as $2/\sqrt{n}$.

Fig. A.4: Diagnostic Visualizations for ELA Proficiency Rates



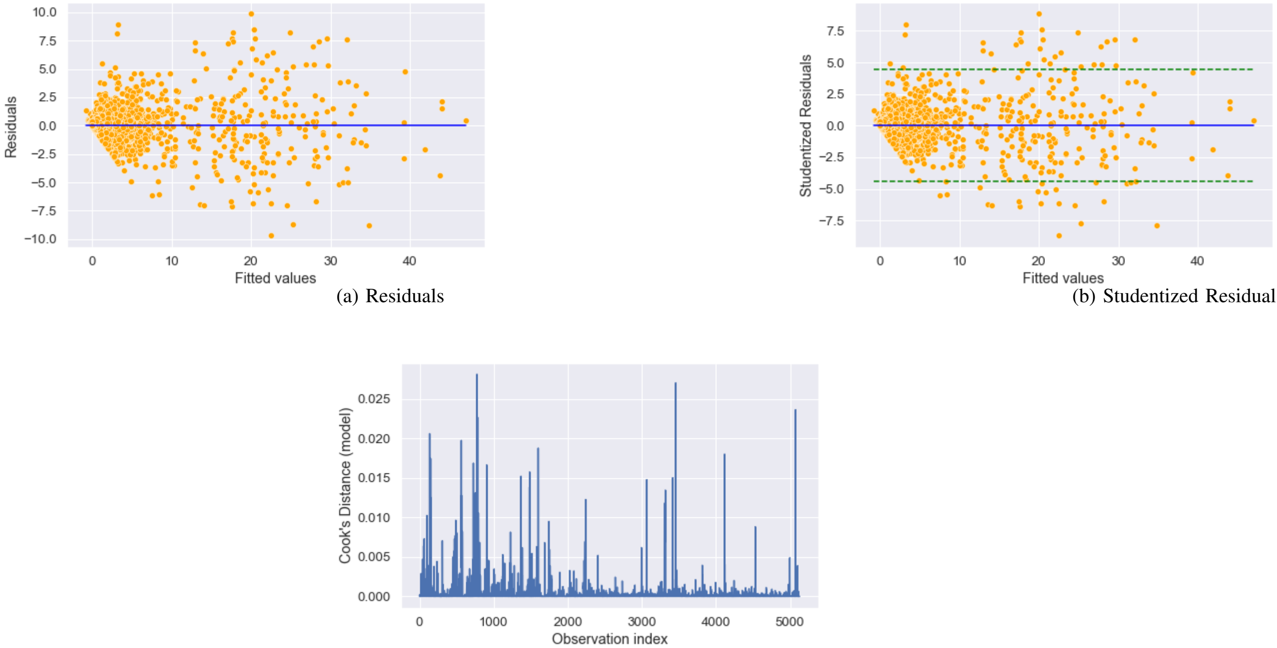
(a) Residuals

(b) Studentized Residuals

(c) Cook's Distance

Note: This figure displays residuals against fitted values, studentized residuals against fitted values, and Cook's distance for each observation in regression analyses for ELA proficiency rates. All plots are generated using the final dataset after removing influential points. In the studentized residual plot, the green dashed line represents critical values for outliers, calculated at the 95% confidence interval. The blue bars in the Cook's distance plot denote Cook's distance of each observation included in the final model. The cutoff for highly influential points was calculated as 0.99, which is significantly larger than the Cook's distance of any observation.

Fig. A.5: Diagnostic Visualizations for Dropout Rates



(a) Residuals

(b) Studentized Residuals

(c) Cook’s Distance

Note: This figure displays residuals against fitted values, studentized residuals against fitted values, and Cook’s distance for each observation in regression analyses for dropout rates. All plots are generated using the final dataset after removing influential points. In the studentized residual plot, the green dashed line represents critical values for outliers, calculated at the 95% confidence interval. The blue bars in the Cook’s distance plot denote Cook’s distance of each observation included in the final model. The cutoff for highly influential points was calculated as 0.99, which is significantly larger than the Cook’s distance of any observation.

Appendix B: Tables

TABLE B.1: Summary Statistics by Educational Outcomes and by State

	Instruction Mode and Demographic Characteristics							
	# School	Virtual	Hybrid	Black or Hispanic	ED	Math	ELA	Drop
Panel A: Math								
Indiana (IN)	306	13.84	7.16	17.71	41.40	35.20		
Illinois (IL)	608	33.74	47.69	39.94	45.89	31.14		
Wisconsin (WI)	354	13.80	22.30	13.01	31.48	33.87		
Overall	1268	24.38	31.48	28.66	41.93	32.79		
Panel B: ELA								
Indiana (IN)	317	14.76	7.66	17.53	39.97		62.12	
Illinois (IL)	656	34.32	50.10	40.24	43.66		35.64	
Wisconsin (WI)	355	15.89	22.27	14.72	32.74		38.36	
Overall	1328	25.95	33.97	29.89	40.84		43.18	
Panel C: Dropout								
Arizona (AZ)	282	41.54	56.68	53.98	50.58			2.83
Colorado (CO)	366	29.72	42.80	36.89	36.37			1.68
Georgia (GA)	228	33.36	21.57	53.44	51.96			2.73
Wisconsin (WI)	406	18.48	29.21	16.48	33.03			1.10
Overall	1282	31.61	36.82	42.47	44.21			2.19

Note: This table shows the summary statistics of % Virtual, % Hybrid, % Black or Hispanic, % Economically Disadvantaged (ED), Math and ELA proficiency rates, weighted by the number of students tested, and Dropout rates, weighted by enrollment. Demographics characteristics and academic outcomes variables are weighted mean from the 2016-17 to the 2020-21 school year. Instruction mode variables pertain to the 2020-21 school year only, since % Virtual and % Hybrid are expected to be 0 during pre-pandemic years.

TABLE B.2: Correlations between Instruction Mode and Student Demographics

	Pairwise Correlation with Virtual Learning		
	Math	ELA	Dropout
Black %	0.61	0.60	0.44
	(0.56, 0.67)	(0.54, 0.65)	(0.38, 0.49)
Hispanic %	0.51	0.54	0.39
	(0.45, 0.56)	(0.48, 0.59)	(0.34, 0.45)
ED %	0.52	0.51	0.41
	(0.47, 0.58)	(0.46, 0.57)	(0.36, 0.47)

Note: This table shows the pairwise correlation between the share of virtual learning during the 2020-21 school year and demographic characteristics. 95% confidence intervals are reported in parenthesis. All correlation coefficients have p-values below 0.01.

TABLE B.3: Results of Model Assumption Check

	Model Assumption Check (Before Influential Points Analysis)		
	Math	ELA	Dropout
Panel A: VIF			
% Virtual	9.18	9.18	18.19
% Hybrid	5.27	5.27	9.74
IN * % Virtual	1.57	1.56	
IN * % Hybrid	1.21	1.21	
WI * % Virtual	1.79	1.79	1.97
WI * % Hybrid	1.58	1.58	2.36
CO * % Virtual			2.68
CO * % Hybrid			2.77
GA * % Virtual			9.13
GA * % Hybrid			4.36
% Black * % Virtual	6.06	6.05	9.12
% Black * % Hybrid	2.26	2.26	5.18
% Hispanic * % Virtual	9.84	9.84	7.01
% Hispanic * % Hybrid	3.51	3.51	11.58
% ED * % Virtual	18.33	18.32	11.82
% ED * % Hybrid	7.73	7.73	12.83
Panel B: Durbin-Watson			
Test Statistics	2.29	2.37	2.28

	Model Assumption Check		
	(After Influential Points Analysis)		
	Math	ELA	Dropout
Panel A: VIF			
% Virtual	9.48	9.09	21.62
% Hybrid	6.94	5.05	10.85
IN * % Virtual	1.53	1.63	
IN * % Hybrid	1.32	1.31	
WI * % Virtual	1.81	1.67	2.19
WI * % Hybrid	1.56	1.46	2.64
CO * % Virtual			2.94
CO * % Hybrid			2.82
GA * % Virtual			11.11
GA * % Hybrid			4.59
% Black * % Virtual	10.56	6.01	14.22
% Black * % Hybrid	2.84	2.33	5.42
% Hispanic * % Virtual	13.96	9.59	10.86
% Hispanic * % Hybrid	3.82	3.66	11.76
% ED * % Virtual	30.20	16.29	16.59
% ED * % Hybrid	10.80	8.19	13.05
Panel B: Durbin-Watson			
Test Statistics	2.31	2.39	2.34

	Model Assumption Check		
	(Final Model Specification)		
	Math	ELA	Dropout
Panel A: VIF			
% Virtual	8.97	8.84	19.83
% Hybrid	6.50	4.92	10.41
IN * % Virtual	1.52	1.59	
IN * % Hybrid	1.31	1.31	
WI * % Virtual	1.64	1.52	2.19
WI * % Hybrid	1.55	1.45	2.62
CO * % Virtual			2.89
CO * % Hybrid			2.81
GA * % Virtual			9.36
GA * % Hybrid			4.38
% Black * % Virtual	2.98	2.80	5.99
% Black * % Hybrid	2.72	2.31	5.01
% Hispanic * % Virtual	4.94	4.31	8.02
% Hispanic * % Hybrid	3.74	3.52	11.38
% ED * % Hybrid	9.32	7.46	11.15
Panel B: Durbin-Watson			
Test Statistics	2.31	2.39	2.34

*Note: This table shows the results of model assumption check from all obtained data (before influential points removals), final dataset used for regression analyses (after influential points removal), and final model specifications with final dataset (removed %ED * %Virtual due to high VIF and its high correlation with %Black * %Virtual and %Hispanic * %Virtual). Panel A shows the VIF values of all parameters of interest and Panel B shows the test statistics on the Durbin-Watson test for serial correlation detection. No severe multicollinearity or serial correlations are detected from the final model specification used for conclusions and recommendations.*

Assessing the Impacts of Residency Restrictions on Sex Crimes

Erin Lillis

Grinnell College

Abstract—**This paper analyzes the efficacy of residency restriction laws placed on sex offenders post-incarceration in Montana, North Dakota, and South Carolina. These laws give a specific radius within which those they deem "high-risk" sex offenders cannot live. A Callaway-Sant'Anna difference-in-difference model with state and year fixed effects is utilized as the identification strategy. The main finding is that the law's enactment led to an increase of 23.9 sex crimes per 100,000. It also led to a substantial increase in the proportion of sex crime victims under the age of 18. Results from a synthetic control model and an alternative difference-in-difference specification also found that the law did not substantially decrease sex crimes. The recommendation based on these results is that residency restriction laws are not effective policy initiatives to reduce sex crimes, especially against children.**

I. Introduction

Although there are numerous federal laws governing sex offender registration and community notification, there is no uniformity when it comes to residency restrictions for sex offenders (Office of Justice Programs, nd). As of 2018, 29 states have passed a sex offender residency restriction law (Savage and Windsor, 2018). These laws provide specific guidelines on how close sex offenders can live to places that are expected to have large populations of children. The exact locations restricted and the number of feet vary greatly depending on the state. In this paper, I evaluate the impact that state-level sex offender residency restriction laws implemented on or after 2008 have had on the rate of sex crimes. 95% of sex offenses are committed by first-time offenders and sex offenders have one of the lowest recidivism rates at only 13.4%, but it is unclear if this is a result of successful policy or if this group would have a low rate without strict restrictions (Perillo, 2016). I also examine how the law impacts minor versus adult victimization rates and the rate of sex crimes committed by strangers, since these laws particularly try to curb sex offenses by strangers against children.

Sex offender residency restriction laws have not been adopted at a federal level and continue to be debated at the state level. Therefore, extensive research is crucial for states to make informed policy decisions. Implementing these laws also disrupts the lives of sex offenders and requires significant resources to implement and enforce, so it is paramount that they are effective. There has also been a focus on pedophilia in recent years with the Epstein Island scandal, which involved many public figures and conservative politicians focused on "saving the children." Residency restriction laws are therefore especially relevant from this perspective, since they target reducing sex crimes with child victims.

I utilize a Callaway-Sant'Anna difference-in-difference model to capture a causal relationship between the change in policy and changes in sex crimes. The model uses state and year fixed effects. For states to be part of the untreated group, they must have never passed sex offender residency restriction laws, must not have local municipalities implementing their own laws, and must have National Incident-Based Reporting System (NIBRS) reporting data available for the years of interest. Five states match these criteria: Colorado, Connecticut, Kansas, Utah, and Vermont. For states to be part of the treatment group, they must have passed sex offender residency restriction laws in 2008 or after and also have NIBRS reporting data in the years surrounding the policy change. The states that match these criteria are Montana, North Dakota, and South Carolina. Each of these states has substantial individual differences in whom the law applies to and what the law stipulates, which allows for interesting analysis.

The main analysis finds that sex offender residency restriction laws have the opposite effect than intended. There is an increase of 23.9 sex offenses per 100,000 residents from the implementation of the law, and 20.64 of these offenses have minors as their victims. This result is statistically significant and represents a substantial jump in sex offenses from 132.34 per 100,000 residents pre-treatment. The majority of the increase comes from a jump in fondling offenses. All other categories of sex offenses see an increase after the treatment, which shows that the law is not effective in preventing any type of sex crime. This paper is the first of the known existing literature to look at how this law affects everyone, not just prior offenders. It is also the first to examine the impact on those it was intended to protect: minors.

The rest of the paper is outlined as follows: Section II provides a brief overview of the history of residence restriction laws as well as other research done on it, Section III explains the process of obtaining and cleaning the data used, Section IV outlines the difference-in-difference models used for analysis, Section V describes the results of the difference-in-difference models with a variety of specification, and finally Section VI gives explicit recommendations on the future of residency restriction laws. All of the tables and figures referenced can be found in the Appendix in Section VII.

II. Background and Literature Review

A. Historical Background

Starting in 1994, the laws in the United States governing sex offenders after their release have undergone drastic changes. The majority of this legislation was enacted in response to high-profile criminal cases involving sex offenders and children. The most well-known and impactful are the Jacob Wetterling Crimes Against Children and Sexually Violent Offender Registration Act, Megan's Law, and Adam Walsh Child Protection and Safety Act. These federal laws set specific, mandatory reporting standards for each state's sex offender registry, as well as public notification requirements (Office of Justice Programs, nd).

Residency restriction laws prohibit sex offenders from establishing residency within a specified distance from where children congregate (Bratina, 2013). The first was passed by Alabama in 1995, and a growing number of states have implemented them. As of 2008, 30 states had passed laws, but this number dropped to 29 in 2018 (Savage and Windsor, 2018). This is due to the constitutionality of these laws being debated and even repealed in some states including Massachusetts, California, New York, Rhode Island, and New Jersey (ACLU Massachusetts, 2015; Savage and Windsor, 2018; Nobles et al., 2012). In these cases, state supreme courts banned towns from implementing these laws. Their reasoning was that the laws were ineffective, expensive, and burdensome. Judges decided that federal sex offender registries and community notification were sufficient (ACLU Massachusetts, 2015; FOX News Network, 2015).

The exact specifications of sex offender residency restriction laws vary greatly from state to state. The main elements are what areas the laws apply to, the specified distance, and what type of sex offenders it applies to. The prohibited zones generally fall into the categories of "schools, daycare centers, public playgrounds and swimming pools, libraries, parks, and school bus stops" (Bratina, 2013). The distance varies from 300 to 3,000 feet, with the most common specification being 1,000 feet (Savage and Windsor, 2018). Some states only apply residency restrictions to high-risk offenders and/or those with child victims, including Arizona, Illinois, Louisiana, Montana, North Dakota, Ohio, South Carolina, Texas, and Wisconsin (p. 14-15). Other states, like California, have different restrictions based on the designated risk of the sex offender, which expands restrictions by 640 feet for high-risk offenders (p. 14).

Despite these laws being in place for almost 30 years, there is not much data available on the costs of these programs. Many of the costs overlap with those for registration and notification laws (Perillo, 2016). The main costs of eviction and relocation occur when the law is initially enacted. The recurring costs can stem from the incarceration of violators and from monitoring; the state of California, for instance, utilizes GPS monitoring for its high-risk offenders, which is estimated to cost $88.4 million annually. Polk County in Iowa estimated that initially implementing residency restrictions has an estimated annual cost of $2.7 million (Perillo, 2016). It is difficult to find an exact number in the literature for a per-person per year cost of these laws, but what is known supports the fact that they require substantial financial resources.

B. Theoretical Framework and Assumptions

Residency restrictions laws are implemented in order to incapacitate sex offenders and also work in this way according to the economic theory of crime. The main criminology theory used to justify these laws is the routine activity theory. This theory posits that in order for a criminal event to occur, there needs to be "a motivated offender, a suitable target, and the absence of capable guardians" close to where both the victim and criminal have their daily routine (Bratina, 2013). Therefore, not allowing sex offenders to reside in these areas lowers the probability of them encountering potential victims in familiar areas (Huebner et al., 2014). This partially incapacitates them from being able to commit the crime.

There is also a deterrence effect for potential first-time sex offenders because these restrictions add additional punishment to the crime. There is a plethora of literature and anecdotal stories that outline the difficulties these restrictions cause in reintegrating into society. (Savage and Windsor, 2018) cite studies which concluded that "sex offenders paroled under a residence restriction were significantly more likely to move 3 to 5 times than those paroled before the restrictions were put in place" (p. 13). They also discuss a specific example that in Miami-Dade County, the only place outside the limits of the residency restrictions was under a bridge. This fits into (Becker, 1968) economic framework of crime that contextualized criminals as rational actors who commit crimes when the benefit outweighs the cost. Sex offender residency restriction laws add to the costs of a crime, since it increases the punishment if caught. Therefore, in terms of Becker's model, this increases the cost for first-time and repeat offenders and acts as deterrence for both groups.

C. Prior Literature

There is extensive literature examining the efficacy of sex offender residency restriction laws. One such study by (Nobles et al., 2012) examined the impact of Jacksonville, Florida, expanding the buffer zone from the state mandate of 1,000 feet to 2,500 feet with a city ordinance in 2005. They obtained data of arrest records from 2003 to 2007. This allowed them to analyze all offenders and recidivists. They then used binary logistic regression models to estimate whether overall sex crime arrests and/or rearrests for sex crimes are impacted. Both of the coefficients were positive and statistically significant. One mechanism to explain the increase in sex crime arrests and rearrests in the post-treatment period is increased attention to these offenses. Overall, the research shows that this specific policy enhancement in Jacksonville failed to decrease sex crimes and recidivism in sex offenders.

(Socia, 2012; Kang, 2017; Huebner et al., 2014) all studied the impact of transitioning from having no residency restrictions to implementing restrictions, but on different scopes of analysis. (Socia, 2012) compared counties in New York state from 1998 to 2009 where there was no state law regarding residency restrictions, but variation of their presence within counties. They examined the impact of residency restriction on sex crimes on a variety of dimensions. They separated the analyses by both victim type (adult vs. child) and whether the criminal was a recidivist or non-recidivist. The authors utilized a fixed-effects panel model with a variety of controls for the county's overall crime rate. The only significant decrease found was for first-time offenders against adult victims indicating a deterrence effect from the increased punishment. However, they also concluded that residency restrictions did not have any incapacitation effects because the sex crime rates of recidivists were not significantly impacted by the laws. The main mechanism contributing to this is that the residency restriction laws attempt to incapacitate sex offenders from strangers, but the overwhelming majority of sex crimes occur against acquaintances.

Instead of comparing crime rates before and after treatment, (Kang, 2017) compared sex offenders released before and after the implementation of a residency restriction law in North Carolina. This law applies only to sex offenders attempting to establish residency after it was enacted. Using a difference-in-difference approach, the author found that sex offenders released after the law's implementation were significantly more likely to be reconvicted of a felony in a one- to three-year period. The only type of felony with a significant increase was property crimes, which the author notes are usually financially motivated. Another important finding was that the law decreased reconviction of a sex crime for individuals released from prison at or before the age of 30. This was the first instance in the literature of evidence that residency restrictions decreased recidivism among sex offenders.

(Huebner et al., 2014) studied Missouri's and Michigan's residency restriction laws. They matched sex offenders to non-sex offenders released into similar community parole environments as a control group. The authors used logistic regression and survival analysis as their main models. They did not find a significant effect of residency restriction laws on recidivism, whether sexual or in general. They concluded that the current policies are ineffective and need to be altered, and they suggest individualizing residency restrictions by applying it only to those who had minor-aged victims or those otherwise considered high-risk. Also, more broadly, they suggested that more resources need to be allocated to help sex offenders reintegrate themselves after release.

The present research adds to this growing literature by comparing states that did not implement laws to those that did. None of the aforementioned analyses used states without residency restrictions as a control group. Another gap this paper fills is separately analyzing the impact of sex crimes where the victims did and did not know the offender and where the offender was a minor. These laws are made specifically to incapacitate sex offenders from victimizing minor strangers, so assessing their impact on this front is crucial in analyzing their efficacy.

III. Data

A. Data Description

As this paper focuses on sex crimes, data from the National Incident-Based Reporting System (NIBRS) was used because it provides details on the crime and victim crucial to this analysis. NIBRS is a database where individual precincts report crimes, and this data has been collected since 1991 (Kaplan, 2023). This paper uses data from 2004 to 2022, but since there are different treatment years, availability differs greatly by state as well as use within the analysis. Each row of the data represents an individual crime reported to the police with a unique incident ID. The information available on each crime important to this analysis included the type of sex crime, the age of the victim, and the relationship between victim and offender. This was collapsed into panel data so that each row represents the sex crimes in a state in a year. Since the NIBRS dataset only includes information from agencies that chose to report, the proportion of agencies reporting in each state in each year is utilized to transform the number of each pertinent crime to a full state estimate. This was combined with population estimates for each state in each year provided by the (United States Census Bureau, 2023). Therefore, each row of the final data set represents the number of sex crimes per 100,000 residents.

For information on each state's sex offender residency restriction laws, I used the table provided by (Savage and Windsor, 2018) which shows the law in each state in 2008 and 2018. They provided information on whether local municipalities were able to enforce their own residency restriction laws. This enabled me to identify which states never implemented residency restriction laws and which states did between 2008 and 2018. The specifications of the laws in each of the treatment states is described in more detail in the next section.

B. Cleaning Process

Utilizing (Kaplan, 2023)'s publication of NIBRS data, I filtered sex crimes that occurred between 2004 and 2022. States included had either implemented statewide residency restriction laws in or after 2008 or had never implemented residency restrictions, even at the local municipality or county levels. States must also have had sufficient years of NIBRS data to allow analysis. According to (Savage and Windsor, 2018), Colorado, Connecticut, Kansas, Utah, and Vermont never had residency restriction laws and served as the control group. I conducted further research to find the exact timing of implementation for those identified as having laws employed between 2008 and 2018. For Montana and North Dakota, I found the actual code for the law online and found the year there (Montana Legislature, 2023)(Geographic Restrictions Applicable To High-Risk Sexual Offenders, 2023; Chapter 12.1-32: Penalties and Sentencing, n.d.). The law was implemented in Montana in 2015 and in North Dakota in 2008. Finally, for South Carolina, I found an academic resource recounting the history of in-state sex offender laws which asserts that the law took effect in 2011 (University of South Carolina, n.d.). Montana, North Dakota, and South Carolina apply their residency restriction laws to high-risk sex offenders, but the definition of high-risk varies by state. Generally, the "high-risk" designation implies that the offender's victim was a minor, or that they committed a violent offense (Savage and Windsor, 2018).[1] All three states vary in buffer zone size, with Montana having the smallest of 300 feet, North Dakota at 500 feet, and South Carolina with the largest at 1,000 feet (Savage and Windsor, 2018). North Dakota's residency restriction only applies to schools or preschools, while Montana's and South Carolina's each apply to parks, playgrounds, and other recreational facilities that mainly service minors (Savage and Windsor, 2018). The individual differences in each of these laws highlight the sheer amount of variation that exists between states.

The NIBRS data set was reduced to only observations from the states identified above. I then created binary variables for whether the recorded observation was a sex crime, as well as binary variables for each specific type of sex crime. These categories are statutory rape, rape, incest, fondling, sexual assault with an object, and sodomy. I made binary variables for whether the victim was a minor and if they knew the offender. These binary variables were then summed by state by year to have panel data with which to perform the analysis. I also created a binary variable for whether the crime was not a sex offense to be able to control for unobserved changes that affect overall crime in each state. Since state and year fixed effects are utilized, no other control variables were added to the data set.

The panel data was transformed from the total number of crimes that were reported by year by state to the rate per 100,000 residents. The FBI's Crime Data Explorer (n.d.) provides data on agencies' participation in reporting to NIBRS. I turned this into panel data that had the proportion of total agencies in each state each year and merged that with the panel data previously described. Then, after merging population estimates for each state in each year, I was able to find the sex crime rate per 100,000 residents, which is the main dependent variable of interest. This rate is available for each type of sex crime outlined above, as well as victim type and victims' relation to their offenders.

C. Summary Statistics

The descriptive statistics separated by treatment and control group can be found in Table 1 in the Appendix. The average overall sex crime rate for the treatment group is 138.88 crimes per 100,000 residents per year; for the control group, it is 136.45 crimes per 100,000 residents per year. The order of the subcategories by size for the treatment group from largest to smallest is fondling, rape, statutory rape, sodomy, sexual assault with an object, and incest. The only difference in order for the control group is that statutory rape and sodomy trade places in the order. Therefore, the order is fairly consistent between the groups. On average, the treatment group has higher sex crime rates overall and in every subcategory except for sexual assault with an object and sodomy. However, none of these differences are statistically significant. For the rate of sex crimes involving minor victims, the treatment group has 0.12 more of these crimes per 100,000 residents per year. For the stranger-to-offender rate specification, the difference is even smaller, with the treatment group having a rate of 0.83 crimes per 100,000 residents per year less than the control group. The control variable of the non-sex crime rate is also included in the table and is much larger relative to the sex crime rate. The rates are 8,509.13 and 8,002.18 per 100,000 residents per year in the treatment and control groups respectively. However, similar to the sex crime rates, this difference is not statistically significant.

IV. Empirical Model

A. Identification Strategy

Since this paper aims to capture the efficacy of residency restriction laws on decreasing sex crimes, a difference-in-difference model that also includes state and year fixed effects is utilized. The equation for this model is

$$\begin{aligned} SexCrimeRate_{it} = \beta_0 &+ \beta_1(Post_t \\ &\times ResidencyRestriction_i) \\ &+ \beta_2(NonSexCrimeRate_{it}) \\ &+ \alpha_i + \omega_t + \varepsilon_{it} \end{aligned} \tag{1}$$

$SexCrimeRate_{it}$ is the number of sex crimes that occurred in state i in the year t per 100,000 residents. In alternative specifications of the model, this variable continues to represent the rate in state i in the year t, but either for the specific types of sex crimes, for victims who are minors, or for offenders and victims who are strangers. $ResidencyRestriction_i$ is a dummy variable that equals 1 if state i had a residency restriction law implemented and is therefore part of the treatment group in the model. If $ResidencyRestriction_i$ is 0, then state i had no residency restriction law implemented and is in the control group. $Post_t$ is another dummy variable that represents whether the year t occurred on or after the implementation of the residency restriction law. For the control groups, this variable will be determined by the treatment year of the state they are being compared to, since each state has a different treatment year. The interaction of $ResidencyRestriction_i$ and $Post_t$ is the estimate of interest in the difference-in-difference model, since it shows the impact the residency restriction laws have on those with them compared to those who did not have them. $NonSexCrimeRate_{it}$ is the number of non-sex crimes that occurred in state i in the year t per 100,000 residents. This acts as a control for changes that affect crime aside from the residency restriction laws. α_i is a fixed effects term for state i, and ω_t is a fixed effects term for year t. The state and year fixed effect terms, along with the non-sex crime rate, account for any endogeneity in the model and allow for causal inference to be made. ε_{it} represents the standard errors used in this model.

The same difference-in-difference model has its results presented as an event study to see if the effects of the law change

[1] Wisconsin met all criteria except its NIBRS reporting proportion was below 10% for many years which skewed their sex offense numbers greatly and therefore needed to be dropped. Rhode Island was dropped due to decriminalizing indoor prostitution from 2003-2009 (Gordon, 2017). Their residency restriction law was implemented in 2008, so this presented endogeneity.

as time passes from its enactment. This shows the difference between the treated and untreated states in each relative year to capture if/how the impact of treatment changes with time. Nothing in the difference-in-difference model changes for this specification, just the way the results are presented.

Since there are different treatment years for each state, a traditional difference-in-difference may be biased. As demonstrated by (Goodman-Bacon, 2021), traditional difference-in-difference models include states treated early in the sample among the control group for later treated states. This is a problem if the impacts of treatment are dynamic over time.

Moreover, the traditional model ends up putting more weight on states that changed their policies in the middle of the sample window. The (Callaway and Sant'Anna, 2021) method addresses these issues by individually estimating a difference-in-difference model for each treatment year separately, then weighting them to provide an average treatment effect.

B. *Model Assumptions*

In order to employ a difference-in-difference model, four assumptions must be met: common pretreatment trends, nothing else changing after treatment for just one group, no anticipatory effects, and no changing groups. Figure 1 shows the comparison of the treated states to the control group states by their relative year. The pretreatment trends appear similar four years before treatment when the treatment group starts to see an increase in sex crimes while the control group stays the same. This is likely a result of there being only three states and scaling the number of sex crimes based on the percentage of precincts reporting. This is especially an issue since two of the treatment states, North Dakota and Montana, have relatively small populations. Furthermore, Figures 2 and 3 present the graphs for the subcategories of sex offenses and victim types and also have issues meeting the pretreatment trends assumption. Therefore, results from the model must be interpreted with caution. However, a synthetic control model is utilized as an alternative specification to alleviate this issue so that results can be interpreted with greater confidence.

As for the second assumption that nothing else changes across the threshold for only one group, no other laws regarding sex offenses were passed in this time period. As previously noted, sex offender registry laws were established prior to the use of residency restriction laws. Regarding the third assumption of anticipatory behavior, there are two ways that residency laws can impact sex offenders. Some of the laws apply to people if they are released from prison after the law is enacted. With this type of law, it would be extremely difficult for a criminal to time committing a sex offense far enough in advance and with enough certainty to avoid the residency restrictions. The other way residency restrictions can work is by applying to sex offenders even if they were released before their enactment and were forced to move. Anticipatory behavior is impossible in that case. Overall, it is difficult to impossible for potential sex offenders to have any sort of anticipatory behavior. As for the last assumption of no switching between the treatment and control group, any states that allowed for municipalities to pass their own restrictions—or had residency restriction laws that were ruled unconstitutional in state courts and subsequently taken away—were not included in the control group. Therefore, the assumptions of a difference-in-difference model are mostly met, and the analysis can be performed.

V. Results

A. *Callaway-Sant'Anna Difference-in-Difference Results*

Table 2 presents the results for the difference-in-difference model with the different sex crime specifications. These results include state and year fixed effects and a control for non-sex offense crimes. Column 1 includes all sex offenses, while columns 2-7 show the model for specific subcategories. Holding all else constant, passing a residency restriction law is associated with an increase of 23.9 sex offenses per 100,000 residents. This coefficient is significant at the 0.1% level. It also holds practical implications: the mean for pretreatment sex offenses per 100,000 residents of the treatment group was 132.34, meaning that this is a substantial 18% increase.

The specific subcategories show which types of sex offenses drive this overall increase. Columns 4 and 5 of Table 2 display that incest and fondling offenses saw statistically significant increases for the treatment states after the residency restriction laws were passed. Holding all else constant, enacting a residency restriction for sex offenders leads to an increase of 17.57 fondling offenses per 100,000 residents and 0.83 incest offenses per 100,000 residents. The fondling coefficient is significant at the 0.1% level, while the incest coefficient is significant at the 5% level. Fondling offenses account for 74% of the increase in sex offenses. The coefficient for incest offenses is much smaller and only represents 3% of the overall increase. However, incest is a relatively rare offense compared to the others, so even if the increase seems small, it represents a 33% increase in incest crimes. The coefficients for statutory rape, rape, sexual assault with an object, and sodomy are not statistically significant, but it is important to note they have all positive coefficients, so the enactment of the law did not result in a decrease for any type of sexual offense. All these results must be interpreted with great caution, however, since the pre-treatment trends assumption of the difference-in-difference model was not met.

Since these laws are designed to specifically protect minors from being victimized by a stranger, a secondary model was run with these specifications to see if that was accomplished. Similar to the main model, this secondary model had the state and year fixed effects and non-sex crime rate as a control variable. These results can be found in Table 3 in the Appendix. Column 1 shows that holding all else constant enacting the residency restriction law is associated with 20.64 more sex offenses per 100,000 residents where the victim is a minor. This is significant at the 0.1% level. Minors account for the majority of victims of sex crimes and, in this model, account for the large majority of the increase after residency restriction law enactment. 86 out of 100 new victims reported following the enactment of the law were minors. Before the law, only about 69 out of 100 sex crime victims were minors for the treatment group. This shows that, instead of protecting minors, the law leads them to be victimized at even higher rates.

A potential mechanism for this phenomenon is that the law creates a false sense of security among parents and guardians. These laws perpetuate the idea that sex offenders are generally strangers who find their victims in public places when they can be relatives, family friends, and other people children know. Therefore, there is much less vigilance in protecting children following the passage of residency restriction laws, as people believe these laws to solve the issue. The increase in stranger victimization is much smaller, with law enactment leading to 1.17 more sex offenses per 100,000 residents where the victim is a stranger to the offender, holding all else constant. This is not statistically significant, but the fact that it is not a negative number further highlights that the laws are not having the desired impact. Stranger victimization is quite rare within sex offenses, so even if the magnitude of this coefficient appears small, it still represents a 15% increase. Similar to the caveat given for the primary results, the pre-treatment parallel trend assumption is not met for these variables, either, so results must be interpreted with this in mind.

B. Callaway-Sant'Anna Method Difference-in-Difference Event Study

Figure 4 displays the average treatment effects of the treated states on sex offenses for every year relative to the law's enactment. This allows us to see whether the impact of the law changes as years pass. Although relative years 4-6 are statistically significant, the main years driving the impact of the law are years 8-14. There is a large jump in the magnitude of the law's impact in these later years. Further examining what drives this delayed impact, Figure 5 shows the average treatment effects for fondling offenses. Similarly to all sex offenses, there is a large jump in magnitude of average effect that is sustained for the rest of the years. Since fondling accounts for the majority of the increase, 74% of the increase in sex offenses, this explains why overall offenses have a delayed effect. Figure 6 presents the effect of the treatment by relative year for minor victims. The first statistically significant year is year 5, and results in years 8-14 are the greatest in magnitude. The consistency in victimization and offense results indicates that the law has a delayed effect.

C. Synthetic Control

As previously explained, the pre-treatment parallel trends assumption needed for a difference-in-difference model is not met in this analysis. This is likely due to measures of sex crimes being transformed into rates based on proportional reporting to NIBRS. With there only being three states and two of them having relatively small populations (Montana and North Dakota) it makes sense why Figures 1 and 2 show great volatility. Thus, a synthetic control model is appropriate to construct control groups for each state that best match their pretreatment trends. The control group will be constructed using specified weights of the five control states.

Montana's and North Dakota's synthetic controls are constructed from Colorado and Connecticut. Montana's synthetic control is 92.2% of Colorado and 7.8% of Connecticut while, North Dakota's synthetic control is 62.4% of Colorado and 37.6% of Connecticut. Figures 4 and 5 show how their and the synthetic control's sex crime rates change over time. These graphs show that for both states, the synthetic control still fails to meet the pre-treatment trends assumption. Table 6 presents the results from the model, with Montana's average treatment effects in column 1 and North Dakota's in column 2. All coefficients are positive, which is consistent with the results from each difference-in-difference model. However, the respective coefficients on overall sex offenses are each about double the magnitude. In Montana, holding all else constant, there is an increase of 40.86 sex offenses per 100,000 residents on average after the implementation of sex offender residency restrictions, while in North Dakota there is an increase of 39.27 sex offenses per 100,000 residents on average after the restrictions. For both states, fondling experiences the greatest increase in any of the subcategories.[2] Also of note is that in North Dakota, minor victims experience a larger increase in sex offenses per year than all victims, with an increase of 40.09 per 100,000 residents on average.

South Carolina differs from the other two treatment states in the construction of its synthetic control. The control is composed of 77.8% Vermont and 22.2% Utah. Considering Figure 9, it is apparent that South Carolina also differs from the other two treatment states in that it meets the pre-treatment parallel trends assumptions. This means that these are the results that can be interpreted with the greatest confidence in causality. These results can be found in column 3 of Table 6 in the Appendix. For the main outcome measure of all sex offenses, the magnitude of the average treatment effect is much smaller than other states' synthetic control results and the difference-in-difference results previously discussed. There was an increase of 4.48 sex offenses per 100,000 residents per year on average in South Carolina after the residency restriction law was passed. However, in the first year, this effect is much larger at 40.82 sex offenses per 100,000 residents. One explanation for this large impact in the first year is that this law causes many logistical issues for sex offenders previously living in the buffer zones and may lead to them recidivating. There also may be an increased focus on sex offenses with these laws being enacted, so law enforcement simply catches more people rather than an increase in criminality. Also, not every subcategory experiences an increase in offenses: statutory rape shows a decrease of 0.52 offenses per 100,000 residents per year on average, and stranger victims decrease by 0.19 offenses per 100,000 residents per year on average. Minor victims account for 54.0% of the sex-crime increase in South Carolina, which is much smaller than it was prior to the

[2] 24.77 fondling offenses per 100,000 residents per year for Montana, and 16.57 fondling offenses per 100,000 residents per year for North Dakota.

law and in comparison to Montana and North Dakota.

VI. Conclusion

Residency restriction laws have been enacted in many U.S. states in attempts to reduce sex crimes. They are particularly designed to protect children from strangers, since the residency restrictions usually apply to schools, daycares, and parks. This analysis focused on the impact of laws enacted between 2008 and 2015 that have sufficient data available in NIBRS. The states that fit these criteria were Montana, North Dakota, and South Carolina. States without residency restriction laws and sufficient NIBRS data were used as a control group. These states were Colorado, Connecticut, Kansas, Utah, and Vermont. Since there are different treatment years, a difference-in-difference model with the Callaway-Sant'Anna method was used. The model has state and year fixed effects, and it was found that enacting residency restriction laws result in a significant increase in sex crimes. An increase in fondling offenses primarily drove this result; incest also saw a significant increase. These laws specifically seek to decrease the incidence of offenses committed by strangers against minors, but in practice, the opposite occurred. After the institution of the laws, there was a significant increase in sex crimes involving minors as victims. The proportion of sex crime victims who were minors increased by 23.4% after the law was enacted, showing that it is not only ineffective but actually has the opposite impact than intended. Albeit not statistically significant, there was also an increase in stranger victimization. Adding an interaction term to capture whether the impact changes as time passes to this model did not significantly alter results, showing that this impact is fairly consistent over time.

A synthetic control model was also utilized to try to combat the issues with the assumption of pre-treatment parallel trends. Importantly, South Carolina was able to meet this assumption. In that state, I found that the law caused an increase in all measures except for statutory rape. Furthermore, this decrease was small in magnitude—only 0.52 offenses per 100,000 residents on average per year. Having positive coefficients provides further robustness to the results found in the main specification. Although the magnitude of the coefficients is smaller in comparison to the other results, with an increase of 4.48 sex offenses per 100,000 residents per year, this still shows this law is not at all achieving its goal of decreasing sex crimes.

This paper contributes to the growing literature that shows residency restriction laws are an unsuccessful way to reduce sex crimes. These results are similar to (Nobles et al., 2012) in finding that these laws lead to an increase in sex offenses. While (Nobles et al., 2012)—as well as other prior literature—only looked at recidivism, this paper widened the scope to all sex offenses and still found the laws to be ineffective. These findings are therefore especially important in showing that there is not a large deterrence effect. Even though past papers did not find significant impacts on recidivism, one hypothesis was that by increasing punishment for first-time offenders, the deterrence effect of the laws would be large enough to justify their existence.

This paper also adds to the literature by looking specifically at victims before and after the enactment of the law. Not only do they fail to reduce minor victimization, but they actually cause a substantial increase. A possible explanation is that the laws' enactment may make minors and adults in their lives feel safer and, consequently, let their guard down. According to this analysis, only 6.1% of all sex crimes happen to stranger victims. By feeling as if the threat was eliminated, people may be less vigilant against more likely threats, like family and acquaintances. Therefore, these laws cause not only sex offenders' lives to be more difficult but also minors' lives to be more dangerous. If this is the case, no one benefits from their implementation.

An alternative explanation is that the laws increase vigilance surrounding sex crimes and make people more likely to notice and report them. This is evidenced by South Carolina's huge increase of 40.82 sex offenses per 100,000 residents in the first year, when there would be the most public attention surrounding the law, leading to more offenses being caught. However, the Callaway-Sant'Anna difference-in-difference model with event study found that the impact of the law is delayed until seven to eight years after implementation, which contradicts this mechanism. Therefore, there is contradicting evidence on whether this is a valid explanation. Even if it is, an expensive and disruptive law is an ineffective way to achieve this result, and other avenues of sex crime prevention could be explored instead.

Therefore, the policy implications are very clear: residency restriction laws should not be utilized. States that currently have them should work to repeal them, and states that do not should keep it that way. Furthermore, no federal law stipulating that these are adopted should be passed. Instead, people should be educated about the most likely perpetrators of sex crimes. It is important to repeal these laws because they emphasize the stereotype that a sex offender is just a stranger when, in fact, they usually know the victim personally. Residency restrictions are also extremely damaging to sex offenders, making it difficult to find housing and employment and rendering the transition back into normal life more difficult.

Since the main model fails to meet the parallel trends assumption, it would not be right to conclude with certainty that the laws increase sex offenses. However, the results are compelling enough to show that they fail to decrease them and, therefore, are not useful. The synthetic control model for South Carolina provides more confidence in these results, since it did not find any evidence of a decrease in sex offenses. Another limitation is that this analysis only looks at the effects of this law in three states, so it may not be representative of the results for every other state in the country. Also, in all of the states examined, the law applied only to high-risk sex offenders, whereas similar laws in other states apply to all sex offenders regardless of the offense. Future research can incorporate more states for more representative results, which may help to avoid issues with pre-treatment trends.

References

ACLU Massachusetts (2015). Massachusetts high court unanimously strikes down lynn sex offender residency restrictions. Accessed 2025-04-07.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76:169–217.

Bratina, M. (2013). Sex offender residency requirements: An effective crime prevention strategy or a false sense of security? *International Journal of Police Science & Management*, 15:200–218.

Callaway, B. and Sant'Anna, P. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

FOX News Network (2015). New jersey towns cannot ban sex offenders from living near kids, state court rules. Accessed 2025-04-07.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Gordon, E. (2017). Prostitution decriminalized: Rhode island's experiment.

Huebner, B. M., Kras, K. R., Rydberg, J., Bynum, T. S., Grommon, E., and Pleggenkuhle, B. (2014). The effect and implications of sex offender residence restrictions. *Criminology and Public Policy*, 13(1):139–168.

Kang, S. (2017). The consequences of sex offender residency restriction: Evidence from north carolina. *International Review of Law and Economics*, 49:10–22.

Kaplan, J. (2023). Jacob kaplan's concatenated files: National incident-based reporting system (nibrs) data, 1991-2023.

Montana Legislature (2023). Geographic restrictions applicable to high-risk sexual offenders. https://leg.mt.gov/bills/mca/title_0450/chapter_0050/part_0050/section_0130/0450-0050-0050-0130.html. Accessed 2025-04-07.

Nobles, M. R., Levenson, J. S., and Youstin, T. J. (2012). Effectiveness of residence restrictions in preventing sex offense recidivism. *Crime & Delinquency*, 58(4):491–513.

Office of Justice Programs (n.d.). Legislative history of federal sex offender registration and notification. https://leg.mt.gov/content/Committees/Interim/2019-2020/Law-and-Justice/Committee-Topics/SJ-19-Study/sorna-legislative-history.pdf.

Perillo, A. (2016). The economics of sex offender policy and prevention. In Jeglic, E. and Calkins, C., editors, *Sexual Violence*. Springer, Cham.

Savage, J. and Windsor, C. (2018). Sex offender residence restrictions and sex crimes against children: A comprehensive review. *Aggression and Violent Behavior*, 43:13–25.

Socia, K. M. (2012). The efficacy of county-level sex offender residence restrictions in new york. *Crime & Delinquency*, 58(4):612–642.

United States Census Bureau (2023). Population and housing unit estimates tables.

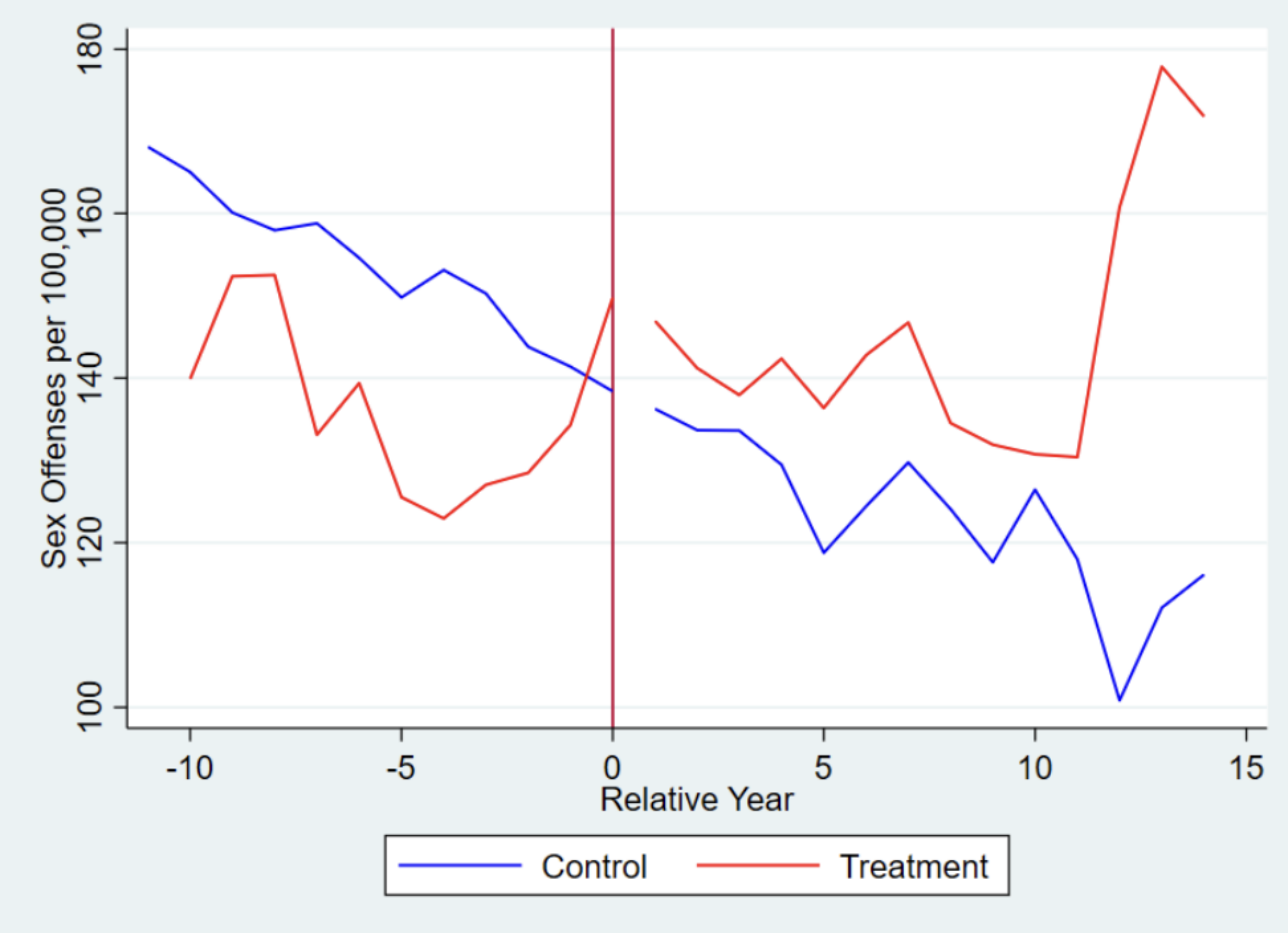


Fig. A.1: Treatment state sex offenses compared to the control group of states

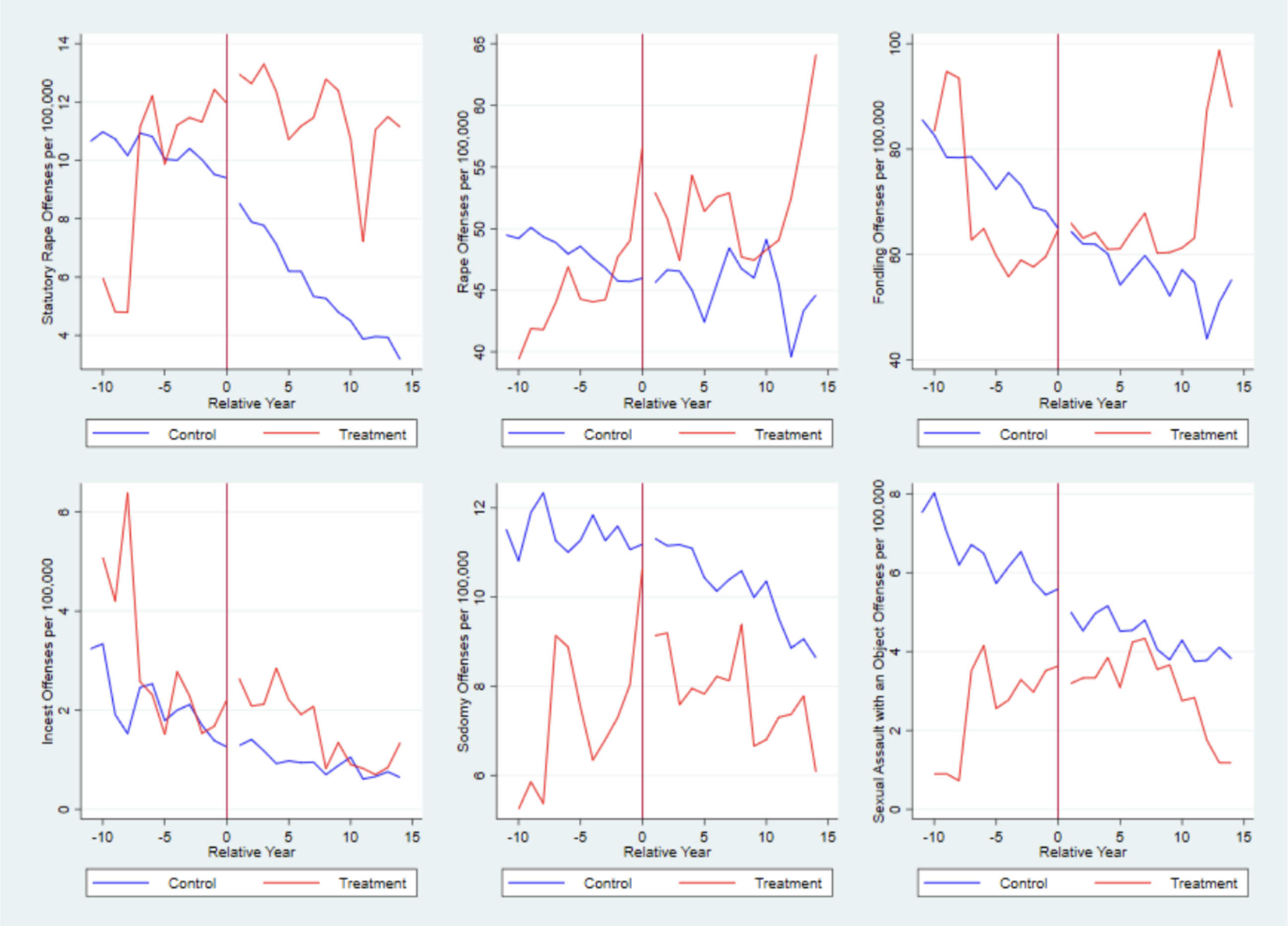


Fig. A.2: Treatment to control state comparison for each category of sex offense

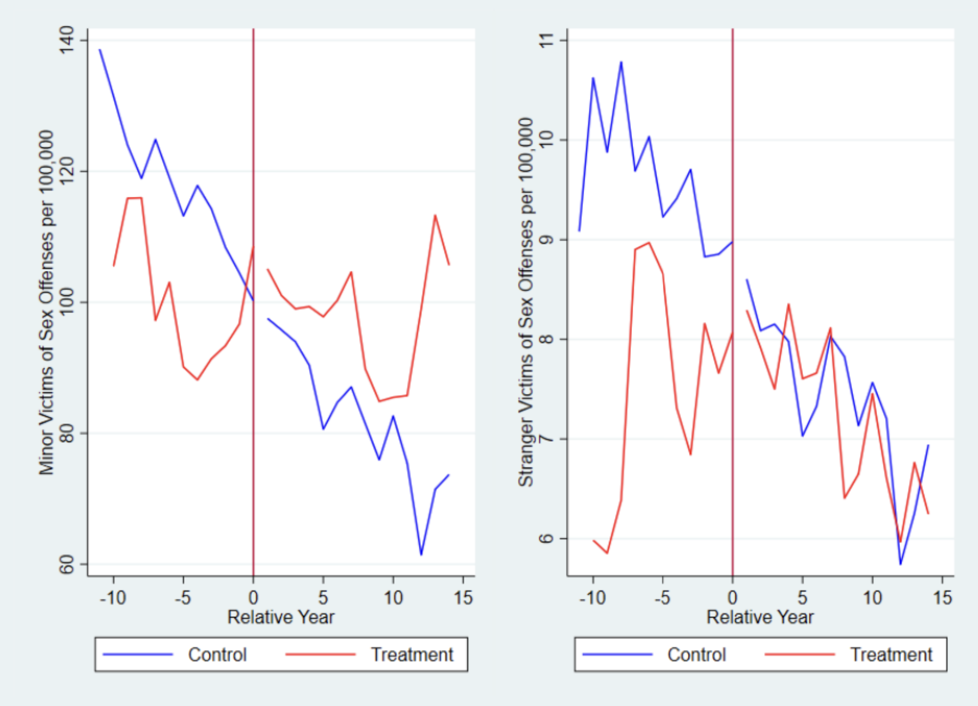


Fig. A.3: Treatment to control state comparison for each victim type

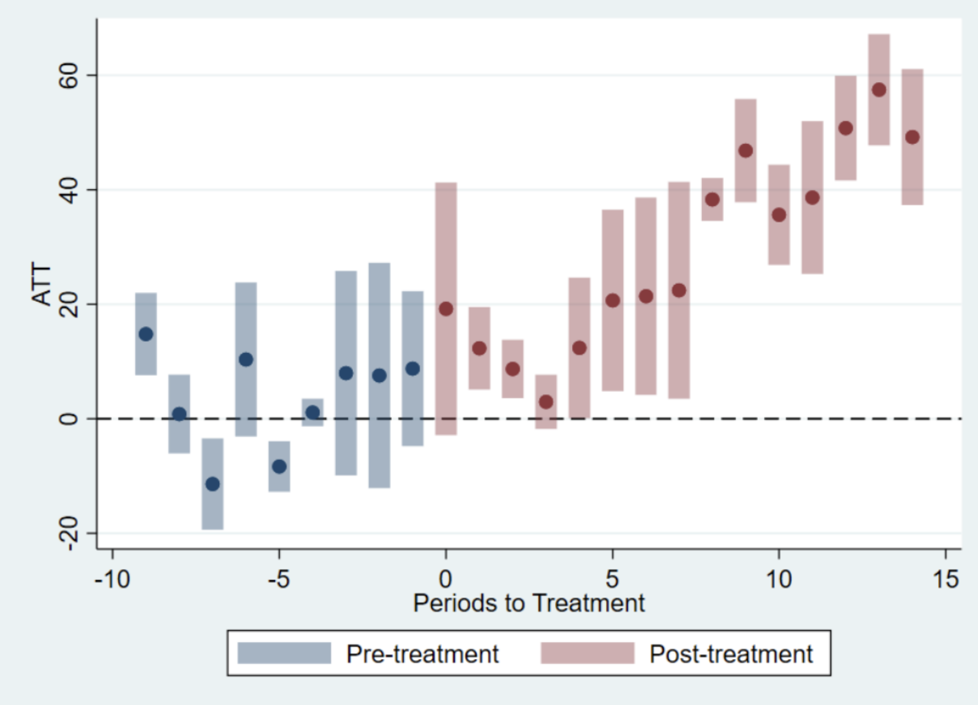


Fig. A.4: Average treatment effect of treated for each relative year for overall sex offenses

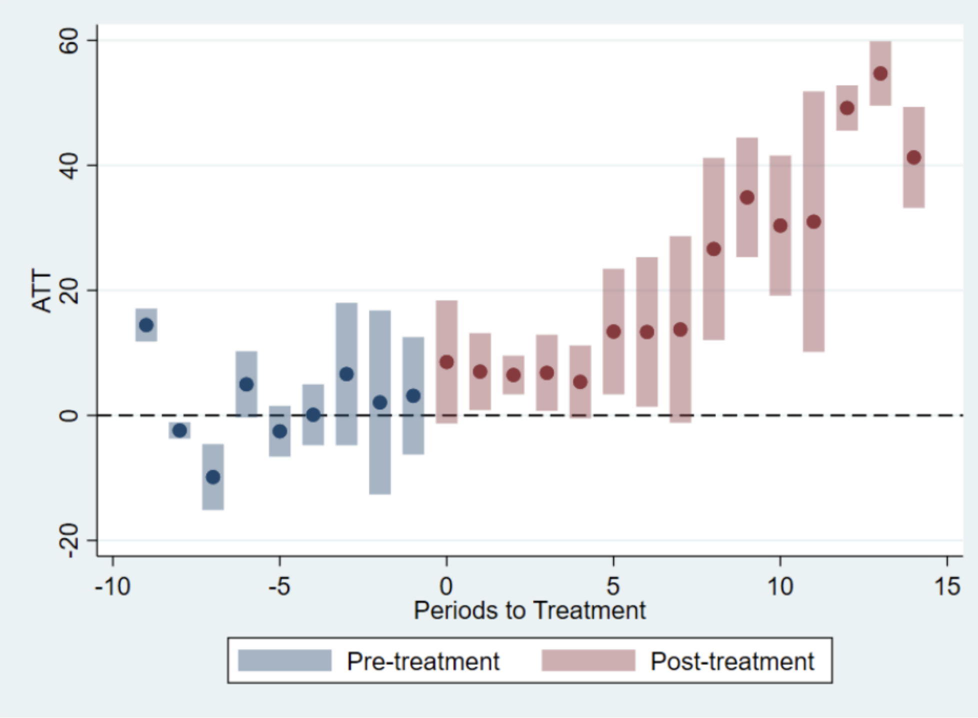


Fig. A.5: Average treatment effect of treated for each relative year for overall sex offenses

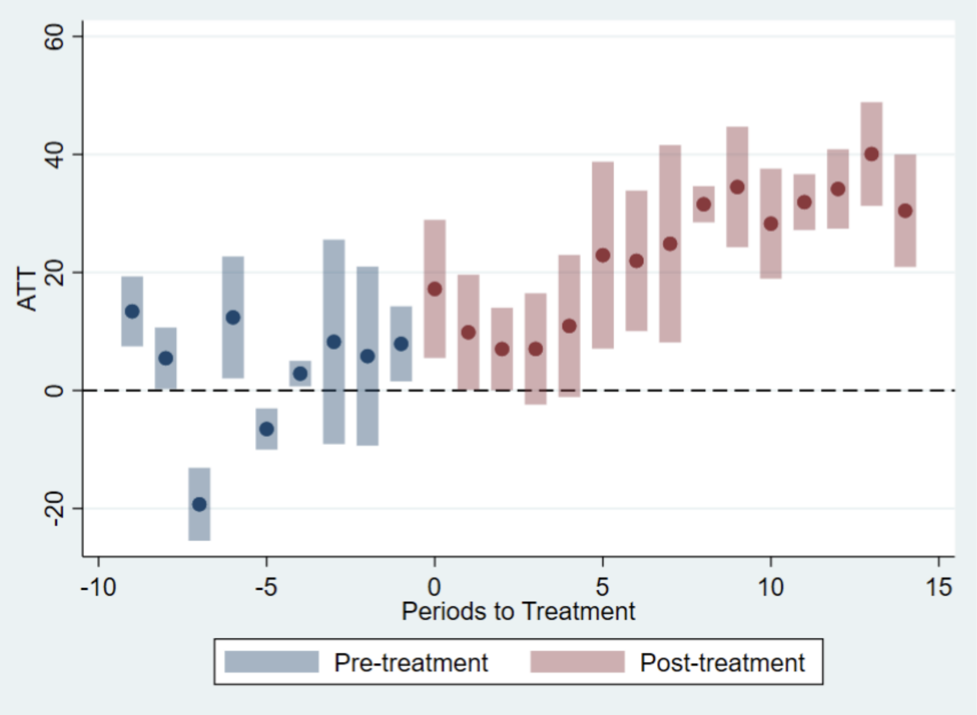


Fig. A.6: Average treatment effect of treated for each relative year for minor victims

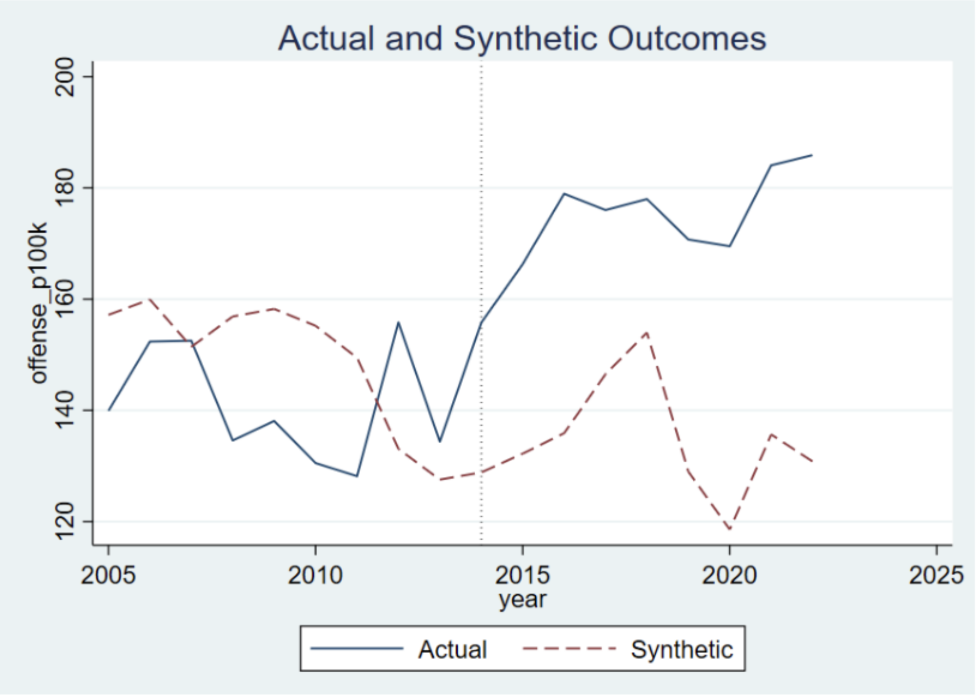


Fig. A.7: Montana synthetic control model

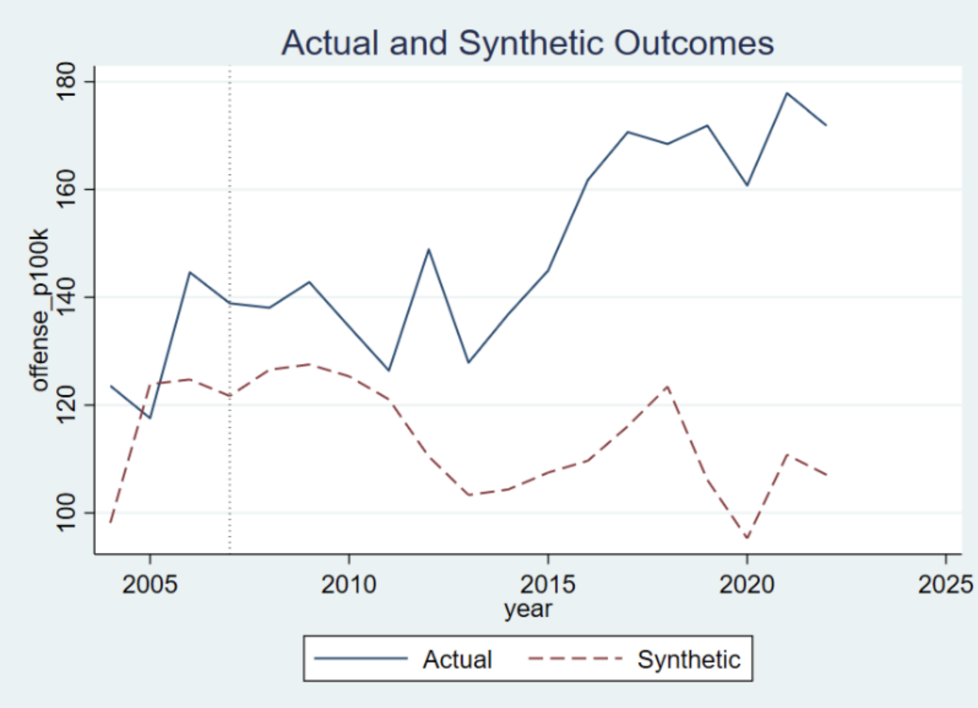


Fig. A.8: North Dakota synthetic control model

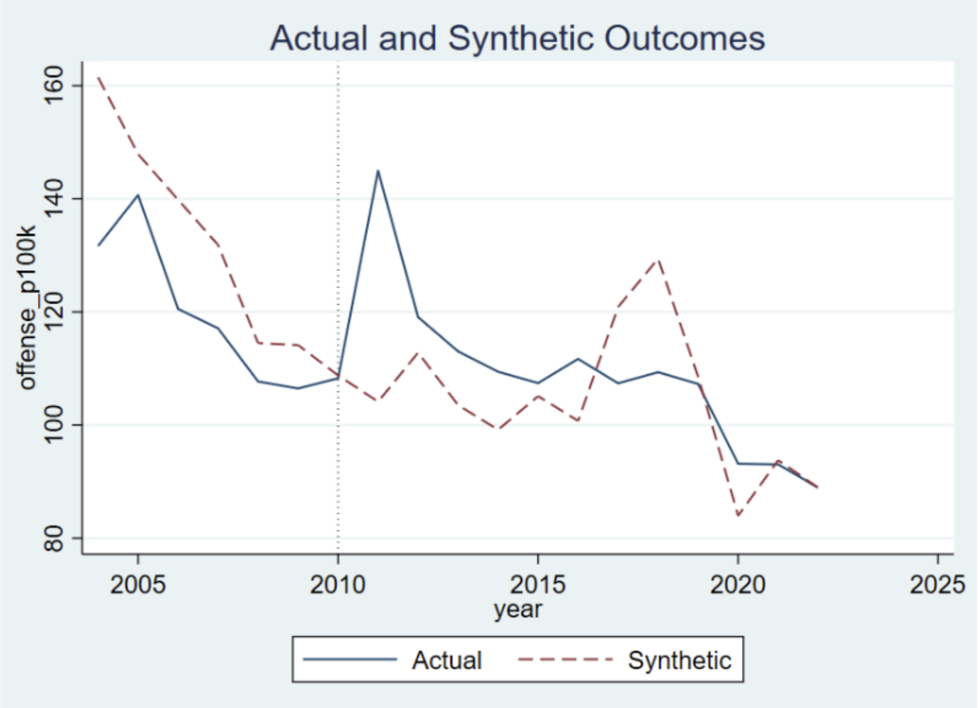


Fig. A.9: North Dakota synthetic control model

Appendix B: Tables

TABLE B.1: *Descriptive statistics per 100,000 residents*

	(1) Treatment (N=56)	(2) Control (N=95)	(3) Difference
Overall sex crimes	138.88	136.45	2.43
	(26.10)	(98.88)	(45.91)
Statutory rape	11.27	7.89	3.38
	(7.23)	(6.29)	(4.42)
Rape	49.37 (7.53)	46.44 (24.39)	2.93 (11.08)
	(7.53)	(24.39)	(11.08)
Incest	2.11 (1.79)	1.41 (1.50)	0.70 (1.09)
	(1.79)	(1.50)	(1.09)
Fondling	65.09 (22.25)	64.60 (57.30)	0.49 (28.24)
	(22.25)	(57.30)	(28.24)
Sexual assault with an object	3.17 (1.82)	5.25 (6.61)	-2.08 (2.89)
	(22.25)	(57.30)	(28.24)
Sodomy	7.87 (2.62)	10.86 (8.49)	-2.99 (3.94)
	(2.62)	(8.49)	(3.94)
Sex crime with minor victim	93.55 (16.83)	93.43 (77.60)	0.12 (34.10)
	(16.83)	(77.60)	(34.10)
Sex crime with offender who is a stranger to victim	7.59 (1.94)	8.42 (5.90)	-0.83 (2.64)
	(1.94)	(5.90)	(2.64)
Non-sex crimes	8509.13 (1798.03)	8002.18 (4564.90)	506.95 (2137.55)
	(1798.03)	(4564.90)	(2137.55)

Notes: * p<0.05, ** p<0.01, *** p<0.001. Standard deviations in parentheses for Columns 1 and 2 while robust standard error for Column 3. Column 3 presents a regression with the dependent variable as the variable of interest and the independent variable a binary indicator of the state being in the treatment group. All variables are rates per 100,000 residents.

TABLE B.2: *Callaway-Sant'Anna method difference-in-difference results by type of sex offense*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Offense	Sex offenses	Statutory rape	Rape	Incest	Fondling	Sexual assault with object	Sodomy
Implementation of law	23.90***	1.23	3.13	0.83*	17.57***	0.36	0.77
	(2.41)	(0.88)	(2.28)	(0.26)	(3.66)	(1.03)	(0.77)
Variable mean	137.35	9.14	47.53	1.67	64.78	4.48	9.75
Num. of observations	151	151	151	151	151	151	151

Note: * p<0.05, ** p<0.01, *** p<0.001. Standard errors in parentheses. All outcome variables are measured in offenses per 100,000 residents.

TABLE B.3: *Callaway-Sant'Anna method difference-in-difference results by victim type*

	(1)	(2)
Victim Type	Minor Victim	Stranger Victim
Implementation of law	20.64***	1.17
	(2.25)	(0.88)
Variable mean	93.47	8.11
Num. of observations	151	151

Note: * p<0.05, ** p<0.01, *** p<0.001. Standard errors in parentheses. All outcome variables are measured in offenses per 100,000 residents.

TABLE B.4: *Synthetic control average treatment effect by state for each measure*

	(1)	(2)	(3)
State	Montana	North Dakota	South Carolina
Sex Offense	40.86	39.27	4.48
Statutory rape	1.78	3.07	-0.52
Rape	8.71	2.73	1.91
Incest	3.15	0.12	0.09
Fondling	24.77	16.57	2.01
Sexual assault with an object	1.54	0.11	0.95
Sodomy	1.58	0.64	0.69
Minor victim	40.09	32.07	2.42
Stranger victim	2.49	0.51	-0.19

Note: All outcome variables are measured in offenses per 100,000 residents.

Permutation Tests for Equality of Variance Applied to the Problem of Clustering

Johnny O'Meara

University of Notre Dame

Abstract—**This paper makes two contributions. First is proposing a permutation test for testing equality of variances in two populations. The test is easy to implement, has exact size in finite samples under the sharp null, and has correct size in large samples. Monte Carlo simulations allow for a comparison between the permutation test and classical tests for equality of variances that point to many settings where the permutation test outperforms classical tests. Second is to demonstrate that testing for clustering in regression analysis can be thought of as testing for equality of variances. Finally, the permutation test is empirically illustrated using data from Tennessee's Project STAR. The permutation test indicates that clustering should be done at the classroom level, which is consistent with previous findings.**

I. Introduction

We often look at a mean as a meaningful summary statistic for a given dataset, but fail to acknowledge measures of dispersion such as variance. More recently, publications have featured a variety of empirical examples in favor of looking at the variance parameter as a beneficial statistical measure. This paper reviews existing tests for equality of variances of two populations and proposes a new permutation test. We assess the performance of the new test using simulations and illustrate its usefulness in the context of inference with clustered data. We demonstrate that testing for the level of clustering can be viewed as testing for equality of variances and thus benefits from our permutation test.

Empirical studies where the variance parameter is the main interest have grown in popularity in the past decades. Studies in economics that do so abound in a variety of sub-fields of the discipline. Jensen (2007) has looked at price volatility in Indian markets to study arbitrage opportunities and inefficiencies in the fishing industry. Engle (1982) introduced the ARCH model which was the first time series model to account for conditional volatility in the regression. The ARCH model has been heavily used in economics and finance, with applications in risk management (Xekalaki and Degiannakis, 2010), the international stock market (Fornari and Mele, 1997), and the Black-Scholes model (Hull and White, 1987). This important body of work exemplifies the key role of the variance parameter in many contexts.

This paper tests proposes a new permutation test for testing equality of variances between two populations based on the theory of robust permutation tests of Chung and Romano (2013) and Bertanha and Chung (2023). There are several classical tests available in the literature: the F test, Bartlett's test, Levene's test, and Brown-Forsythe's test. However, these tests rely on strong distributional assumptions. This paper studies the performance of the permutation test in simulations and finds that the permutation test exhibits superior size control in many settings, at the cost of increased computational complexity. A second contribution is to show that the problem of testing for the right level of clustering in a regression amounts to the problem of testing equality of variances. Using our derivations, researchers can use a test for equality of variances to determine the best level of clustering.

The paper features an empirical illustration that combines the problem of clustering with testing for equality of variance. To do so, we utilize data from Tennessee's Project STAR, an initiative brought about by Tennessee Governor Alexander in the 1980s that aimed to study if class size impacted student performance, as measured through test scores. This example is reasonable because it imposes three potential levels of clustering: student, classroom, or school. In previous literature related to Project STAR, MacKinnon et al. (2023) created novel tests to determine the right level of clustering. They find that the classroom level is the appropriate level. We first replicate MacKinnon's results and then use various equality of variance tests to assess the proper level of clustering. Generally speaking, tests that were found to behave well in simulations tend to agree that clustering should be done at the classroom level, and this is also the finding of the permutation test.

The rest of this paper is organized as follows. Section 2 reviews the classical tests for equality of variance. Section 3 presents Monte Carlo simulations that we design to assess the performance of the various tests including the permutation test that we propose. Section 4 bridges the gap between testing for equality of variance and testing for clustering by deriving new variables whose variances we want to compare in order to determine the right level of clustering. Section 5 introduces Project STAR and provides results on the level of clustering using our tests for equality of variance. Finally, Section 6 summarizes the paper and provides future steps for investigation.

II. Testing Equality of Variances

This section reviews the classical tests for equality of variance. It also proposes a natural test that we construct based on an asymptotically normal approximation and the section ends with our proposed permutation test. Historically, there have been a wide variety of statistical tests used for testing equality of variance. Each of these tests follows the same null hypothesis that two or more population variances are equal. We focus on the case of two populations for simplicity. Given two independent samples, $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} F_X$ and $Y_1, ..., Y_m \overset{\text{i.i.d.}}{\sim} F_Y$, the researcher specifies a nominal significance level $\alpha \in (0, 1)$ and desires to test the following null hypothesis:

$$H_0 = \sigma_X^2 = \sigma_Y^2;$$
$$H_1 = \sigma_X^2 \neq \sigma_Y^2.$$

A. The F Test

First and foremost, the F-test is the most basic test and can be computed given two independent and identically distributed samples drawn from normal populations. The F-test's test statistic uses the sample variances from the samples to produce the equation

$$F = S_X^2/S_Y^2$$

where

$$S_X^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2 \text{ and } S_Y^2 = \frac{1}{m-1}\Sigma_{i=1}^{m}(Y_i - \bar{Y})^2$$

and where

$$\bar{X} = \frac{1}{n}\Sigma_{i=1}^{n}X_i \text{ and } \bar{Y} = \frac{1}{m}\Sigma_{i=1}^{m}Y_i.$$

The F test is a two-sided test due to the fact that one needs to reject for both significantly small ratios produced from the test statistic as well as significantly large ratios. Let $G(f; n-1; m-1)$ be the CDF of a F-distribution with n-1 numerator degrees of freedom and m-1 denominator degrees of freedom. The upper critical value c_2 is such that $G(c_2; n-1; m-1) = 1 - \alpha/2$ and the lower critical value c_1 is such that $G(c_1; n-1; m-1) = \alpha/2$. The test rejects when F $< c_1$ or F $> c_2$.

B. Bartlett's Test

Another test for equality of variance originating not long after the F-test was Bartlett's test. Maurice Bartlett designed Bartlett's Test as a means of testing equality of variance for k samples as opposed to only two. The Bartlett test has been proven to be consistent (Brown, 1939) and unbiased (Pitman, 1939). However, the Bartlett test lacks compared to other tests because it is non-robust, thus making it extremely sensitive to departures from normality. It is so sensitive to departures from normality that some economists (Box, 1953) have deemed it to be a test for population normality on its own.

In our case, for two populations, its test statistic can be defined as

$$T = \frac{(N-2)\ln s_p^2 - \Sigma_{i=1}^{2}(N_i - 1)\ln s_i^2}{1 + (1/3)((\Sigma_{i=1}^{2}1/(N_i - 1)) - 1/(N-2))}$$

where

$$s_p^2 = \frac{1}{N-2}\Sigma_i^2(N_i - 1)s_i^2$$

and s_i^2 is the variance of the i-th group for $i \in (X, Y)$.

The test statistic utilizes both the individual sample variance as well as the pooled variances, where the pooled variance takes the weighted average of the group variances. N is the total number of samples, such that N = n + m. If $G(t; k-1)$ is the CDF of a Chi-square distribution with k-1 degrees of freedom, then the critical value c_1 comes from the upper tail such that $G(c_1; k-1) = 1 - \alpha$. The test rejects when T $> c_1$. It is a one-sided upper tail test due to the fact that the ratios only increase for different variances. Thus, as variances become more and more different across populations, the test statistic increases.

C. Levene's Test

One problem with both the F-test and Bartlett's test is that they perform poorly when the populations deviate from a normal distribution. Realizing that all these previous statistical tests of the equality of variance relied on the assumption of normality, Professor Howard Levene proposed a new test in 1960 that applied the F-test to observations' absolute deviations from group means. Since being developed, Levene-type tests have been applied to a wide range of data including fields such as climate change (Henriksen, 2003), food quality (François et al., 2006), and sports (Cumming and Hall, 2002).

Levene's test statistic can be defined as

$$W = \frac{N-2}{k-1}\frac{\Sigma_{i=1}^{2}N_i(Z_i. - Z..)^2}{\Sigma_{i=1}^{2}\Sigma_{j=1}^{N_i}(Z_{ij} - Z_i.)^2}$$

where

$$Z_i. = \frac{1}{N_i}\Sigma_{j=1}^{N_i}Z_{ij}$$
$$Z.. = \frac{1}{N}\Sigma_{i=1}^{2}\Sigma_{j=1}^{N_i}Z_{ij}$$
$$Z_{ij} = |Y_{ij} - \bar{Y_i}|$$

$\bar{Y_i}$ is the mean of group i for i $\in (X, Y)$ and Y_{ij} is the specific value of the data point in the j-th case of group i. For our case of two populations, j is either $\in (X_1, X_2, ..., X_n)$ or $\in (Y_1, Y_2, ..., Y_n)$. Let $G(w; N-k)$ be the CDF of a F-distribution with N-2 degrees of freedom. The upper critical value c_1 is such that $G(c_1; N-k) = 1 - \alpha$. The test rejects when W $> c_1$.

D. Brown-Forsythe's Test

Despite solving the problem of normality, Levene's test lacked in other areas according to Morton Brown and Alan Forsythe. They argued that the test statistic was not robust if the underlying distribution was skewed. Therefore, they proposed using the median in the Z_{ij} term instead. The Brown-Forsythe's test statistic remains the same as the test statistic, W, from Levene's test, with the exception of defining Z_{ij} in a new manner:

$$W = \frac{N-k}{k-1}\frac{\Sigma_{i=1}^{k}N_i(Z_i. - Z..)^2}{\Sigma_{i=1}^{k}\Sigma_{j=1}^{N_i}(Z_{ij} - Z_i.)^2}.$$

Here, we can define Z_{ij} using the median, where $\tilde{Y_i}$ is the median of the i-th group:

$$Z_{ij} = |Y_{ij} - \tilde{Y_i}|.$$

Brown-Forsythe's test works well with skewed distributions as well as distributions with heavy tails relative to Levene's test. If $G(w; N-k)$ is the CDF of a F-distribution with N-k degrees of freedom, then the upper critical value c_1 is such that $G(c_1; N-k) = 1 - \alpha$. The test rejects when W $> c_1$.

E. Asymptotically Normal Test

It is also helpful to think of a way to construct a test statistic that does not rely on any specific distribution assumption. The previous tests above fail due to the assumptions they make about the distributions of the data. The test we propose here utilizes the Central Limit Theorem to approximate the large sample distribution of the test statistic.

By applying the Central Limit Theorem, we find that the limiting distribution will be the normal distribution in large samples. However, the Central Limit Theorem only works for sample averages. Thus, we need to construct a test statistic that resembles a sample average. In this case, the sample variance can be seen as a sample average and allows us to use the Central Limit Theorem to construct an asymptotically normal distribution. The test statistic, S, can be defined as follows

$$S = \sqrt{n+m}\frac{\hat{W} - \hat{U}}{\sqrt{\frac{n+m}{n}V\hat{a}r_w + \frac{n+m}{m}V\hat{a}r_u}}$$

where

$$\hat{W}_i = (x_i - \bar{x})^2, \quad \hat{W} = \frac{1}{n}\Sigma\hat{W}_i$$

$$V\hat{a}r_w = \frac{1}{n}\Sigma(\hat{W}_i - \hat{W})^2$$

and

$$\hat{U}_i = (y - \bar{y})^2, \quad \hat{U} = \frac{1}{m}\Sigma\hat{U}_i$$

$$V\hat{a}r_u = \frac{1}{m}\Sigma(\hat{U}_i - \hat{U})^2.$$

The test statistic, S, is a two-tailed test because the numerator, $\hat{W} - \hat{U}$, produces both negative and positive outcomes depending on the difference of sample variances in the two samples. Let $G(s)$ be the CDF of a standard normal distribution. The upper critical value c_2 is such that $G(c_2) = 1 - \alpha/2$ and the lower critical value c_1 is such that $G(c_1) = \alpha/2$. By symmetry of G, we have that $c_2 = -c_1$. The test rejects when S $< c_1$ or S $> c_2$.

F. Permutation Test

This paper advocates for the permutation test because it is both practical and intuitive to implement. It exhibits exact size control in small samples, in the particular case of the sharp null; outside of that particular case, recent work demonstrates how to construct the test such that it also controls size in large samples. The permutation test relies on the same test statistic of the previous section, that is

$$T_n^{\pi} = S = \sqrt{n+m}\frac{\hat{W} - \hat{U}}{\sqrt{\frac{n+m}{n}V\hat{A}R_w + \frac{n+m}{m}V\hat{A}R_u}}$$

The permutation test originated as a test for equality of distributions and was introduced by E.J.G. Pitman in 1937. His paper described tests of significance that could be used for populations where the distribution of the sample was unknown (Pitman, 1937). Unlike the null used for the previous tests above, the original permutation test's null states that the two samples are drawn from the same underlying distribution. The "sharp null" can be defined as:

$$H_0 : F_X = F_Y.$$

The permutation test can be constructed by permuting the order of observations and stacking the X's and Y's such that $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ equals $Z_1, Z_2, ..., Z_m, Z_{m+1}, ..., Z_N$, where $N = m + n$. Next, compute a new test statistic from the unique mixture of F_X and F_Y observations. For a given permutation π, the permuted sample can be defined as $Z_n^{\pi} = (Z_{\pi(1)}, ..., Z_{\pi(n)})$ and the test statistic can be defined as T_n^{π}. The total number of permutations can be defined as P_n, where the number of elements in P_n is n!. After re-computing the test statistic for every permutation in P_n, rank the test statistic across all permutations: $T_n^1 \leq T_n^2 \leq ... \leq T_n^{n!}$. Next, set the upper, $\lfloor k^+ \rfloor$, and lower, $\lfloor k^- \rfloor$, bounds such that $\lfloor k^- \rfloor = \frac{n!\alpha}{2}$ and $\lfloor k^+ \rfloor = n!\, k^-$. Then, find the number of values greater than T_n^{k+}, the number of values less than T_n^{k-}, and the number of values equal to either T_n^{k-} or T_n^{k+}. These will be called A^+, A^-, and A^0, respectively. Finally, compute $a = (\alpha n! - A^+ - A^-)/A^0$. The outcome is based on the following system:

$$\Omega(T_n) = \begin{cases} 1, & \text{if } T_n > T_n^{k+} \text{or } T_n < T_n^{k+} \\ a, & \text{if } T_n = T_n^{k+} \text{or } T_n = T_n^{k+} \\ 0, & \text{if } T_n < T_n^{k+} \text{or } T_n > T_n^{k+} \end{cases}$$

When Ω = 1, reject the null hypothesis; when Ω = a, randomly reject with probability a; and, when Ω = 0, fail to reject the null. If the sharp null is true, such that $F_X = F_Y$, then $E[\Omega(T_n)] = \alpha$. A major advantage of the classical permutation test is that it is exact in size, under the sharp null. This means that regardless of sample size, the permutation test will always reject precisely at the set significance level. One downside to the permutation test is the cost of re-computing the test statistic for all n! permutations. However, to maintain the expectation such that $E[\Omega(T_n)] = \alpha$, Romano and Lehmann (2005) show that it suffices to use a a random sample of P_n, with or without replacement. This way, the permutation test is as expensive as the bootstrap to compute.

Since Pitman designed the permutation test and Fisher et al. (1960) used it to test the difference in heights of self-fertilized and cross-fertilized Zea may plants, researchers have expressed a desire to test a null hypothesis that is very different from the sharp null. That is, testing that just one parameter is equal as opposed to the whole distribution. If we now consider the case where $\theta(F_k)$ is a real-valued parameter, such as mean or variance, the new null hypothesis can be defined as:

$$H_0 : \theta(F_X) = \theta(F_Y).$$

In this case, the classical permutation test often fails to control for size so a new approach is required. Recent literature (Romano, 1990; Neuhaus, 1993; Chung and Romano, 2013; Bertanha and Chung, 2023) has studied this null hypothesis and determined that to control for size in finite and large samples, the test statistic must be studentized; in other words, the test statistic takes the form of the estimator for the difference in parameters divided by the standard error of this estimator. Using this new test statistic, permutation tests would attain asymptotic rejection probability equal to α even if the $F_X \neq F_Y$ under the null hypothesis that $\theta(F_X) = \theta(F_Y)$. Given a total number of observations, N, and a sample parameter, $\hat{\theta}_k = \theta_{N_k,N}(Z_{k,1}, ..., Z_{k,N_k})$, where Z_N is the combination of F_X and F_Y, the test statistic can be defined as:

$$T_N(Z_N) = \frac{(\hat{\theta_1} - \hat{\theta_2})}{\sqrt{\hat{\sigma_1}^2 + \hat{\sigma_2}^2}}$$

where $\hat{\theta_1} - \hat{\theta_2}$ is the difference of respective estimators.

The reason why we care to look at permutation test is because the permutation test promises to have good properties in small samples. In particular, for our case of equality of variances, the test statistic is the same as the test statistic for the asymptotically normal test above. If you happen to be in the case where the distributions are the same, the test is exact. More so, the robustness property says with a large enough number of observations, the distributions can be different and the test should control size. In the next section, we will use Monte Carlo simulations to compare the finite sample performance of the permutation test to the other five tests listed in this section.

III. Simulations

Let's investigate the hypothesis that the permutation test works well compared to other tests for equality of variance. We conduct Monte Carlo simulations to verify our theoretical predictions as well as explore Data Generating Process (DGP) variations that illustrate the performance of the permutation test in a variety of scenarios. The simulation results will output the rejection rates of the six tests of equality of variance from Section 2. There are three factors that will vary throughout the simulations: sample size, imbalance, and distribution. The sample sizes are N $\in$ {50, 100, 1000, 10000 }. The imbalances are $\lambda \in$ {0.5, 0.25, 0.05}, where in the case of a non-integer split, the value will be rounded to the nearest integer. With a sample size of 50 and an imbalance of 0.05, the size of sample 1 is n = 3, and the size of sample 2 is m = 47. The baseline case will be when the imbalance equals 0.5, meaning both samples receive an equal number of observations.

Additionally, there are 4 distributions that will be used: Normal, Log-normal, Chi-squared, and generalized Pareto. Monte Carlo simulations will be performed on seven specific combinations of the distributions (F_X, F_Y) $\in$ { (Normal, Normal), (Log-normal, Log-normal), (Chi-squared, Chi-squared), (generalized Pareto, generalized Pareto), (Normal, Log-normal), (Normal, Chi-squared), (Normal, generalized Pareto)}. Normal samples will be drawn from a normal distribution with μ = 0 and σ^2 = 1 while Log-normal samples will be drawn from a Log-normal distribution with μ = 0 and σ^2 = 1. Chi-squared samples will be drawn from a Chi-squared distribution with 1 degree of freedom. Finally, generalized Pareto samples will be drawn from a generalized Pareto distribution with location μ = 2, scale σ = 1/2, and shape ξ = 1/4. For the cases where the pair of distributions are different, we use parameter values such that the variances are the same, so we are always under the null hypothesis. We constructed the simulation using 10000 trials and a significance level of 0.05. For the specific case of the permutation test, 1000 permutations will be used for each trial. The goal is to provide a holistic comparison of the permutation test against other tests in both small and large samples.

		Normal-Normal					
Size	Imbalance	F Test	Bartlett	Levene	Brown-Forsythe	Normal	Permutation
50	0.5	0.0491	0.0497	0.0539	0.0397	0.0552	0.0498
50	0.25	0.048	0.0493	0.0913	0.073	0.1087	0.0491
50	0.05	0.0482	0.0468	0.4273	0.4295	0.4829	0.048817
100	0.5	0.048	0.0486	0.0482	0.0412	0.0482	0.0466
100	0.25	0.0491	0.0499	0.088	0.079	0.0814	0.04965
100	0.05	0.0512	0.0513	0.4284	0.4366	0.3567	0.05255
1000	0.5	0.0488	0.0489	0.0506	0.0501	0.0494	0.0477
1000	0.25	0.0503	0.0513	0.0895	0.0893	0.0574	0.0512
1000	0.05	0.0495	0.0489	0.3908	0.3898	0.088	0.05
10000	0.5	0.0482	0.0482	0.0473	0.0473	0.049	0.0471
10000	0.25	0.0483	0.0479	0.0895	0.0896	0.0488	0.0478
10000	0.05	0.0523	0.0525	0.3954	0.3953	0.0549	0.05035
		Pareto-Pareto					
Size	Imbalance	F Test	Bartlett	Levene	Brown-Forsythe	Normal	Permutation
50	0.5	0.4821	0.4837	0.1843	0.0398	0.0411	0.04775
50	0.25	0.4492	0.4548	0.2306	0.074	0.0831	0.0486
50	0.05	0.216	0.2466	0.4701	0.2785	0.3871	0.05245
100	0.5	0.5351	0.5362	0.1833	0.0432	0.0361	0.0493
100	0.25	0.516	0.518	0.2339	0.0782	0.0818	0.04615
100	0.05	0.3354	0.3676	0.5302	0.3555	0.3835	0.0523
1000	0.5	0.6724	0.6724	0.1634	0.043	0.0305	0.0446
1000	0.25	0.6523	0.6536	0.2284	0.0833	0.0691	0.0464
1000	0.05	0.6025	0.6077	0.5317	0.3777	0.2569	0.0502
10000	0.5	0.7402	0.7402	0.1627	0.0479	0.0368	0.0476
10000	0.25	0.7371	0.7371	0.2254	0.0841	0.0581	0.0471
10000	0.05	0.7071	0.7069	0.5387	0.3873	0.1624	0.0476

TABLE I: Simulation results for Normal-Normal and Pareto-Pareto distributions

In the simulation of two Normal distributions, all tests reject at approximately the 5 percent level for any sample size, but when samples are balanced. Theory suggests that the F Test is exact for two Normal distributions and our results support such a claim. The permutation test overcomes any imbalances as well and rejects around the 5 percent level no matter size or imbalance. Comparably, when the distributions are no longer normal, we see the robustness property affecting several other tests. For instance, the F Test and Bartlett's test perform poorly in the simulation of two generalized Pareto distributions, with rejection levels upwards of 0.7. Other tests such as the Brown-Forsythe's test suffered under unequal imbalances, but rejected around 5 percent for equal imbalances. Brown-Forsythe's test statistic relies on the median allowing it to work well even under skewed distributions. The permutation test performs great here, rejecting at approximately 5 percent for all sizes and imbalances; more so, the asymptotically normal test rejects close to 5 percent as the sample size grows which is to be expected according to the Central Limit Theorem. However, the asymptotically normal test rejects at 0.3871 in finite and imbalanced samples while the permutation test rejects at approximately 0.05. This illustrates the permutation test's strong performance in both finite and large samples.

Another part of the simulations experiment involved testing the performance of tests when sample distributions varied. When distributions are the same, theory tells us the permutation test will control size. With two different distributions however, theory says that in large enough samples the permutation

Normal-Chi-Squared							
Size	Imbalance	F Test	Bartlett	Levene	Brown-Forsythe	Normal	Permutation
50	0.5	0.2806	0.2822	0.2185	0.203	0.1572	0.1511
50	0.25	0.1905	0.187	0.185	0.1492	0.0671	0.10215
50	0.05	0.0811	0.0674	0.3271	0.144	0.2364	0.026367
100	0.5	0.2943	0.2957	0.2533	0.2915	0.1355	0.13255
100	0.25	0.2019	0.1999	0.2109	0.2193	0.0634	0.10365
100	0.05	0.0849	0.0801	0.3352	0.2062	0.2012	0.0357
1000	0.5	0.3197	0.3198	0.7105	0.9689	0.0752	0.07685
1000	0.25	0.2101	0.2092	0.6748	0.9478	0.0556	0.0703
1000	0.05	0.0875	0.087	0.5793	0.7786	0.0674	0.0596
10000	0.5	0.3226	0.3226	1	1	0.0529	0.0532
10000	0.25	0.2148	0.215	1	1	0.0476	0.05145
10000	0.05	0.0832	0.084	0.9985	1	0.0494	0.0498

TABLE II: Simulation results for Normal-Chi Squared Distributions

test will be exact. However, for the case of the finite sample with two different distributions, theory does not necessarily tell us anything about the performance of the asymptotically normal and permutation tests.

In the case of a Normal and Chi-squared distribution, the permutation test rejects around the 5 percent level in large samples and under every type of imbalance. We see the Central Limit Theorem in the asymptotically normal test where it rejects around the 5 percent level in large enough samples, but is not as consistent in smaller samples. These results align with the theories, where we see the permutation test and normal test are exact in large samples, while all other tests perform poorly. The appendix reports the four other Monte Carlo simulations that consist of comparisons of different distributions.

IV. Testing Clustering Level is Testing Equality of Variances

Recently, econometricians discovered that they can perform hypothesis tests using the variance of two groups to determine the best level of clustering in a population. This section first illustrates that testing for level of clustering is the same as testing for equality of variances and then describes the sample splitting methods used to create two independent samples of the data. Suppose you have a dataset, with individual i and group g, where i ∈ [1, N] and g ∈ [1, G]. Using OLS, where x_i is a matrix of explanatory variables, y_i is the vector of observed values, and u_i is the random error, the regression $y_i = x_i\beta + u_i$ betas can be computed as follows:

$$\beta = (\Sigma x_i' x_i)^{-1}(\Sigma x_i' y_i) = (X'X)^{-1}(X'Y)$$

Now, we need to transform the problem of clustering into a problem of testing for equality of variance. To do so, we need to create two variables: a variable for the case of no clustering, ψ^{NC}, and a variable for the case under clustering, ψ^{C}. The experiment in the first part of this section is to test if we can create a variable, ψ, such that if we take the variance of that variable it gives the same variance as the asymptotic variance of the estimator, $\hat{\beta}$. Under usual cluster asymptotics (Hansen, 2000), for the case of one β and clustering, there is an expansion that is useful because it is asymptotically normal and can be defined as follows:

$$\sqrt{G}(\hat{\beta}_j - \beta_j) = \frac{1}{\sqrt{G}}\Sigma e_j' Q^{-1} X_g' U_g + R$$

where e_j is a vector of zeros except for a value of 1 assigned to the row in the vector corresponding to the β of interest, and where X_g contains all x_i predictors for everyone in a group g. More so, Q is equal to $E[x_i' x_i]$, U is the stacked errors u_i of individuals inside group g, and R is the error of the expansion such that $R \xrightarrow{p} 0$ (Hansen, 2000). In the clustering case, $\psi_g^C = e_j' Q^{-1} X_g' U_g$. The expansion from above converges in distribution to $N(0, Var(\psi_g^C))$.

Now, we must study how the variance changes in the case where we do not have clustering. Under this condition, it is known that each individual is i.i.d., whereas with clustering, individuals may or may not be i.i.d. No clustering implies that the variance of the sum of individual variables equals the sum of the variances. Under this condition, we can redefine and simplify our variance equation using the following steps:

$$Var[e_j' Q^{-1} X_g' U_g] = e_j' Q^{-1}(\Sigma_{i=1}^{n_g} Var[X_{ig}' U_{ig}])Q^{-1} e_j$$

$$Var[e_j' Q^{-1} X_g' U_g] = \Sigma_{i=1}^{n_g} Var[e_j Q^{-1} X_{ig}' U_{ig}]$$

$$Var[e_j' Q^{-1} X_g' U_g] = n_g Var[e_j Q^{-1} X_{ig}' U_{ig}]$$

From here, we can move the n_g back into the variance equation by taking the square root to obtain our ψ_{ig}^{NC} equation:

$$Var[e_j' Q^{-1} X_g' U_g] = Var[\sqrt{n_g} e_j Q^{-1} X_{ig}' U_{ig}]$$

where

$$Var[e_j' Q^{-1} X_g' U_g] = \psi_{ig}^{NC}$$

and

$$\psi_{ig}^{NC} = \sqrt{n_g} e_j Q^{-1} X_{ig}' U_{ig}$$

If we analyze our two ψ variables, we see that clustering does in fact affect the variables as the equations are different. Next, it is important to ensure that there are two independent samples. In this thesis, we satisfy this condition using sample splitting. Our data in the following section consists of one sample with many clusters. Given that we created two ψ variables from the same data, we confirm that they are independent by splitting the sample at the level of the group.

For example, in the case of our empirical example in Section 5, let us suppose we are testing students versus classroom for level of clustering. We first split the classrooms in half meaning there would be 165 classrooms in one group and 165 classrooms in another. To compute ψ_g^C, look at the classroom level and use the equation above to compute 165 ψ_g^C variables. To compute ψ_g^{NC}, look at the remaining classrooms and then look at the number of students in each of those remaining 165 classrooms. There will be as many ψ_g^{NC} variables as the total number of students in all the remaining classrooms: the second sample will have 2034 ψ_{ig}^{NC} variables for the case of student versus classroom.

Now, we have solved two problems: namely, we transformed the problem of testing whether clustering is true into a problem of testing equality of variance and we verified that we have two independent samples using the process of sample splitting. Our third and final problem results in part because we do not observe ψ_g^C and ψ_{ig}^{NC} directly as Q and U are not known. Therefore, we need to estimate ψ_g^C and ψ_{ig}^{NC}.

For $\hat{\psi}_g^C$, we can define $\hat{U}_g$ as a vector of stacked residuals from the regression and $\hat{Q}$ as follows:

$$\hat{Q} = \frac{1}{G}\Sigma x_i' x_i$$

Using these new variables, we obtain

$$\hat{\psi}_g^C = e_j' \hat{Q}^{-1} X_g' \hat{U}_g$$

For $\hat{\psi}_{ig}^{NC}$, we can define $\hat{U_{ig}}$ as the residual value for each individual and $\hat{Q}$ as follows:

$$\hat{Q} = \frac{1}{n}\Sigma x_i' x_i$$

Using these new variables, we obtain

$$\hat{\psi}_{ig}^{NC} = \sqrt{n_g} e_j \hat{Q}^{-1} X_{ig}' \hat{U}_{ig}$$

Since $\hat{Q}$ and $\hat{U}$ are consistent estimators for Q and U, the estimated ψ variables are approximately true ones.

V. Empirical Example

As economics develops, it has become harder to justify the claim that random errors are uncorrelated in a regression. Rather, it is becoming more popular to cluster standard errors that are robust against variation within clusters. Beginning in the 1990s, the assumption of uncorrelated errors became less widely accepted in empirical work. Rather, programs such as Stata began to incorporate cluster-robust standard errors to allow for patterns for correlations within clusters. In health studies, disturbances of outcomes could be clustered by doctor, hospital, or group of hospitals (MacKinnon, 2019). In housing economics, disturbances could be clustered at the village, county, state or country level. In our case, clustering for disturbances can be at the student, classroom, or school level.

In academic literature, Moulton (1990) writes about the pitfall of seriously biased standard errors that result from incorrect clustering. He uses an empirical illustration of a wage regression that requires clustering correctly by the geographic state to prevent spurious regression. Others such as Bertrand et al. (2004) argue that difference-in-difference standard errors may severely understate the standard deviation which can cause inflated rejection results. He mentions that many papers involving difference-in-difference fail to acknowledge the serial correlation in disturbances. Cameron and Miller (2015) provide a comprehensive review of cluster-robust inference. In particular, Section 3 of their paper describes the consequences of clustered errors. Most recently, MacKinnon et al. (2023) published a paper proposing two tests for the correct level of clustering: the asymptotic and bootstrap tests. They applied their results to an empirical example involving data from Tennessee's Project STAR. This thesis uses the same dataset, but ultimately tests the permutation test's performance in determining the level of clustering for students in Project STAR.

The Tennessee class size project, created by Governor Lamar Alexander, began in 1985 as a three-phase study with the goal of determining the effect of small classes in early grades on pupil performance. This four year study measured test scores from the Stanford Achievement Test (SAT) and Tennessee Basic Skills First (BSF) test for students in Kindergarten, First Grade, Second Grade and Third Grade. Three kinds of groups were used for the study: small classes with 13 to 17 students, regular classes with 22 to 25 students, and regular classes with a teacher's aide that contained the same number of pupils as a regular class. Our main focus will be on Phase 1, the Project STAR, but two others phases, the Lasting Benefits Study and Project Challenge, were incorporated afterwards to study long-term effects of being in smaller classes. Project STAR ultimately concluded that small classes have a positive impact on student performance.

For the case of this thesis, however, we should turn our focus to the problem of clustering involved in Project STAR. The dataset provided includes samples of 3989 students, 330 classrooms, and 79 schools from all around Tennessee. For the purpose of this thesis, three comparisons will be made for the proper clustering level: student vs classroom, classroom vs school, and student vs school. To begin, we first used OLS to compute the Betas of our regression. The regression includes 17 explanatory variables and can be written as follows:

$$\text{read-one}_{sgi} = \alpha + \beta_s \text{small-class} + \beta_a \text{aide-class}_{sg} + x_{sgi}^T \delta + u_{sri}$$

The outcome variable is the test score of a student, i, in grade one from school s and classroom g. Small-class and aide-class are dummy variables: if a student was in a small-class in grade one then small-class equals 1 and 0 otherwise. Aide-class is defined in the same manner, but depending on whether a student was in a class with a teacher's aide. All additional control variables are stored in the vector x_{sgi}. Such controls include dummy variables regarding whether the pupil was non-white, male, on free lunch, and whether their teacher was non-white. Other controls included the teacher's years of experience and the student's test scores from kindergarten. Lastly, there are dummy variables depending on which quarter of the year a student was born, the student's birth year, and the highest degree obtained by the teacher. The OLS estimates for the regression model can be found in the table

Estimates		Student	Classroom	School
small	$\hat{\beta}_s$	9.211	9.211	9.211
	s.e.	1.627	3.191	3.150
aide	$\hat{\beta}_a$	6.245	6.245	6.245
	s.e.	1.658	3.248	2.765

TABLE III: Estimates and Standard Errors

Three standard errors can be found for each coefficient along with the estimates for each coefficient in Table 3. Classroom and school levels use cluster-robust standard errors while the student level uses robust standard errors. Initially, clustering at the classroom level seemed logical considering treatment was assigned at that level. However, given students from the same school most likely had similar characteristics, clustering at the school level also seems plausible. The estimated effect of being in a small class on test scores is 9.211, holding all other variables constant. Likewise, holding all other variables constant, the effect of being in a class with an aide on a student's test score is 6.245.

Cluster Tests	Coefficient: $\hat{\beta}_S$					
p-values	F Test	Bartlett	Levene	Brown Forsythe	Normal	Permutation
H_N vs H_G	0	0	0	0	0.0001	0
H_N vs H_S	0.0001	0.0002	0	0	0.0339	0.004
H_G vs H_S	0.1143	0.10232	0.9100	0.9425	0.11577	0.364
	Coefficient: $\hat{\beta}_A$					
p-values	F Test	Bartlett	Levene	Brown Forsythe	Normal	Permutation
H_N vs H_G	0	0	0	0	4.57E-07	0
H_N vs H_S	0	1.11E-16	0	0	0.0014	0
H_G vs H_S	0.21531	0.237	0.0216	0.0198	0.3758	0.254

TABLE IV: P-Values for Cluster Tests

Table 4 reports the p-values of various variance tests and three clustering level hypotheses. We focus on the asymptotic variances of two coefficients, small-class (β_s) and aide-class (β_a). For each specification, we analyze three hypotheses: H_N is no clustering which is the same as clustering at the student level, H_G is clustering at the classroom level, and H_S is clustering at the school level. All six tests for equality of variance have p-values listed in Table 4.

When testing H_N vs H_G, we strongly reject the null of no clustering for both the aide and small coefficients, under every test for equality of variance. This indicates we need to cluster at a more aggregate level than the student level. More so, when testing H_N vs H_S at the 5% significance level, we likewise strongly reject the null in favor of an alternative of a coarser level of clustering. To determine the proper coarser level of clustering, we test H_G vs H_S. When testing H_G vs H_S, we fail to reject under all tests for equality of variance for $\hat{\beta}_S$, but fail to reject for most tests except for Levene and Brown-Forsythe in the case of $\hat{\beta}_A$. According to the permutation test, if we fail to reject the null here, this likely indicates that the classroom level is the proper level for clustering which aligns with the results from MacKinnon et al. (2023).

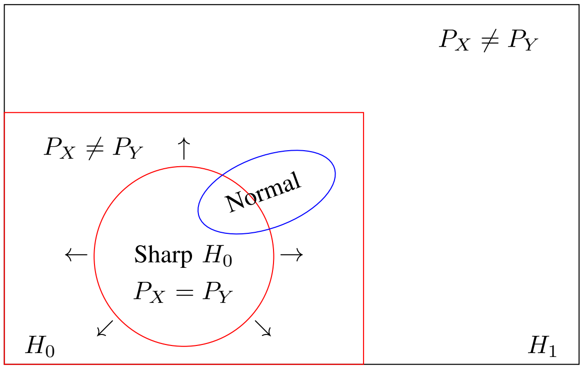


Fig. 1: Equality of Variance and Distribution Field

VI. Conclusion

This paper provides a thorough review of the permutation test's performance for testing equality of variances. More so, it provides building blocks for future studies on tests for equality of variance. In Section 4, the six tests for equality of variance that we used reacted in different manners depending on the two distributions chosen for the simulations. Figure 1 represents the complete range of scenarios presented in our simulations study. Inside the red box represents situations where the variances are equal so our null hypothesis is true. However, given the equality of a single parameter, we cannot certainly say that the distributions will be equal. The sharp null inside the red circle represents cases where the distributions are equal as well as the variance, while cases outside the red circle but inside the red box imply that the variances are equal but the distributions are not. The blue circle represents the set of normal distributions that fall under the null of equality of variances. However, this does not mean the distributions will be necessarily be equal, because the two normal distributions can have different means but the same variance. In the case of the F test, it is exact when two distributions are normal. For cases outside the red box, neither the distributions nor the variances are equal.

In the past, solutions to the problem of clustering typically relied on assumptions as opposed to statistical tests. As a result, clustering at the wrong level often resulted in smaller standard errors and over-rejecting the null. While MacKinnon et al. (2023) proposed asymptotic and bootstrap tests for the problem of clustering, this paper proves that tests for equality of variance, in particular the permutation test, can be used to determine the proper level of clustering. After proving this, we applied our theory to an empirical example using data from Tennessee Project STAR and found that our results aligned with those of MacKinnon et al. (2023). Ultimately, we found many benefits of the permutation test which come at a computational cost that is no bigger than that of the bootstrap.

For future investigation, it could be useful to test the boundaries of the sharp null circle. One could investigate different departures from the sharp null: for instance, consider two populations, one normally distributed, and another distributed as mixture of Normal and Chi-squared. Here, we could test for how long until the test fails for the Normal and Chi-squared distributions by weighting each differently. We could assign a weight to the normal distribution, λ, and another weight to the Chi-square distribution, $1 - \lambda$, and alter the weights accordingly to find the limit at which the test fails. If the weight λ equals 1, then we are in the normal case and the sharp null. If the weight λ is 0, then we are in the Chi-square case and outside the sharp null. This investigation can be expanded to all distributions used in this paper.

References

Bertanha, M. and Chung, E. (2023). Permutation tests at nonparametric rates. *Journal of the American Statistical Association*, 118(544):2833–2846.

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

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335.

Brown, G. W. (1939). On the power of the l 1 test for equality of several variances. *The Annals of Mathematical Statistics*, 10(2):119–128.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests.

Cumming, J. and Hall, C. (2002). Athletes' use of imagery in the off-season. *The Sport Psychologist*, 16(2):160–172.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Fisher, R. A. et al. (1960). The design of experiments. *The Design of Experiments.*, (7th Ed).

Fornari, F. and Mele, A. (1997). Sign-and volatility-switching arch models: theory and applications to international stock markets. *Journal of Applied Econometrics*, 12(1):49–65.

François, N., Guyot-Declerck, C., Hug, B., Callemien, D., Govaerts, B., and Collin, S. (2006). Beer astringency assessed by time–intensity and quantitative descriptive analysis: Influence of ph and accelerated aging. *Food Quality and Preference*, 17(6):445–452.

Hansen, B. E. (2000). Econometrics. *University of Wisconsin*, pages 219–240.

Henriksen, H. (2003). The role of some regional factors in the assessment of well yields from hard-rock aquifers of fennoscandia. *Hydrogeology Journal*, 11:628–645.

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, 42(2):281–300.

Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the south indian fisheries sector. *The Quarterly Journal of Economics*, 122(3):879–924.

MacKinnon, J. G. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics/Revue canadienne d'économique*, 52(3):851–881.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2023). Testing for the appropriate level of clustering in linear regression models. *Journal of Econometrics*, 235(2):2027–2056.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, pages 334–338.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, pages 1760–1779.

Pitman, E. (1939). Tests of hypotheses concerning location and scale parameters. *Biometrika*, 31(1/2):200–215.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.

Romano, J. P. and Lehmann, E. (2005). Testing statistical hypotheses.

Xekalaki, E. and Degiannakis, S. (2010). *ARCH models for financial applications*. John Wiley & Sons.

Appendix A: Tables

		Chi-square-Chi-square					
Size	Imbalance	F Test	Bartlett	Levene	Brown-Forsythe	Normal	Permutation
50	0.5	0.4001	0.4012	0.1974	0.0476	0.0549	0.05225
50	0.25	0.3818	0.3842	0.2532	0.0781	0.1067	0.04665
50	0.05	0.2326	0.253	0.5387	0.3042	0.4455	0.055233
100	0.5	0.4162	0.4168	0.1855	0.0486	0.0492	0.05045
100	0.25	0.4073	0.4098	0.2534	0.0867	0.0964	0.0509
100	0.05	0.3016	0.3214	0.5699	0.3833	0.4371	0.04985
1000	0.5	0.4464	0.4464	0.1747	0.0541	0.0478	0.04975
1000	0.25	0.4541	0.4543	0.2447	0.0919	0.0677	0.05
1000	0.05	0.4138	0.4173	0.5436	0.3913	0.1952	0.04865
10000	0.5	0.4625	0.4625	0.1773	0.0491	0.0518	0.0518
10000	0.25	0.4562	0.4563	0.2404	0.0877	0.0564	0.0492
10000	0.05	0.4573	0.4579	0.5638	0.3904	0.0845	0.05215

TABLE A.1: Simulation results for Chi-square-Chi-square distributions

		Log-normal-Log-normal					
Size	Imbalance	F Test	Bartlett	Levene	Brown-Forsythe	Normal	Permutation
50	0.5	0.5118	0.5127	0.1899	0.0414	0.0435	0.05025
50	0.25	0.4824	0.4883	0.2457	0.0762	0.0879	0.0542
50	0.05	0.2161	0.2527	0.4652	0.2584	0.3824	0.050317
100	0.5	0.5607	0.5609	0.1852	0.0439	0.035	0.05195
100	0.25	0.5437	0.5476	0.2493	0.0854	0.0817	0.05525
100	0.05	0.3567	0.3895	0.5111	0.3377	0.3677	0.05185
1000	0.5	0.6873	0.6873	0.1764	0.0455	0.033	0.0483
1000	0.25	0.6722	0.6719	0.2379	0.0887	0.0689	0.0517
1000	0.05	0.613	0.6183	0.5415	0.381	0.2555	0.046
10000	0.5	0.7361	0.7361	0.1705	0.0499	0.0418	0.0504
10000	0.25	0.7333	0.7329	0.2346	0.088	0.0594	0.05095
10000	0.05	0.7089	0.7094	0.5436	0.3816	0.1649	0.0464

TABLE A.2: Simulation results for Log-normal-Log-normal distributions

		Normal-Log-normal					
Size	Imbalance	F Test	Bartlett	Levene	Brown-Forsythe	Normal	Permutation
50	0.5	0.4586	0.4601	0.3849	0.382	0.2961	0.28815
50	0.25	0.3551	0.3488	0.3435	0.3089	0.1325	0.20435
50	0.05	0.1378	0.1091	0.3386	0.2021	0.1504	0.054533
100	0.5	0.5072	0.5076	0.4857	0.5384	0.298	0.2918
100	0.25	0.3882	0.3818	0.4205	0.4426	0.1562	0.2277
100	0.05	0.1447	0.129	0.3652	0.2861	0.1019	0.06525
1000	0.5	0.6021	0.6021	0.9596	0.9957	0.1933	0.202
1000	0.25	0.4931	0.4914	0.9555	0.9951	0.1369	0.20805
1000	0.05	0.2135	0.2106	0.86	0.9518	0.0533	0.14955
10000	0.5	0.658	0.658	1	1	0.1148	0.11975
10000	0.25	0.5482	0.5478	1	1	0.0995	0.1392
10000	0.05	0.2548	0.2539	1	1	0.0569	0.1292

TABLE A.3: Simulation results for Normal-Log-normal distributions

		Normal-Pareto					
Size	Imbalance	F Test	Bartlett	Levene	Brown-Forsythe	Normal	Permutation
50	0.5	0.4289	0.4308	0.3592	0.3522	0.2786	0.27395
50	0.25	0.3317	0.324	0.3187	0.2838	0.121	0.15475
50	0.05	0.1307	0.1056	0.3407	0.1944	0.1613	0.02495
100	0.5	0.4782	0.4789	0.4548	0.5045	0.2757	0.27705
100	0.25	0.3695	0.3637	0.4055	0.4219	0.1447	0.1877
100	0.05	0.1432	0.1285	0.38	0.2897	0.1234	0.08
1000	0.5	0.5958	0.5959	0.9506	0.9936	0.2065	0.2091
1000	0.25	0.4797	0.4792	0.9453	0.9913	0.1541	0.19135
1000	0.05	0.2132	0.2098	0.8405	0.9355	0.0562	0.13725
10000	0.5	0.6662	0.6662	1	1	0.1378	0.13955
10000	0.25	0.5671	0.5666	1	1	0.1166	0.14445
10000	0.05	0.2827	0.2813	1	1	0.0688	0.1457

TABLE A.4: Simulation results for Normal-Pareto distributions

Beyond the Headlines: Monetary Policy Transmission to Equity Markets

Ezra Polesky

Department of Economics, Occidental College

Abstract—**This paper examines the impact of monetary policy announcements on equity markets, with a particular focus on how dynamics have shifted since the rise of forward guidance. A new method is used to estimate the surprise component of monetary policy announcements. Notably, this paper finds that with the rise of forward guidance, equity markets demonstrate heightened sensitivity to surprise monetary policy announcements. Capital-intensive and cyclical industries are the most responsive segments of the market to surprise monetary policy announcements, both before and after the rise of forward guidance.**

I. Introduction

While a large body of evidence suggests that adjustments to the federal funds rate (FFR) influence equity markets, little is known about how this relationship has changed in the era of forward guidance. In April 2011, the Fed Chair began holding post-Federal Open Market Committee (FOMC) meeting press conferences, and, in December 2012, the Fed officially adopted forward guidance to its monetary policy toolkit. The purpose of these actions is to increase the transparency of the Fed's macroeconomic outlook and the expected future path of the FFR. Doing so reinforces the effectiveness of monetary policy by increasing transparency, enhancing accountability, and reducing unnecessary market volatility.[1] Recent studies prove the effectiveness of forward guidance, finding a positive relationship with equity market returns.[2] In theory, given the efficient market hypothesis, an expected rate change should be priced into equity markets prior to the Fed's announcement, and only a surprise change should cause markets to move.[3] Considering this and the purpose and proven effectiveness of the Fed increasing its transparency, surprise FFR changes should be, on average, smaller in magnitude post-2011, translating to more frequently accurate market expectations. However, with more Fed transparency, equity markets may be more confident in pricing in expectations and, therefore, more sensitive if an announcement contradicts expectations. This hypothesis raises the following question, which guides this study: what effect do FFR announcements have on equity markets since the rise of Fed transparency?

Understanding the links between monetary policy and asset prices—in this case, the FFR and equity markets—is critical to understanding the effectiveness of monetary policy transmission, thus helping inform future policy approaches. Public equities are some of the most liquid assets and, therefore, can quickly respond to monetary policy announcements. While the Fed's long-term goals are stable prices and maximum employment—not day-to-day fluctuations in the equity market—equity movement may indicate whether policy announcements are on track to achieve their intended outcomes. Equity prices reflect the present value of expected future cash flows and, accordingly, the cost of capital; when these components change, so does business strategy, which, over time, impacts inflation and employment. However, because equity markets react to countless factors, isolating the relationship between FFR announcements and equity market returns is not simple. Following the conventional approach, this study applies an event study methodology, narrowing the window of analysis to the announcement day to minimize equity market exposure to unrelated factors and the possibility of reverse causality. To distinguish between expected and surprise components of an announcement, Federal funds futures data is used, applying the method proposed by Kuttner (2001) and a novel method, labeled Polesky's Method.

The equity market's response to FFR actions is measured and analyzed in aggregate and at the level of industry portfolios. Relative to analyzing the aggregate, industry-level analysis provides a more nuanced evaluation of which parts of the market are most and least sensitive to FFR actions, which could be valuable in directing future policy approaches. Moreover, investors often diversify portfolios for industry and asset-class exposure. Therefore, industry-level results are valuable for informing investment strategy and risk management.

The findings in this study indicate that the FFR remains an effective tool in impacting equity prices, with market responses to surprise changes amplified in the presence of

Thank you to Professor Harris for his invaluable guidance, mentorship, and support on this project. Thank you to Professor Jalil for inspiring me to explore this research topic, and engaging in excellent discussion of this study and its methodology. Thank you to my peers, Francesca, Grace, and Leo, for their camaraderie and input. Thank you to the Occidental College Economics Department for engaging with and encouraging me throughout my undergraduate journey. The views expressed herein are those of the author and not necessarily those of Occidental College. The author declares that he has neither relevant nor material financial interest that relates to the research described in this paper. Any errors or omissions are the responsibility of the author. Author contact: ezrapolesky@gmail.com.

[1] Large, sudden, unexpected shifts in the FFR can cause financial instability beyond the intentions and control of the Fed. For example, households with variable-rate mortgages and firms with variable-rate loans would be severely impacted by a sudden FFR shift.

[2] For example, Neuhierl and Weber (2019), Gardner et al. (2022), and Gorodnichenko et al. (2023).

[3] For example, if one day prior to a rate change, investors expect a 25bp increase in the FFR, then equity prices should already reflect this information prior to the announcement. However, if the Fed increases the FFR by 50bp, there is a 25bp surprise component. Controlling for all else, this surprise component should be the only component of a FFR change that markets react to.

forward guidance. From 2000-2019, a surprise 25-basis-point (bp) cut in the FFR is associated with an increase in total equity market returns of about one percent. This relationship is statistically and economically significant and is consistent with previous literature. Pre-2011 observations show similar results; however, post-2011 results—specifically for FOMC meetings followed by press conferences, controlling for forward guidance—suggest that a surprise 25 bp cut in the FFR is associated with an increase in total equity market returns of 2.25%, more than twofold greater than the full sample. This study also finds that capital-intensive and cyclical industries are the most sensitive segments of the market to FFR surprises: a finding consistent with previous literature. The results are robust to asymmetries and the choice of windows for measuring equity market returns.

The structure of this paper is as follows. Section II surveys the related literature. Section III presents the data used in this research and provides a descriptive analysis of the data series. Section IV introduces the empirical strategy used for calculating the surprise and expected components of FFR announcements and a conventional model for estimating the effects of FFR announcements on equity markets. Section V presents the empirical results. Section VI concludes.

II. Literature Review

This research builds upon the current body of literature in three primary ways. First, a new method for measuring the surprise and expected components of FFR changes is proposed. Second, this paper is the first to analyze the effect of FFR announcements on equity returns in the era of forward guidance and outside of the zero lower bound (ZLB). A majority of the literature studying this topic was published pre-2011 and, therefore, pre-forward guidance. The literature post-2011 use samples pre-dating the Fed's rate hikes in 2016, hence restraining their research to the extended period of the FFR at the ZLB.[4] Third, controls for the voice tone of the Fed chair during press conferences and text sentiment of the Fed's press releases are used to separate the Fed's forward guidance from the direct effect of FFR announcements. Exploring this topic may provide evidence of 1) the effectiveness of monetary policy in the presence of forward guidance and 2) effective approaches to future monetary policy actions outside of the ZLB.

Early studies in the literature relied on raw changes to the FFR and used daily data and the event study methodology with a one-day event window on the day of FOMC meetings to analyze market responses to FFR announcements. Cook and Hahn (1989) published the seminal work on the effect of FOMC announcements on asset prices. They examine the reaction of the bond market to raw changes in the FFR from 1974-1979. They find that increases in the FFR caused large positive movements in short-term rates and smaller but still significant movements in long-term rates. Roley and Sellon (1995) apply Cook and Hahn's approach to a later period (1987-1995) and find much weaker links between changes in the FFR and changes in other market interest rates. Similar studies examining the effect of the raw change in the FFR on equity markets find mixed results. For example, Tarhan (1995) and Reinhart and Simin (1997) find no evidence that the Fed influences equity markets, whereas Thorbecke and Alami (1994) and Thorbecke (1997) find a significant negative effect from changes in the FFR. Thorbecke (1997) finds that the response of equity returns to FFR changes differs significantly across industries and that small firms' returns react much more strongly than those of large firms.

Although expectations of Fed policy announcements are not directly observable, research has attempted to disentangle the surprise and expected components of policy announcements. Krueger and Kuttner (1996) show that the Federal funds futures rate yields an efficient forecast of the FFR and, therefore, an appropriate measure of policy expectations. Since this seminal study, the related literature unanimously agrees on the importance of disentangling the surprise and expected components of policy announcements, and using the futures market to do so. Kuttner (2001) estimates the surprise component of a rate change as the change in the futures price on the day of the Fed policy action relative to the day prior to the action: a method universally applied throughout the related literature, and assessed in Section IV of this paper. Kuttner (2001) provides evidence that asset prices respond significantly to surprise. changes and not to expected changes in the FFR. This finding upholds the efficient market hypothesis, which states that asset prices should reflect all information available at any point in time, as expected changes should be priced into markets prior to the announcement. Bernanke and Kuttner (2004) find a significant response of the equity market to surprise changes in the FFR. They find that, on average, a surprise 25bp increase in the FFR is associated with about a one percent decline in broad equity indexes. They use a VAR model to explain the reaction, estimating that the largest reactions come from revisions to expectations of future excess returns and to expectations of future dividends; real interest rates have a very small direct impact. Bomfim (2003) finds that the conditional volatility of the S&P500 increases after surprise FFR changes, with positive surprises (higher-than-expected values of the FFR) tending to have a larger effect on volatility than negative surprises. Ehrmann and Fratzscher (2004) show that capital-intensive and cyclical industries (such as technology and durable goods, respectively) react two to three times stronger to monetary policy than less capital-intensive or cyclical industries (such as nondurable goods). These studies use daily data and the event study methodology with a daily response window on the day of FOMC announcements.

Several studies apply the event study method using intraday data in intervals of 5-30 minutes, enabling them to narrow the event window to an hour or less, thus further isolating the effect of FFR changes. Gürakınak et al. (2004), Zebedee et al. (2007), D'Amico and Farka (2011), and Bauer and Swanson

[4] For example, Gorodnichenko and Weber (2016), Neuhierl and Weber (2018), and Bauer and Swanson (2022).

(2022) find results largely unchanged relative to studies that use daily data.

A strand in the literature analyzes market responses to forward guidance. Kurov et al. (2021) document that pre-2011 asset returns appreciated in the 24 hours before the FOMC announcement: a measure of pre-FOMC announcement drift; however, post-2011, after the Fed began to implement forward guidance, the pre-FOMC drift disappears, indicating a reduction in uncertainty surrounding FOMC meetings and effectiveness of forward guidance. Neuhierl and Weber (2019), Gardner et al. (2022), and Gorodnichenko et al. (2023) find that, respectively, positive speeches given by the Fed Chair, text sentiment in FOMC statements, and voice tone during the Fed Chair's speeches positively affect equity returns via their effect on the market's expectations about the speed of future monetary policy loosening or tightening. When the Fed is positive and optimistic in its macroeconomic outlook and the expected future path of monetary policy, markets respond positively, and vice versa. Research has yet to analyze the effect of FFR announcements on asset prices in the era of forward guidance.

III. Data

There are four primary data sources used in this study: FOMC announcement data, Fed funds futures data, equity data, and forward guidance data. The analysis covers the period from February 2, 2000 through June 19, 2019.[5] A description of each data source is provided below.

A. Announcement Data

FOMC meeting dates are collected from the FOMC meeting calendar on the Fed's website.[6] This source is used to confirm changes to the FFR, too. The Fed began to formally announce its policy changes in February 1994 (prior to this study's sample); therefore, the announced date of a policy change is used, as opposed to the effective date. The FFR target and changes to it are collected from The Fed of New York's Effective FFR website: a daily time-series dataset expressed as percentages.

Table A1 in the Appendix lists FOMC meeting dates over the sample period, with the FFR target, changes to the FFR, and whether a press conference followed the meeting.[7] The sample period includes 162 announcements: 156 following regularly scheduled meetings and six intermeeting (unscheduled) announcements. All of the scheduled announcements occurred during market hours, after open. 29 announcements resulted in a positive change in the FFR, 18 resulted in a negative change, and 109 resulted in no change.[8] Given that the FOMC's decision to maintain (not change) the FFR can surprise markets, these dates cannot be overlooked. Of the intermeeting announcements, five resulted in negative changes and one resulted in no change. All six intermeeting announcements occurred in response to (or anticipation of) a negative shock to the macroeconomy: the dot-com bubble bursting in 2001, 9/11, and the Great Recession.

B. Fed Funds Futures Data

Daily data on the CBOT's Fed funds futures contract trades from February 2, 2000 through June 19, 2019 are acquired from Investing.com.[9] Contract prices settle based on 100 minus the relevant month's average effective FFR: the sum of all the daily effective FFRs divided by the total number of days in that month.[10] A price of $98.75 implies an effective FFR of 1.25%. This will be referred to as the implied rate. The implied rate is used to calculate the surprise and expected components of FFR changes, as explained in Section IV.A.

The time series of the implied rate and the FFR over the sample period is illustrated in Figure 1. The figure shows that 1) the implied rate closely tracks the FFR and 2) throughout the sample, monetary policy has undergone periods of tightening and easing, as well as prolonged periods of unchanged target rates, particularly at the ZLB. Dates on which the implied rate varied from the FFR by more than 25bp primarily occurred surrounding FOMC meetings, lagging the actual FFR given the nature of these contracts.[11] Dates on which the implied rate varied from the FFR by more than 50bp occurred during the rate cuts in 2008.

C. Equity Data

Daily data on equity returns from February 2, 2000 through June 19, 2019 are gathered from Kenneth French's website.[12], [13] Two datasets are used, each constructed from CRSP returns as in Fama and French (1988). Equity returns are reported as percentages.

[5] Following the approach of Gardner et al. (2022), the sample period starts in February 2000. The analysis could start in February 1994, when the FOMC first announced the outcome of a meeting and policy decision. However, the initial announcements were inconsistent and not particularly informative. In January 2000, the Committee announced that it would issue a statement following each regularly scheduled meeting, regardless of whether there had been a change in monetary policy. Following this, the first FOMC meeting and policy announcement occurred on February 2, 2000. Moreover, the available futures data dates back to January 3, 2000, preventing analysis of FOMC meetings and policy decisions pre-2000.

[6] See www.federalreserve.gov/monetarypolicy/fomccalendars.htm

[7] The information in Table A1 is confirmed using Table A12 in Gardner et al. (2022), which includes the sample period February 2, 2000 through December 16, 2020.

[8] Although one intermeeting announcement (which occurred on August 17, 2007) resulted in no change to the FFR, this meeting resulted in a 50bp cut to the Fed's discount rate, bringing it to 5.75: a response to concerns about the subprime lending crisis.

[9] Many studies (e.g. Bernanke and Kuttner (2004)) use data directly from the CBOT exchange. However, without access CBOT's data, Investing.com is the best free alternative with equivalent data. Free data prior to January 3, 2000 is not available.

[10] For example, if in a month with 30 days, the sum of effective FFRs were 37.500, then the implied rate at expiration = 37.500 / 30 = 1.250%, and the final settlement price = 100 - 1.250 = 98.750.

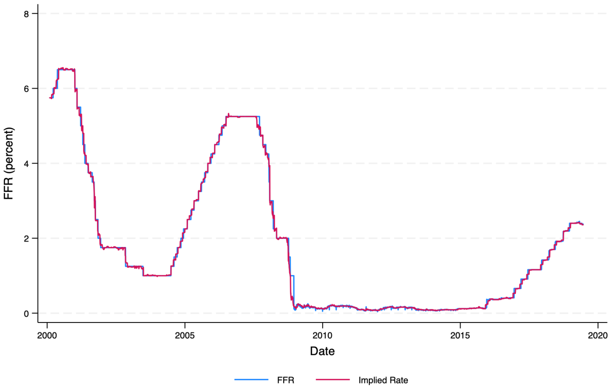
[11] The first contract to fully reflect a rate change would be the next deferred contract month, not the contract month in which the meeting takes place.

[12] Data downloaded from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/datalibrary.html Kenneth French is the Roth Family Distinguished Professor of Finance at the Tuck School of Business, Dartmouth College.

[13] Many studies (e.g. Gorodnichenko and Weber (2016)) use data from the Wharton Research Data Services CRSP database, which has minute-by-minute data on all equities and NAICS industries. However, without access the CRSP database, Fama-French data is the best free alternative. Fama-French data is sourced from CRSP, and studies analyzing industry-level data use French's website (e.g. Bernanke and Kuttner (2004)).

The first dataset contains value-weighted returns for the 10 industry portfolios. All equities listed on the NYSE, AMEX, and NASDAQ are assigned to an industry portfolio.[14]

Fig. 1: FFR and Implied Rate Over Time



Notes: The time series graph presents the implied rate (in red) imposed over the FFR target rate (in blue) over the sample period (February 2, 2000 – June 19, 2019).

The second dataset contains total market excess returns, which is merged with the industry-level dataset, allowing for a comparison of each industry's performance relative to the total market.[15]

Industry-level equity returns are used for two reasons instead of firm-level returns or a total market index. First, with its dual mandate of maximum employment and stable prices, the Fed's goal is to guide the macroeconomy as opposed to individual firms. Estimating which industries are most and least sensitive to monetary policy could be valuable in directing future policy approaches. For example, if a specific industry is driving inflation, understanding whether the FFR effectively targets the specific industry could be vital in optimizing policy approaches and mitigating potential unintended effects on the broader economy. As inflation shocks (such as those caused by global supply disruptions, shifts in energy prices, and the rise of AI) and employment shocks (such as those caused by industry-specific unionization) increasingly impact the macroeconomy, industry-level analysis becomes critical. Second, investors and portfolio managers often diversify portfolios for industry and asset-class exposure, as opposed to specific firm exposure. Therefore, industry-level results are valuable for informing investment strategy and risk management.

Value-weighted returns are used for two reasons instead of equal-weighted returns. First, equal-weighted returns weigh each constituent the same, regardless of market capitalization, whereas valued-weighted returns weigh each constituent based on market capitalization, allowing for a more nuanced analysis that mirrors the impact of monetary policy on larger firms within the market; since larger firms have a greater impact on market movements, value-weighted returns provide a more realistic representation of a given industry's response to monetary policy changes. Second, nearly all studies in the associated body of literature use value-weighted returns.

D. *Forward Guidance Data*

The text sentiment of FOMC press releases and the voice tone of the Fed Chair during post-FOMC meeting press conferences can be used to quantify forward guidance. Gardner et al. (2022) measure text sentiment on a scale from -1 to 1 (hawkish sentiment versus dovish sentiment, respectively) using large language models (LLMs). Gorodnichenko et al. (2023) measure voice tone on a scale from -1 to 1 (pessimistic tone versus optimistic tone, respectively) using LLMs. The data used in Gorodnichenko et al. (2023) are acquired from openICPSR. The dataset contains the text sentiment data used in Gardner et al. (2022), too.

Data for neither voice tone nor text sentiment span the full sample period. Voice tone data spans April 27, 2011 to June 19, 2019, covering 36 FOMC meetings (only those followed by a press conference). See Table A1 in the Appendix for a list of specific dates on which press conferences occurred. Text sentiment data spans January 26, 2011 to June 19, 2019, covering all 68 FOMC meetings over the period. Table I exhibits that the average voice tone and text sentiment over the sample period are both positive, meaning optimistic and dovish, respectively.

IV. Empirical Strategy

This section introduces a new method for calculating the surprise component of FFR changes and a conventional model for estimating the effects of FOMC announcements on equity markets. All theory and equations below rely on the efficient market hypothesis: asset prices should reflect all information available at any point in time.

A. *The Surprise Component of FFR Announcements*

For each FOMC announcement, the surprise component of the FFR change is measured using the implied rate by Fed funds futures. Investors trade the futures contracts based on what they expect the average effective FFR to be at the end of the month when the contracts settle. Krueger and Kuttner (1996) show that the futures price yields an appropriate measure for estimating expected and surprise components of FFR changes. While some surveys measure the expected component of a rate change, 1) these surveys only cover the latter portion of this study's sample period, and 2) futures data allows for the calculation of expectations on specific event days, rather than having to use stale survey expectations.

[14] Given that firms may change what industry they operate in or equities may be listed (e.g. an IPO) or delisted from an exchange, equities are assigned to an industry at the end of June of year t based on its four-digit SIC code at that time. Returns are computed from July of t to June of t+1. See French's website for industry definitions and additional information on calculating industry returns.

[15] Excess returns are calculated by subtracting the one-month Treasury bill rate (the "risk-free rate") from the value-weight return of all firms listed on the NYSE, AMEX, or NASDAQ. See Fama and French (1993) for additional information on calculating the total market excess returns.

TABLE I: Forward Guidance on Press Conference Dates

	Meetings (1)	Changes (2)	Voice Tone (3)	Text Sentiment (4)
No Change	27	0.00	0.27	0.43
		(0.00)	(0.72)	(0.20)
Positive	9	25.00	-0.45	0.10
		(0.00)	(0.57)	(0.17)
Total	36	6.25	**0.09**	**0.35**
		(10.98)	(0.75)	(0.24)

Notes: The table reports descriptive statistics for the dates over the sample period on which a press conference occurred. See Table A1 in the Appendix for a list of specific dates. The columns are separated by (1) the number of FOMC meetings, (2) the actual change of the FFR, (3) the voice tone, and (4) the text sentiment. Voice tone and text sentiment are on a scale from -1 to 1, as described in Section III.D. Mean changes are reported in basis points, followed by standard deviations in parenthesis. The rows are categorized by positive, no change, and negative to reflect the direction of change in the FFR. The totals are drawn down by meetings resulting in no change. Excluding intermeeting events has no effect on the statistics of positive changes, a marginal effect on no change, and increases the magnitude of the statistics of negative changes and, therefore, the totals; this is expected given that five of the six intermeetings resulted in a negative rate change of 50bp or more.

Differentiating between the surprise and expected components is critical to this study as expected FFR changes should be priced into financial markets prior to the announcement and, therefore, have no effect on asset prices following the announcement: a hypothesis confirmed by Kuttner (2001).[16] Hence, using raw changes to the FFR as the independent variable would impart a bias on estimates of β to the extent that the policy announcements were correctly expected by financial markets. Two separate methods are used to calculate the surprise and expected components of FFR changes.

B. Kuttner's Method

First, following [citation], the surprise component of a rate change is derived from the change in the implied rate from the market's open to close, for an event taking place on day d of month m. Given that the futures contracts settle on the monthly average effective FFR, the change in the implied rate must be scaled to account for the timing of the announcement within the month,

$$\Delta i^s = \frac{D}{D-d}(f^0_{m,d} - f^0_{m,d(open)}), \tag{1}$$

where Δi^S is the one-day surprise component of a rate change, $f^0_{m,d}$ is the current-month implied rate at the market's close, $f^0_{m,d(open)}$ is the current-month implied rate at the market's open, and D is the number of days in the month.[17] The change in the implied rate is used as opposed to the difference between the raw FFR and the implied at the market's open on day d because the implied rate is based on the effective FFR rather than the target, resulting in discrepancies between the two rates on a daily basis. [18]The expected component of a rate change is defined as the difference between the actual change and the surprise component, or

$$\Delta i^e = (\Delta i - \Delta i^s). \tag{2}$$

This method is applied universally throughout the literature in this field. [19] Because FFR announcements occur prior to the close of the futures market in every event of the sample, the closing futures price on day d should incorporate the change, making this method effective, in theory.

However, in practice, a flaw arises in Kuttner's Method. Many months within the sample period experience a surge in futures contract trading on the first few business days of the month, often followed by a reduced trade volume and little to no change in the implied rate over the remainder of the month, regardless of FOMC announcements. Derived from a one-day window, Kuttner's Method calculates expectations based on days with potentially little trade volume, in which the implied rate may not reflect revised market expectations given the new FFR. Therefore, Kuttner's Method may understate surprise components and overstate expected components for some FOMC events. The bottom of Section IV.A provides evidence supporting this claim.

The stagnation in futures trade volume may occur for a variety of reasons. First, with a fixed expiration date at the end of a month, traders may roll their positions into contracts for the following months. This rolling process could contribute to price stability within the current month, followed by an adjustment at the start of the new month when traders transition to new contracts. Second, there may be too few active traders of futures contracts to move prices on a day-to-day basis. With the implied rate reflecting the average price of all contracts issued in a given month, if the dominant investors in the market take their positions at the start of a month and do not trade throughout, then all other trades would be insignificant; moreover, there may be insufficient liquidity and infrequent trading, such that ask prices are not met by bids, and vice versa. Finally, in some instances, although the data appears in daily intervals, the price data may be weekly: possibly an error of the exchange in correctly recording price data. All free sources have the same futures data.

Regardless, the potential flaw in Kuttner's Method has yet to be addressed.

C. Polesky's Method

Unlike Kuttner's Method, Polesky's Method calculates the expected component of a rate change first and the surprise component second. For an event taking place on day d of month m, the expected component is derived from the change in the implied rate on the market's open relative to the final day of the previous month. Given that the futures contracts

[16] This assumption is also supported by the efficient market hypothesis.

[17] To minimize the effect of month-end noise in the effective FFR, the unscaled change in the implied rate is used to calculate the surprise component when the event falls on one of the last three days of the month. See Kuttner (2001) for details.

[18] This is particularly relevant beginning on December 16, 2008, when the FOMC moved from a single target rate to a target range with an upper and lower limit.

[19] For example, Bernanke and Kuttner (2004), Gorodnichenko and Weber (2016), and Gardner et al. (2022).

settle on the monthly average effective FFR, the change in the implied rate must be scaled to account for the timing of the announcement within the month,

$$\Delta i^e = \frac{D}{D-d}(f^0_{m,d(open)} - f^1_{m-1,D}), \tag{3}$$

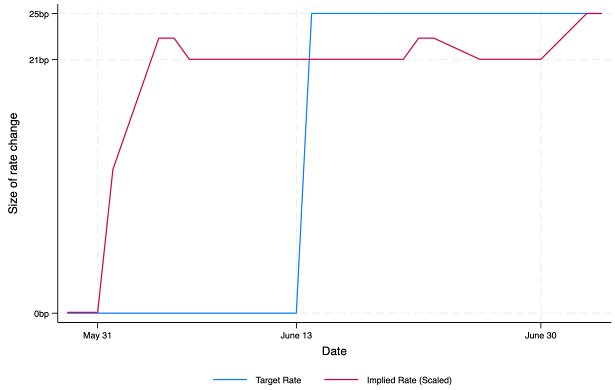
where Δi^e is the expected component of a rate change, $f^0_{m,d(open)}$ is the current-month implied rate at the market's open, D is the number of days in the month, and $f^1_{m-1,D}$ is the implied rate on the final day of the previous month.[20] The change in the implied rate is used as opposed to the difference between the raw FFR and the implied at the market's open on day d for the same reason as mentioned above, regarding equation (1). The surprise component of a rate change is defined as the difference between the actual change and the expected component, or

$$\Delta i^s = (\Delta i - \Delta i^e). \tag{4}$$

Derived from a wider window, and thus greater trade volume, Polesky's Method accounts for changes in expectations on the days and weeks leading up to FOMC meetings; this makes it robust to event days with potentially little trade volume, in which the implied rate may not immediately reflect revised market expectations given the new FFR. Kuttner's Method is susceptible to such occurrences. Figure 2 exemplifies such an instance, contrasting the two methods for a specific event month. The scaled implied rate remains unchanged for around one week on either side of the event. Accordingly, Kuttner's Method calculates an expected component of 25bp and a surprise component of 0bp.[21] Polesky's Method calculates what appears to be a more accurate estimation of expectations: an expected component of 21bp and a surprise component of 4bp.[22]

More comprehensive analysis affirms Polesky's Method as more robust than Kuttner's Method. For a majority of the events, Kuttner's Method calculates a surprise component of near zero, implying that markets accurately expect a majority rate decisions. Table II reflects this, reporting that, on average, Kuttner's Method calculates surprise components closer to zero than Polesky's Method; importantly, with a standard deviation of around 3.5bp, the distribution of Kuttner's Method is nearly three times smaller than Polesky's Method. This supports the claim that Kuttner's Method may understate surprise components and overstate expected components for some FOMC events. Figures A3-A6 in the Appendix further illustrate the differences between Kuttner's Method and Polesky's Method, and that Kuttner's Method clumps much of the surprise components around 0bp. Given the evidence in favor of Polesky's Method, it is the primary approach for calculating the expected and surprise components of FFR announcements throughout the remainder of the paper unless otherwise noted.

Fig. 2: Target Rate and Scaled Implied Rate Over Time (Event on June 13, 2018)



Notes: The time series graph exhibits the scaled implied rate (in red) imposed over the target rate (in blue) from May 29, 2018 through July 3, 2018. The implied rate is scaled by the timing of the FOMC announcement within the month, using the scale factor from equations (1) and (3). On June 13, 2018, the FOMC held a scheduled meeting in which it increased the FFR target by 25bp. Markets were closed for the weekend of July 1 and 2, explaining the gradual increase in the implied rate from June 30 to July 3.

TABLE II: Forward Guidance on Press Conference Dates

	(1)	(2)	(3) Surprise		(4) Expected	
	Meetings	Change	Kuttner	Polesky	Kuttner	Polesky
Negative	17	-43.06	-4.07	-7.21	-38.99	-35.85
		(16.73)	(8.50)	(20.84)	(12.59)	(18.94)
No Change	109	0.00	0.17	-0.36	-0.17	0.36
		(0.00)	(1.62)	(5.80)	(1.62)	(5.80)
Positive	29	25.86	0.22	1.22	25.65	24.65
		(4.64)	(2.19)	(11.57)	(4.40)	(12.08)
Total	156	-0.16	**-0.31**	**-0.86**	0.15	0.70
		(19.36)	**(3.53)**	**(10.04)**	(17.95)	(18.75)

Notes: The table reports descriptive statistics for the scheduled FOMC announcement dates over the sample period. The columns are separated by (1) the number of FOMC meetings, (2) the actual change of the FFR, (3) the surprise component of Kuttner's Method versus Polesky's Method (using equations 1 and 4, respectively), and (4) the expected component of Kuttner's Method versus Polesky's Method (using equations 2 and 3, respectively). Mean changes are reported in basis points, followed by standard deviations in parenthesis. The rows are categorized by positive, no change, and negative to reflect the direction of change in the FFR. The totals are drawn down by meetings resulting in no change. Excluding intermeeting events has no effect on the statistics of positive changes, a marginal effect on no change, and increases the magnitude of the statistics of negative changes and, therefore, the totals; this is expected given that five of the six intermeetings resulted in a negative rate change of 50bp or more.

D. Baseline Model: Event Study Approach

The baseline regression model for this study is in the form of a multivariate Ordinary Least Squares (OLS) regression

[20] For months with more than one FOMC meeting (specifically, months with an intermeeting), the first event of the month uses equation (3), and the subsequent use $f^0_{m,d-(n-1)}$ instead of $f^1_{m-1,D}$. n indicates the number of days since the previous event. Therefore, $f^0_{m,d-(n-1)}$ indicates the implied rate on the day after the previous event.

[21] $\Delta i^s = \frac{30}{30-13}(1.82 - 1.82) = 0$ and $\Delta i^e = \Delta i - \Delta i^s = 0.25 - 0 = 0.25$.

[22] $\Delta i^s = \frac{30}{30-13}(1.82 - 1.70) = 0.21$ and $\Delta i^e = \Delta i - \Delta i^s = 0.25 - 0.21 = 0.04$.

estimator, as it is the most efficient linear regression estimator when the assumptions hold true. Methods such as regression discontinuity and difference-in-differences are unnecessary given the random walk theory.[23] The approach of this paper involves analyzing market reactions to specific events, categorizing it as an event study. As discussed in Section III.A, the sample contains 156 events (FFR announcements), excluding the six intermeeting events. Following the conventional approach, narrowing the window of analysis to the day of the event minimizes equity market exposure to unrelated factors and accounts for potential issues of reverse causality and endogeneity, supporting the claim for causality.[24] Data on voice tone and text sentiment (as measures of forward guidance) do not yet exist for FOMC meetings after June 19, 2019. Given the positive relationship between forward guidance and equity returns, including such meetings without controls for forward guidance could impose a substantial positive bias on the results.[25] Markets may be responding to the Fed's macroeconomic outlook and transparency on the future path of monetary policy, as opposed to exclusively the surprise component of the FFR announcement. Hence, to further isolate the relationship between FFR announcements and equity returns, FOMC meetings after June 19, 2019 are excluded from this analysis.

For an event taking place on day d, the baseline regression model follows:

$$R_{i,d} = \beta_1^i \Delta i_d^e + \beta_2^i \Delta i_d^s + \beta_3^i Voice_d + \beta_4^i Text_d + \epsilon_{i,d}, \quad (5)$$

where $R_{i,d}$ represents the equity return for industry i and Δi_d^e and Δi_d^s represent the expected and surprise components of a FFR change, using the decomposition described in Section IV.A. $Voice_d$ and $Text_d$ represent controls for the voice tone of the Fed Chair during post-FOMC meeting press conferences and the text sentiment of FOMC press releases, respectively. The average voice tone and text sentiment from the sample are applied to all FOMC meetings without such data. Doing so allows regressions to cover the full sample period, while controlling for the meetings with voice tone and text sentiment data, and having no additional effects on regression results. $\epsilon_{i,d}$ denotes the stochastic error term, which represents factors other than FFR changes that affect equity returns on event days. These factors are assumed to be orthogonal to FFR changes. Section IV.C discusses the validity of this assumption.

The regression is run for each industry, estimating industry-specific coefficients. Doing so reveals whether there are industry-specific effects of FFR announcements: the purpose of this study. Estimating one coefficient for all industries raises concerns of heteroskedasticity. For example, cyclical industries and capital-intensive industries (such as durables and technology, respectively) should be more sensitive to FFR announcements than those that are not (such as nondurables). The findings of Ehrmann and Fratzscher (2004) and Thorbecke (1997) support this claim. Applying industry-specific fixed effects is unnecessary as the standard errors of high-volatility industries would be dampened by low-volatility industries, and vice versa, which could lead to inaccurate estimations of statistical significance.

Intermeeting announcements are excluded from the core analysis and included only in alternative models. Faust et al. (2004) find that intermeeting policy decisions may reflect new information about the state of the economy, and hence the equity market reacts to this new information rather than changes in monetary policy. Moreover, while the market may expect a change in the FFR, the timing of intermeeting announcements is a surprise. Gorodnichenko and Weber (2016) and Neuhierl and Weber (2018) exclude intermeeting announcements from their analyses. Zebedee et al. (2007) find that, relative to scheduled FOMC meetings, markets are highly sensitive to intermeeting FOMC announcements. This finding is supported by Figures A3 and A4 in the Appendix, in which the most extreme observations occur on intermeeting dates.

The causal relationship targeted through this model assumes that FFR surprises move equity markets by impacting expected future cash flows and demand for equities. Markets may perceive an increase in the FFR as an indicator of the Fed attempting to slow the macroeconomy to its long-run steady state and decrease the money supply, potentially leading to a policy-induced recession. First, the decline in consumer and business confidence associated with such an announcement may lead to a decline in spending and investment. Second, higher interest rates increase the cost of capital, reducing investment spending and increasing expenses. Both of these lead to a decline in expected future growth and sales, and therefore expected profit and dividends, resulting in a decline in equity prices, assuming that an equity price equals the present value of expected future cash flows discounted over time. Moreover, as risk-free returns increase, the risk-adjusted returns from investing in equities decrease, lowering the fundamental value of the given equity. Higher interest rates also increase bond yields, making them an appealing alternative investment to equities, lowering the demand for and, therefore, the price of equities. In short, a surprise increase in the FFR should lead to a decrease in equity prices, and vice versa. Figure A4 in the Appendix and the findings of Bernanke and Kuttner (2004) support this hypothesis.

E. Orthogonality Revisited

The integrity of the event study approach depends on the assumption that the error term is orthogonal to FFR announcements, and therefore the expected and surprise components. One violation of this condition would occur if monetary policy and equity markets simultaneously respond to a third factor. For example, prior to 1994, the FOMC often adjusted the

[23] The random walk theory states that movements in equity markets are unpredictable and lack any pattern.

[24] For example, unrelated factors affecting equity markets include earnings reports and forecasts, inflation and employment data releases, political news, and world events.

[25] Previous research supports this hypothesis. For example, see Neuhierl and Weber (2019), Gardner et al. (2022), and Gorodnichenko et al. (2023).

FFR hours after (and in response to) the Bureau of Labor Statistics' employment report release, which simultaneously moved equity markets. An employment release indicating a weaker-than-expected job market (or any other negative news release) would likely cause a negative shock to equity markets, and simultaneously lead to a cut in the FFR, and vice versa. However, such an event—in which a rate cut was associated with an equity market decline—would cause a downward bias on or flip the sign of β_2. Moreover, narrowing the event window to one day and excluding intermeeting announcements provides strong preventative measures for such simultaneity.[26]

A second violation of the orthogonality condition would be a response in monetary policy to the movement of equity markets: reverse causality. By narrowing the event window to one day, the risk of reverse causality is minimized. A violation would only occur if the FOMC reacted to same-day market movements, which there are no clear instances of. Even monthly data offers doubtful evidence for such a relationship.[27] If the FOMC were to respond to movements in equity markets, it would likely cause a downward bias on β_2, therefore understating the relationship between FFR surprises and equity markets. Scheduled meetings are announced one year in advance, on approximately the same days every year, unrelated to equity market conditions.

Previous studies propose two solutions to address issues of simultaneity and endogeneity. The first solution uses intraday data in intervals of 5-15 minutes surrounding FOMC announcements, narrowing the event window to an hour or less; doing so isolates the effect of FOMC announcements from the impact of news released earlier or later in the day.

Gürakınak et al. (2004) and Zebedee et al. (2007) use high-frequency intraday data and find results largely unchanged relative to studies that use daily data, aside from improvements in R^2.

The second solution involves a more statistical approach. Bauer and Swanson (2022), who use high-frequency intraday data, orthogonalize FFR surprises with respect to macroeconomic and financial data that pre-date the announcement. Rigobon and Sack (2002) use an estimator based on the heteroskedasticity that exists in high-frequency data. Their estimates of the effects of FOMC announcements on asset prices are largely unchanged from those of the conventional event studies.

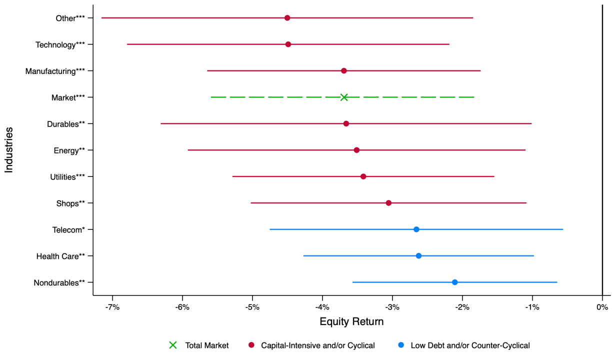
A third violation of the orthogonality condition would be if the surprise component contains any information effects.[28] Information effects would be present if the FOMC's announcement revealed new information about the macroeconomic outlook that directly impacted market expectations and equity prices, separate from the actual expectations of a rate change. For equity markets, information effects would likely counteract the effect of the surprise component. For example, a more positive outlook of the macroeconomy from the Fed than the public could lead to a surprise rate hike: . As discussed above, a surprise increase in the FFR hypothetically should decrease equity prices. However, such an information effect could boost macroeconomic growth forecasts, thus increasing equity prices in anticipation of greater-than-expected future cash flows, and vice versa. Therefore, information effects could cause a downward bias on or flip the sign of β_2. However, Faust et al. (2004) and Bauer and Swanson (2022) find that the responses of equity markets, macroeconomic surveys, and exchange rates to scheduled FOMC announcements show little evidence of information effects. Moreover, controlling for forward guidance and excluding meetings after June 19, 2019, for which forward guidance data does not yet exist, provides strong preventative measures against potential information effects.

To summarize, the alternative methods used to account for potential biases in measuring the effects of FFR surprises offer comparable results to those relying on the event study approach discussed in Section IV.B. Even if the event study results are biased, the bias likely understates the actual response of equity markets to FFR announcements. Hence, the baseline model appears to be the most effective method for this study, noting that it may marginally understate the effects of FFR announcements on equity markets.

V. Empirical Results

Table III presents the results for equation (5) estimated using daily returns data on the total market and industry portfolios. Focusing on the full sample (column 1), a surprise increase in the FFR results in a drop in equity prices for the total market and all industry portfolios. The results indicate that a surprise one percentage point tightening in FFR is associated with a decline of 3.69% in total equity market returns. Translated to traditional policy moves, a surprise 25bp increase in the FFR results in an average decline in equity returns of 0.92%. These findings are consistent with previous literature (e.g. Bernanke and Kuttner (2004)).

Fig. 3: Full Sample—Industry Effects of FFR Surprises



Notes: The figure visualizes the correlation coefficients and standard errors—represented by the dots and whiskers, respectively—from column 1 in Table III. See notes under Table III for additional details.

[26] The intermeeting event on March 16, 2020 exemplifies such a joint response: the Fed's 100bp rate cut and the equity market's steep drop were both prompted by the Covid-19 lockdowns.

[27] For example, Bernanke and Gertler (1999) and Fuhrer and Tootell (2008). Rigobon and Sack (2002) find contradictory evidence.

[28] Romer and Romer (2000) find evidence in support of information effects.

TABLE III: Polesky's Method—Industry Effects of FFR Surprises

	(1) Full Sample	(2) pre-2011	(3) post-2011	(4) PC: No Controls	(5) PC: Controls
Other	-4.50***	-4.10**	-6.75	-6.40	-8.04*
	(1.34)	(1.55)	(3.98)	(3.04)	(3.77)
Technology	-4.49***	-4.32**	-5.39	-7.66*	-8.73*
	(1.17)	(1.39)	(3.24)	(3.04)	(3.79)
Manufacturing	-3.69***	-3.29**	-6.13	-6.88*	-9.86**
	(0.99)	(1.03)	(3.46)	(3.01)	(3.66)
Market	**-3.69*****	**-3.41****	**-5.61**	**-6.54***	**-9.00****
	(0.96)	**(1.06)**	**(3.17)**	**(2.73)**	**(3.32)**
Durables	-3.66**	-3.28*	-6.04	-4.95	-9.46*
	(1.34)	(1.43)	(4.56)	(3.69)	(4.40)
Energy	-3.51**	-2.83*	-10.54*	-12.17*	-18.75**
	(1.22)	(1.17)	(4.70)	(4.78)	(5.63)
Utilities	-3.42***	-2.95**	-8.09*	-7.29	-13.54**
	(0.95)	(0.93)	(3.59)	(4.04)	(4.66)
Shops	-3.05**	-2.84*	-3.74	-4.73	-6.50
	(1.00)	(1.13)	(3.10)	(2.79)	(3.45)
Telecom	-2.66*	-2.69*	-2.16	-1.44	-4.09
	(1.06)	(1.24)	(3.15)	(2.65)	(3.17)
Health Care	-2.63**	-2.55**	-2.48	-5.67	-8.65*
	(0.83)	(0.85)	(3.07)	(2.81)	(3.39)
Nondurables	-2.11**	-1.84*	-3.85	-4.32	-6.84
	(0.74)	(0.73)	(2.81)	(3.02)	(3.70)
Observations	156	90	66	36	36
Avg. Adj. R-Square	0.05	0.06	-0.01	0.05	0.06

Standard errors in parentheses

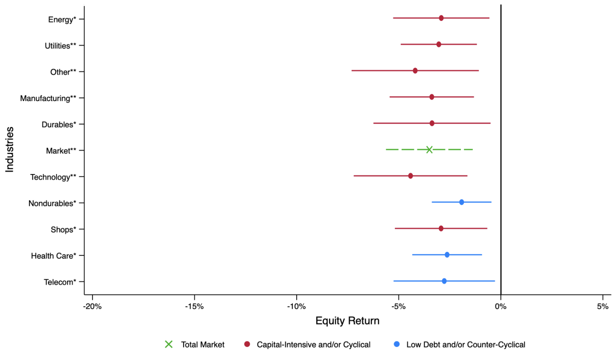
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The table reports the results from regressions of one-day returns of the total market and Fama-French industry portfolios (indicated in the row headings) on one-day FFR surprises, all expressed in percentage terms. The regressions include an intercept, the expected component of the one-day change in the FFR, and forward guidance controls, whose coefficients are not reported. The full sample includes 156 FOMC meetings spanning February 2, 2000 to June 19, 2019. The subsample used in column 2 includes 90 FOMC meetings spanning February 2, 2000 to March 15, 2011. The subsample used in column 3 includes 66 FOMC meetings spanning April 27, 2011 to June 19, 2019. The subsamples used in columns 4 and 5 include the 36 FOMC meetings followed by a press conference, which are denoted by the abbreviation PC. See Table A1 in the Appendix for a list of specific dates on which press conferences occurred. All samples exclude intermeeting events.

Capital-intensive and cyclical industries (such as technology and durables, respectively) are more sensitive to FFR surprises than industries that are not: a finding consistent with previous literature (e.g. Ehrmann and Fratzscher (2004)). Figure 3 provides a visual support, reflecting this dynamic for the full sample results. Theoretically, this finding makes sense; a surprise increase in the FFR increases the cost of capital and decreases expectations about consumer demand and future cash flows, which particularly harms capital-intensive and cyclical industries. The most sensitive industry to FFR surprises is other, which contains firms operating mostly in finance, hotels, and entertainment. Technology is a close second. Ehrmann and Fratzscher (2004) estimate near-zero coefficients on energy and utilities, and find them among the only statistically insignificant industries: contradictory to the findings of this study. In defense of Ehrmann and Fratzscher (2004)), 1) energy performance tracks to the price of oil and gas, which is not necessarily linked to the FFR and cost of capital 2) utilities are largely rate regulated and a consumer staple, and, thus, generate extremely consistent revenue, regardless of the stage in business and credit cycles. Arguably, these factors could shield energy and utilities equities from the effects of surprise FFR announcements. However, both industries are highly capital intensive, and energy quite cyclical, suggesting that a significant relationship with FFR surprises should be present. Thus, this particular finding likely contradicts Ehrmann and Fratzscher (2004)) due to alternative methodologies.

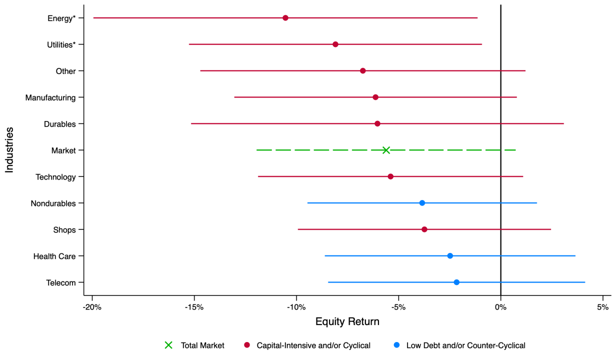
Columns 2 and 3—pre- and post-2011—reveal contrasting effects under different Fed policy regimes. Figures 4 and 5 illustrate these results. Post-2011 relative to pre-2011, the relationship between FFR announcements and equity returns substantially increases in magnitude for the total market and all industries, excluding telecommunications and healthcare, which are both historically non-cyclical and relatively low debt, and among the least sensitive to FFR surprises. Moreover, based on the standard errors, the total market and all industries are far more volatile in response to FFR surprises post-2011. These findings indicate that, with the rise of Fed

Fig. 4: Pre-2011 Sample—Industry Effects of FFR Surprises



Notes: The figure visualizes the correlation coefficients and standard errors—represented by the dots and whiskers, respectively—from column 2 in Table III. See notes under Table III for additional details.

Fig. 5: Post-2011 Sample—Industry Effects of FFR Surprises



Notes: The figure visualizes the correlation coefficients and standard errors—represented by the dots and whiskers, respectively—from column 3 in Table III. See notes under Table III for additional details.

transparency, equity markets demonstrate heightened sensitivity to surprise announcements. Economic theory supports this finding. Pre-2011, in the absence of forward guidance, markets had less concrete information to form expectations on rate changes and, therefore, were likely more conservative in pricing in such expectations. Post-2011, with the rise of Fed transparency, markets likely became more confident in pricing in expectations and, therefore, more sensitive if an announcement was contradictory. For example, if the Fed communicates tentative plans to cut the FFR by 25bp at the next FOMC meeting, markets would confidently price in expectations for such a rate change; however, when the meeting occurs, if the Fed decides against cutting the FFR, the surprise component and, accordingly, equity returns may be greater in magnitude than had the Fed not communicated its tentative plan. In short, equity markets are less sensitive to FFR surprises without Fed transparency and more sensitive with Fed transparency.

Columns 4 and 5 suggest that equity markets are substantially more sensitive to surprise FFR changes on days with post-FOMC meeting press conferences than on days without. The FOMC's press conferences are pre-scheduled and not necessarily associated with major policy announcements, therefore eliminating such a claim as a potential explanation for this finding.[29] Hence, two potential explanations justify this finding. First, given that FOMC press conferences attract significant media attention and are closely watched by market participants, trading activity may be higher on such days, potentially amplifying market movements. This claim is indirectly supported by 1) Savor and Wilson (2013), who find that over 60% of annual equity returns are earned on FOMC event days, and 2) Lucca and Moench (2015), who find that about 80% of annual equity returns are accounted for in the 24 hours before the FOMC announcement, leading up to the 15 minutes before the announcement. Second, many of the FOMC meetings during this subsample occurred while the FFR was constrained by ZLB, resulting in many surprise components near zero, thus amplifying the relationship with equity returns.

The results in column 5 are greater in magnitude than in column 4, suggesting a positive information effect from forward guidance and a negative FFR effect. Column 4 excludes controls for Fed Chair voice tone and FOMC text sentiment, whereas column 5 controls for such factors. The average voice tone and text sentiment over the sample period are both positive, meaning optimistic and dovish, respectively (see Table I). Hence, omitting these variables imposes a positive bias on the results of column 4. Moreover, when including the controls, the coefficients for the total market and nearly all industries increase by a considerably larger magnitude than the standard errors, suggesting that the controls explain some of the variance in the equity returns that was previously unaccounted for. The model's adjusted R^2 increases, too, indicating that the controls improve the model and are important for understanding the relationship between equity returns and FFR surprises. The most sensitive industry in both columns is energy: a surprise 25bp increase in the FFR results in an average decline in equity prices of 4.69%, when controlling for voice tone and text sentiment. This relationship is more than sixfold greater in magnitude than pre-2011.

For comparison to Polesky's Method, Table A2 in the Appendix reports the results from identical regressions to those of Table III, except that Kuttner's Method is used to calculate the surprise and expected components of FFR announcements. The results indicate that pre-2011, a surprise 25bp increase in the FFR is associated with an average decline in equity returns of more than 3%. This finding is nearly threefold greater than previous findings. Kuttner's Method indicates that post-2011, the relationship between FFR surprises and total market equity returns disappears, with the coefficient on the total market narrowing in on zero and becoming statistically insignificant.

To investigate concerns of asymmetries potentially biasing the baseline results, interactive dummy variables are applied to the regression. Broadly defined, asymmetries include the

[29] Only 9 of the 36 FOMC meetings followed by a press conference resulted in a rate change. See Table A1 in the Appendix for a list of specific dates on which press conferences occurred.

possibility that the equity price response to monetary policy depends on the context in which a FFR announcement occurred. The baseline results are robust to asymmetries, as reported in Table A3 in the Appendix. The statistically insignificant coefficients on the "reversal," "no change," and "intermeeting" interaction variables suggest that reversal dates, inaction by the Fed, and intermeetings, respectively, are not important determinants of the market's reaction. However, the coefficient on surprise increases in magnitude when controlling for "no change," indicating that when a change does occur, market responses are greater than when no change occurs. The positive and statistically significant coefficient on the pre-2011 "SEP" (Summary of Economic Projections) interaction variable indicates that the market favored the release of SEPs; excluding this from the full sample of the baseline results has a minor downward bias on the results. "Dummying out" these observations confirms that the baseline results are not dependent on such events.

In the last column in Table A3, the positive and statistically significant coefficient on the post-2019 interaction variable supports the theory behind excluding the post-2019 period from the full sample; it indicates that, with no controls for forward guidance, a material positive bias is present in the period.

The results are robust to placebo tests: specifically, to the choice of windows for measuring equity market returns. Table A4 in the Appendix reports the results from identical regressions to those of Table III, except that rather than use equity returns from the same day as FOMC events, returns from seven business days prior to the event are used. The results show no significant relationship between FFR surprises and equity returns seven days prior to the event; however, the standard errors are comparable between Tables 3 and A4, indicating similar levels of dispersion and volatility on event days versus non-event days. Unlike the baseline results, the coefficients in Table A4 appear relatively random in magnitude across industries and regressions, no longer suggesting capital-intensive and cyclical industries as more responsive. The dynamic seen in the baseline results between pre- and post-2011 (columns 2 and 3) and press conferences with and without controls (columns 4 and 5) effectively disappears with the placebo test. The adjusted R2 decreases for each model except post-2011, indicating that the placebo test is a worse fit for modeling the relationship between FFR surprises and equity returns. The placebo results hold for equity returns used from 4-10 business days prior to FOMC events.[30] Theoretically, all of this makes sense, as markets should not respond to FFR surprises that have not yet occurred. The placebo tests provide further evidence supporting the baseline results as valid.

VI. Conclusion

The findings in this study indicate that equity prices respond inversely to FFR surprises. The relationship is statistically and economically significant. The regressions show that 1) on average, a surprise 25bp increase in the FFR results in about a one percent decrease in total equity market returns and 2) capital-intensive and cyclical industries are more sensitive to FFR surprises than industries that are not. These findings are consistent with previous literature. Returning to the question guiding this research: what effect do FFR announcements have on equity markets since the rise of Fed transparency? Post-2011, since the rise of Fed transparency, equity markets demonstrate heightened sensitivity to FFR surprises. The results are robust to asymmetries and the choice of windows for measuring equity market returns.

These findings present several policy implications. First, and most obvious, FFR announcements remain an effective tool in impacting equity prices. This demonstrates the continued credibility of the Fed and its ability to influence expectations of consumer demand, future cash flows, and the cost of capital. Second, market responses to FFR announcements are amplified in the presence of Fed transparency, evidencing the effectiveness of Fed transparency as a tool for influencing market expectations. This policy implication may be particularly useful when the FFR is constrained by the ZLB, as transparency provides the Fed further control over longer-term rates and, more broadly, perceptions of the future state of the economy. Simultaneously, the Fed must consider the potential ramifications of excessive transparency: heightened sensitivity to FFR surprises could introduce consequential market instability, and even distrust in the Fed if announcements frequently or materially contradict previous guidance. Third, if the U.S. is prioritizing advancements in and investment into renewable energy and AI, leaning into unconventional monetary policy tools and fiscal policy may be more strategic than increasing the FFR when the economy is overheating. When the Fed increases the FFR, industries such as energy and technology—those driving innovations in renewable energy and artificial intelligence—suffer the most. That said, the Fed must remain independent of the government and make decisions accordingly, prioritizing its dual mandate.

This study expands on the traditional analysis of monetary policy announcements by proposing and applying a new method for estimating the expected and surprise components of FFR announcements: Polesky's Method. It further builds upon the literature by analyzing the effectiveness of FFR announcements since the rise of Fed transparency. Moreover, this study is the first in the field to use data on the voice tone of the Fed chair during press conferences and the text sentiment of the Fed's press releases to separate the Fed's transparency from the direct effect of FFR announcements.

While this study offers meaningful contributions to the literature, it also opens the door for future research. As surveys measuring the expected component of FFR announcements become increasingly prevalent, comparing Polesky's Method to such survey results may provide further evidence supporting Polesky's Method as effective. The period from 2019-2024 sustained considerable adjustments to the FFR—both easing and tightening—which could provide a valuable extension to

[30] Equity returns more than 10 days and fewer than four days prior to FOMC announcements were not checked as additional placebo tests.

this research. However, this period should be studied only once data on voice tone and text sentiment is available. This study assumes that markets price in surprise FFR announcements on the same day of the event. Though, it is possible that markets continue to process and price in information in the days following the event, indicating that 1) markets may initially exhibit herd behavior, irrationally overreacting to the surprise announcement, and subsequently settling back to normal, pre-surprise levels or 2) post-FOMC announcement drift exists. Research evaluating a wider window of equity returns post-announcement may establish whether initial market reactions are rational, appropriate in size, and timely. Finally, an analysis of the link between monetary policy and a wider class of assets presents an intriguing topic for future research, and could further develop the conclusions of this study.

References

Bauer, M. D. and Swanson, E. T. (2022). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, 37(1):87–155.

Bernanke, B. and Kuttner, K. (2004). What explains the stock market's reaction to federal reserve policy? Technical Report 10402, NBER Working Paper Series.

Bernanke, B. S. and Gertler, M. (1999). Monetary policy and asset price volatility. *Economic Review*, 84(Q IV):17–51.

Bomfim, A. N. (2003). Pre-announcement effects, news, and volatility: Monetary policy and the stock market. *Journal of Banking and Finance*, 27(1):133–151.

Cook, T. and Hahn, T. (1989). The effect of changes in the federal funds rate target on market interest rates in the 1970s. *Journal of Monetary Economics*, 24(3):321–351.

D'Amico, S. and Farka, M. (2011). The fed and the stock market: An identification based on intraday futures data. *Journal of Business and Economic Statistics*, 29(1):126–137.

Ehrmann, M. and Fratzscher, M. (2004). Taking stock: Monetary policy transmission to equity markets. *Journal of Money, Credit, and Banking*, 36(4):719–737.

Fama, E. F. and French, K. R. (1988). Permanent and temporary components of stock prices. *Journal of Political Economy*, 96(2):246–273.

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

Faust, J., Swanson, E. T., and Wright, J. H. (2004). Do federal reserve policy surprises reveal superior information about the economy? *Contributions in Macroeconomics*, 4(1).

Fuhrer, J. C. and Tootell, G. M. B. (2008). Eyes on the prize: How did the fed respond to the stock market? *Journal of Monetary Economics*, 55(Q IV):796–805.

Gardner, B., Scotti, C., and Vega, C. (2022). Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements. *Journal of Econometrics*, 231(2):387–409.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The voice of monetary policy. *American Economic Review*, 113(2):548–584.

Gorodnichenko, Y. and Weber, M. (2016). Are sticky prices costly? evidence from the stock market. *American Economic Review*, 106(1):165–199.

Gürakınak, R. S., Sack, B., and Swanson, E. T. (2004). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. Technical Report 2004-66, Finance and Economics Discussion Series.

Krueger, J. T. and Kuttner, K. N. (1996). The fed funds futures rate as a predictor of federal reserve policy. *Journal of Futures Markets*, 16(8):865–879.

Kurov, A., Wolfe, M. H., and Gilbert, T. (2021). The disappearing pre-fomc announcement drift. *Finance Research Letters*, 40.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of Monetary Economics*, 47(3):523–544.

Lucca, D. O. and Moench, E. (2015). The pre-fomc announcement drift. *The Journal of Finance*, 70(1):329–371.

Neuhierl, A. and Weber, M. (2018). Monetary momentum. Technical Report w24748, National Bureau of Economic Research.

Neuhierl, A. and Weber, M. (2019). Monetary policy communication, policy slope, and the stock market. *Journal of Monetary Economics*, 108:140–155.

Reinhart, V. and Simin, T. (1997). The market reaction to federal reserve policy action from 1989 to 1992. *Journal of Economics and Business*, 49(2):149–168.

Rigobon, R. and Sack, B. (2002). The impact of monetary policy on asset prices. *Journal of Monetary Economics*, 51(8):1553–1575.

Roley, V. V. and Sellon, G. H. (1995). Monetary policy actions and long-term interest rates. *Federal Reserve Bank of Kansas City Economic Quarterly*, 80(4):77–89.

Romer, C. D. and Romer, D. H. (2000). Federal reserve information and the behavior of interest rates. *American Economic Review*, 90(3):429–457.

Savor, P. and Wilson, M. (2013). How much do investors care about macroeconomic risk? evidence from scheduled economic announcements. *Journal of Financial and Quantitative Analysis*, 48(2):343–375.

Tarhan, V. (1995). Does the federal reserve affect asset prices? *Journal of Economic Dynamics and Control*, 19(5-7):1199–1222.

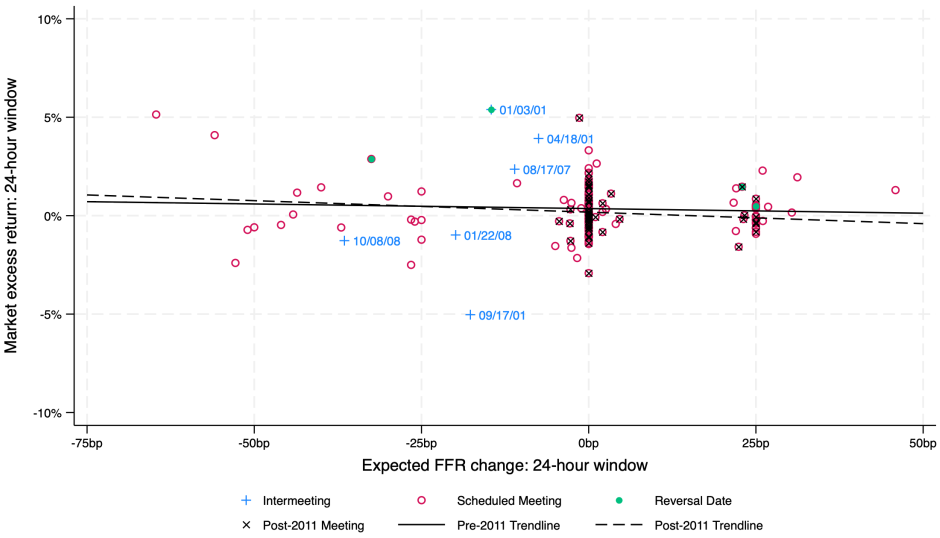
Thorbecke, W. (1997). On stock market returns and monetary policy. *The Journal of Finance*, 52(2):635–654.

Thorbecke, W. and Alami, T. (1994). The effect of changes in the federal funds rate target on stock prices in the 1970s. *Journal of Economics and Business*, 46(1):13–19.

Zebedee, A. A., Bentzen, E., Hansen, P. R., and Lunde, A. (2007). The greenspan years: An analysis of the magnitude and speed of the equity market response to fomc announcements. *Financial Markets and Portfolio Management*, 22(1):3–20.

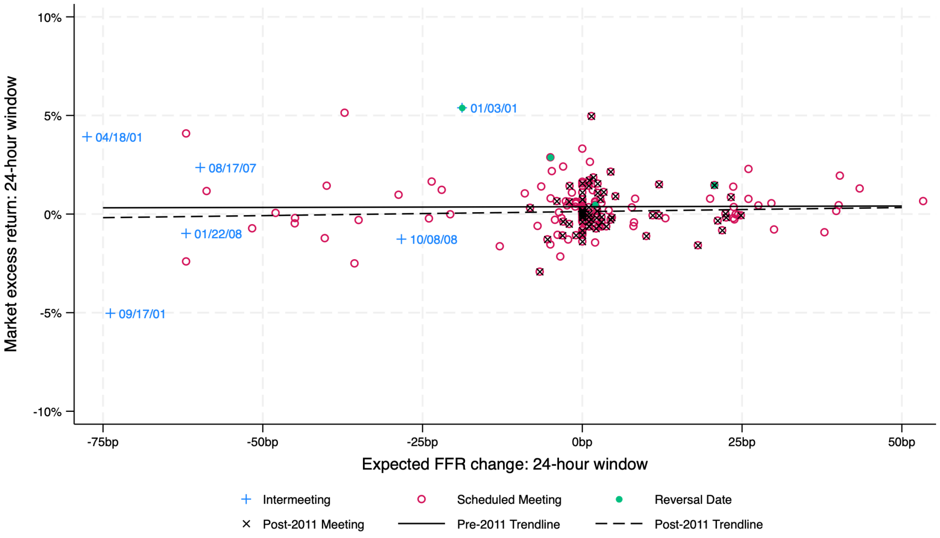
APPENDIX A: FIGURES

Fig. A.1: Kuttner's Method—Market Excess Returns vs. Expected FFR Announcements



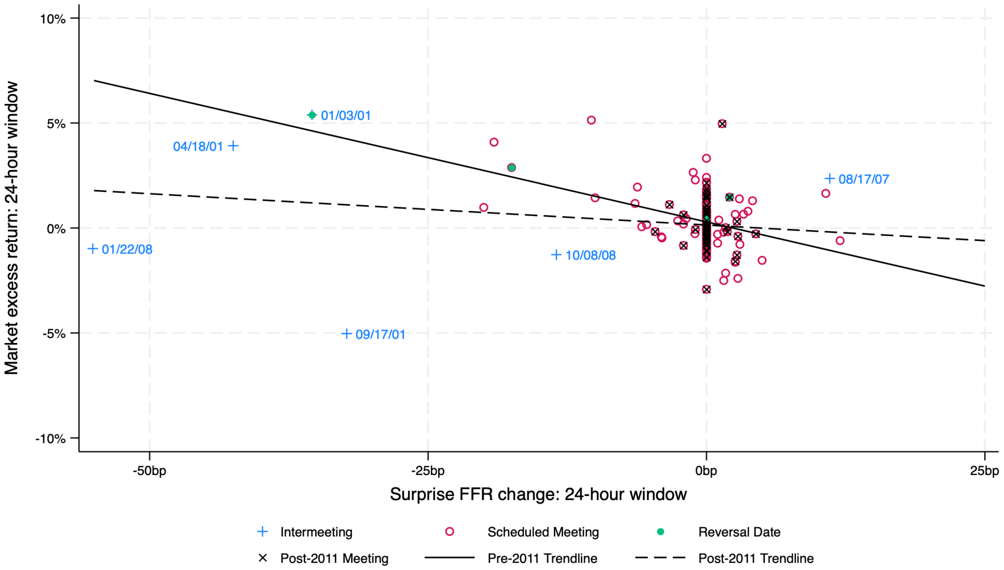
Notes: The scatterplot exhibits one-day returns on the CRSP value-weighted total equity market versus expected components of FFR announcements, according to equation (2). All 162 events of the sample period are included in the plot. The plot distinguishes between intermeetings, regular scheduled FOMC meetings, reversal dates, and post-2011 announcements. Reversal dates coincide with an event in which the Fed reversed the direction of the previous change. The trendlines exclude intermeetings.

Fig. A.2: Polesky's Method—Market Excess Returns vs. Expected FFR Announcements



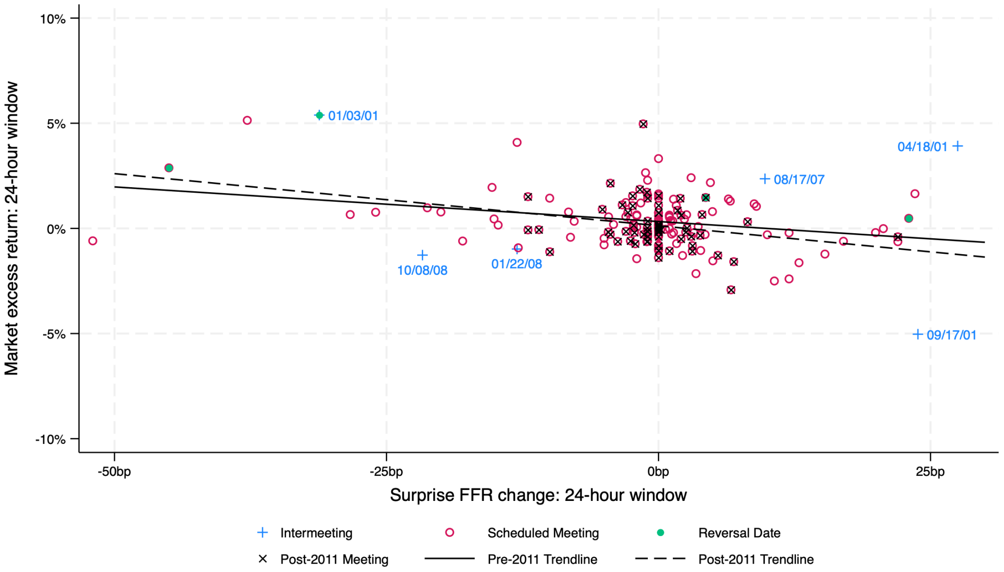
Notes: The scatterplot exhibits one-day returns on the CRSP value-weighted total equity market versus expected component of FFR announcements, according to equation (3). All 162 events of the sample period are included in the plot. The plot distinguishes between intermeetings, regular scheduled FOMC meetings, reversal dates, and post-2011 announcements. Reversal dates coincide with an event in which the Fed reversed the direction of the previous change. The trendlines exclude intermeetings.

Fig. A.3: Kuttner's Method—Market Excess Returns vs. FFR Surprises



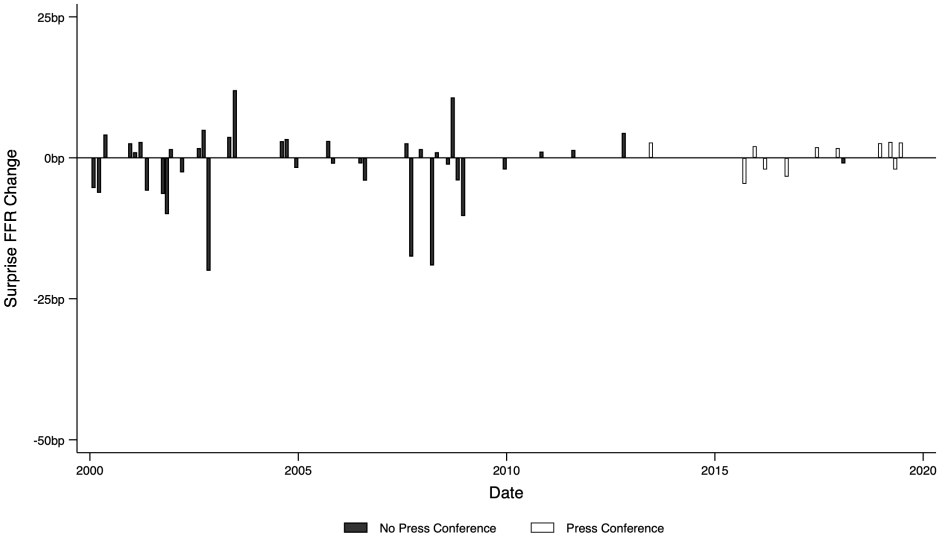
Notes: The scatterplot exhibits one-day returns on the CRSP value-weighted total equity market versus one-day FFR surprises, according to equation (1). All 162 events of the sample period are included in the plot. The plot distinguishes between intermeetings, regular scheduled FOMC meetings, reversal dates, and post-2011 announcements. Reversal dates coincide with an event in which the Fed reversed the direction of the previous change. The trendlines exclude intermeetings.

Fig. A.4: Polesky's Method—Market Excess Returns vs. FFR Surprises



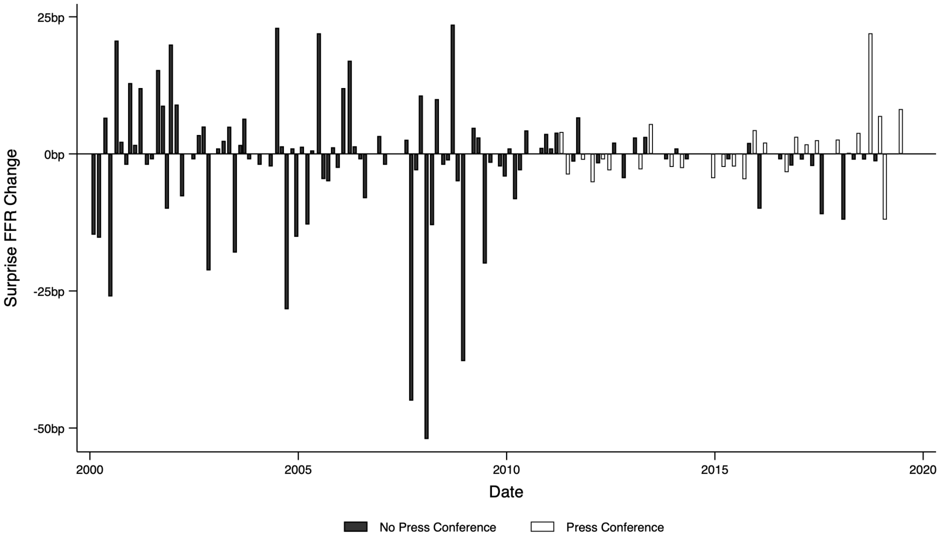
Notes: The scatterplot exhibits one-day returns on the CRSP value-weighted total equity market versus one-day FFR surprises, according to equation (4). All 162 events of the sample period are included in the plot. The plot distinguishes between intermeetings, regular scheduled FOMC meetings, reversal dates, and post-2011 announcements. Reversal dates coincide with an event in which the Fed reversed the direction of the previous change. The trendlines exclude intermeetings.

Fig. A.5: Kuttner's Method—FFR Surprises Over Time



Notes: The time series graph plots the FFR surprises, according to equation (1). All 156 scheduled FOMC events of the sample period are included in the plot. Intermeeting events are excluded. The plot distinguishes between FOMC meetings followed by a press conference, plotted as hollow bars, and those which there are not, plotted as solid bars. The figure exhibits that few of Kuttner's Method's surprise components are noticeably different from zero.

Fig. A.6: Polesky's Method—FFR Surprises Over Time



Notes: The time series graph plots the FFR surprises, according to equation (4). All 156 scheduled FOMC events of the sample period are included in the plot. Intermeeting events are excluded. The plot distinguishes between FOMC meetings followed by a press conference, plotted as hollow bars, and those which there are not, plotted as solid bars. The figure exhibits that many of the largest FFR surprises occurred pre-2011, with the post-2011 period sustaining fewer surprises, and of smaller magnitude. This may suggest the effectiveness of forward guidance. However, this claim may be biased by the extended period of the FFR constrained by ZLB, in which markets expected the FFR to remain unchanged.

Appendix B: Tables

TABLE B.1: FOMC Meetings

Date	(1)	(2)	(3)	Date	(1)	(2)	(3)	Date	(1)	(2)	(3)
2/2/2000	5.75	0.25	0	6/29/2006	5.25	0.25	0	10/24/2012	0.25	0	0
3/21/2000	6	0.25	0	8/8/2006	5.25	0	0	12/12/2012	0.25	0	1
5/16/2000	6.5	0.5	0	9/20/2006	5.25	0	0	1/30/2013	0.25	0	0
6/28/2000	6.5	0	0	10/25/2006	5.25	0	0	3/20/2013	0.25	0	1
8/22/2000	6.5	0	0	12/12/2006	5.25	0	0	5/1/2013	0.25	0	0
10/3/2000	6.5	0	0	1/31/2007	5.25	0	0	6/19/2013	0.25	0	1
11/15/2000	6.5	0	0	3/21/2007	5.25	0	0	7/31/2013	0.25	0	0
12/19/2000	6.5	0	0	5/9/2007	5.25	0	0	9/18/2013	0.25	0	1
1/3/2001 *	6	-0.5	0	6/28/2007	5.25	0	0	10/30/2013	0.25	0	0
1/31/2001	5.5	-0.5	0	8/7/2007	5.25	0	0	12/18/2013	0.25	0	1
3/20/2001	5	-0.5	0	8/17/2007 *	5.25	0	0	1/29/2014	0.25	0	0
4/18/2001 *	4.5	-0.5	0	9/18/2007	4.75	-0.5	0	3/19/2014	0.25	0	1
5/15/2001	4	-0.5	0	10/31/2007	4.5	-0.25	0	4/30/2014	0.25	0	0
6/27/2001	3.75	-0.25	0	12/11/2007	4.25	-0.25	0	6/18/2014	0.25	0	1
8/21/2001	3.5	-0.25	0	1/22/2008 *	3.5	-0.75	0	7/30/2014	0.25	0	0
9/17/2001 *	3	-0.5	0	1/30/2008	3	-0.5	0	9/17/2014	0.25	0	1
10/2/2001	2.5	-0.5	0	3/18/2008	2.25	-0.75	0	10/29/2014	0.25	0	0
11/6/2001	2	-0.5	0	4/30/2008	2	-0.25	0	12/17/2014	0.25	0	1
12/11/2001	1.75	-0.25	0	6/25/2008	2	0	0	1/28/2015	0.25	0	0
1/30/2002	1.75	0	0	8/5/2008	2	0	0	3/18/2015	0.25	0	1
3/19/2002	1.75	0	0	9/16/2008	2	0	0	4/29/2015	0.25	0	0
5/7/2002	1.75	0	0	10/8/2008 *	1.5	-0.5	0	6/17/2015	0.25	0	1
6/26/2002	1.75	0	0	10/29/2008	1	-0.5	0	7/29/2015	0.25	0	0
8/13/2002	1.75	0	0	12/16/2008	0.25	-0.75	0	9/17/2015	0.25	0	1
9/24/2002	1.75	0	0	1/28/2009	0.25	0	0	10/28/2015	0.25	0	0
11/6/2002	1.25	-0.5	0	3/18/2009	0.25	0	0	12/16/2015	0.5	0.25	1
12/10/2002	1.25	0	0	4/29/2009	0.25	0	0	1/27/2016	0.5	0	0
1/29/2003	1.25	0	0	6/24/2009	0.25	0	0	3/16/2016	0.5	0	1
3/18/2003	1.25	0	0	8/12/2009	0.25	0	0	4/27/2016	0.5	0	0
5/6/2003	1.25	0	0	9/23/2009	0.25	0	0	6/15/2016	0.5	0	1
6/25/2003	1	-0.25	0	11/4/2009	0.25	0	0	7/27/2016	0.5	0	0
8/12/2003	1	0	0	12/16/2009	0.25	0	0	9/21/2016	0.5	0	1
9/16/2003	1	0	0	1/27/2010	0.25	0	0	11/2/2016	0.5	0	0
10/28/2003	1	0	0	3/16/2010	0.25	0	0	12/14/2016	0.75	0.25	1
12/9/2003	1	0	0	4/28/2010	0.25	0	0	2/1/2017	0.75	0	0
1/28/2004	1	0	0	6/23/2010	0.25	0	0	3/15/2017	1	0.25	1
3/16/2004	1	0	0	8/10/2010	0.25	0	0	5/3/2017	1	0	0
5/4/2004	1	0	0	9/21/2010	0.25	0	0	6/14/2017	1.25	0.25	1
6/30/2004	1.25	0.25	0	11/3/2010	0.25	0	0	7/26/2017	1.25	0	0
8/10/2004	1.5	0.25	0	12/14/2010	0.25	0	0	9/20/2017	1.25	0	1
9/21/2004	1.75	0.25	0	1/26/2011	0.25	0	0	11/1/2017	1.25	0	0
11/10/2004	2	0.25	0	3/15/2011	0.25	0	0	12/13/2017	1.5	0.25	1
12/14/2004	2.25	0.25	0	4/27/2011	0.25	0	1	1/31/2018	1.5	0	0
2/2/2005	2.5	0.25	0	6/22/2011	0.25	0	1	3/21/2018	1.75	0.25	1
3/22/2005	2.75	0.25	0	8/9/2011	0.25	0	0	5/2/2018	1.75	0	0
5/3/2005	3	0.25	0	9/21/2011	0.25	0	0	6/13/2018	2	0.25	1
6/30/2005	3.25	0.25	0	11/2/2011	0.25	0	1	8/1/2018	2	0	0
8/9/2005	3.5	0.25	0	12/13/2011	0.25	0	0	9/26/2018	2.25	0.25	1
9/20/2005	3.75	0.25	0	1/25/2012	0.25	0	1	11/8/2018	2.25	0	0
11/1/2005	4	0.25	0	3/13/2012	0.25	0	0	12/19/2018	2.5	0.25	1
12/13/2005	4.25	0.25	0	4/25/2012	0.25	0	1	1/30/2019	2.5	0	1
1/31/2006	4.5	0.25	0	6/20/2012	0.25	0	1	3/20/2019	2.5	0	1
3/28/2006	4.75	0.25	0	8/1/2012	0.25	0	0	5/1/2019	2.5	0	1
5/10/2006	5	0.25	0	9/13/2012	0.25	0	1	6/19/2019	2.5	0	1

Notes: The table reports FOMC meeting dates over the sample period, with (1) the FFR target, (2) changes to the FFR, and (3) a dummy taking on a value of 1 if a press conference followed the meeting, and 0 if not. Beginning December 16, 2008, the FOMC moved from a single target rate to a target range with an upper and lower limit. The table reports the upper limit. The * denotes intermeeting dates, of which there are six in the sample.

TABLE B.2: Kuttner's Method—Industry Effects of FFR Surprises

	(1) Full Sample	(2) pre-2011	(3) post-2011	(4) PC: No Controls	(5) PC: Controls
Technology	-15.00***	-16.07***	-0.67	-7.21	-6.62
	(3.37)	(3.88)	(11.79)	(11.21)	(11.91)
Durables	-14.05***	-14.26***	-4.01	-13.35	-15.01
	(3.82)	(3.92)	(16.66)	(13.08)	(13.52)
Other	-13.86***	-14.73***	3.81	-5.92	-5.90
	(3.87)	(4.32)	(14.53)	(11.04)	(11.72)
Manufacturing	-12.48***	-12.95***	-0.20	-8.25	-8.74
	(2.81)	(2.77)	(12.75)	(11.12)	(11.73)
Market	**-12.09*****	**-12.79*****	**-0.63**	**-8.67**	**-8.80****
	(2.76)	**(2.90)**	**(11.61)**	**(9.99)**	**(10.55)**
Telecom	-11.26***	-11.99***	-2.18	-11.46	-10.51
	(3.03)	(3.43)	(11.32)	(9.21)	(9.44)
Shops	-10.81***	-11.40***	1.94	-4.61	-5.18
	(2.85)	(3.11)	(11.23)	(10.03)	(10.62)
Health Care	-10.18***	-10.61***	-2.67	-14.82	-14.86
	(2.38)	(2.32)	(11.08)	(10.15)	(10.67)
Utilities	-8.70**	-8.95***	-8.37	-13.04	-16.02
	(2.78)	(2.64)	(13.32)	(14.60)	(14.84)
Nondurables	-8.03***	-8.46***	0.64	-5.17	-4.88
	(2.12)	(1.97)	(10.29)	(10.84)	(11.40)
Energy	-7.25*	-6.96*	-14.11	-19.04	-20.87
	(3.57)	(3.36)	(17.33)	(17.57)	(18.23)
Observations	156	90	66	36	36
Avg. Adj. R-Square	0.08	0.14	-0.05	0.02	-0.02

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The table reports the results from regressions of one-day returns of the total market and Fama-French industry portfolios (indicated in the row headings) on one-day FFR surprises using Kuttner's Method (equation 1), all expressed in percentage terms. Aside from the method for calculating surprise and expected components of FFR announcements, the regressions are identical to those in Table III. See notes under Table III for additional details.

TABLE B.3: Tests for Asymmetries—Market Effects of FFR Surprises

	(1) Full Sample			(2) Full Sample +	(3) Full Sample +
	Reversal	SEP	No Change	Intermeetings	post-2019
Expected	-0.44	-0.41	-0.27	-0.32	-0.35
	(0.52)	(0.51)	(0.53)	(0.59)	(0.45)
Surprise	-3.61***	-4.82***	-4.13***	-3.65**	-3.66***
	(1.06)	(1.12)	(1.13)	(1.10)	(1.00)
Surprise ×					
Reversal	0.28	...	...	...	...
	(2.58)				
SEP	...	6.69**	...	...	...
		(2.32)			
SEP × post-2011	...	-1.52	...	...	...
		(4.44)			
post-2011	...	-1.73	0.56	...	...
		(4.92)	(5.63)		
No Change	...	...	3.13	...	...
			(2.53)		
No Change × post-2011	...	...	-2.15	...	...
			(4.36)		
Intermeeting	...	...	...	-0.38	...
				(2.67)	
post-2019	...	...	...	...	8.29**
					(2.67)
Intercept	0.35	0.38	0.41	0.40	0.40
	(0.21)	(0.25)	(0.26)	(0.24)	(0.22)
Observations	156	156	156	162	191
Adj. R-Square	0.07	0.11	0.06	0.05	0.06

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The table reports the results from regressions of one-day CRSP value-weighted total equity market return on the one-day surprise and expected components of FFR announcements, all expressed in percentage terms. The full sample (for columns 1-4) includes FOMC meetings spanning February 2, 2000 to June 19, 2019. The reversal dummy equals 1 for rate changes that reverse the direction of the previous change. The SEP (Summary of Economic Projections) dummy equals 1 for events in which the Fed released a SEP report. The post-2011 dummy equals 1 for events occurring from April 27, 2011 to June 19, 2019. The no change dummy equals 1 for events in which the FOMC decided to maintain (not change) the FFR. The intermeeting dummy equals 1 for events in which the FOMC held an unscheduled meeting. Only regression (2) includes intermeeting events. The post-2019 dummy equals 1 for events occurring after June 19, 2019 to December 13, 2023. The post-2019 sample includes 191 scheduled FOMC meetings spanning February 2, 2000 to December 13, 2023, with no controls for forward guidance in the post-2019 period. The regressions include forward guidance controls, whose coefficients are not reported.

TABLE B.4: Placebo Test—Industry Effects of FFR Surprises

	(1) Full Sample	(2) pre-2011	(3) post-2011	(4) PC: No Controls	(5) PC: Controls
Technology	-2.35	-2.58	-2.33	-4.35	-4.73
	(1.30)	(1.62)	(2.98)	(3.61)	(4.52)
Other	-1.56	-1.61	-2.70	-4.48	-4.84
	(1.07)	(1.35)	(2.39)	(2.86)	(3.57)
Market	**-1.05**	**-1.09**	**-2.17**	**-3.55**	**-3.99**
	(0.92)	**(1.13)**	**(2.20)**	**(3.15)**	**(3.32)**
Shops	-0.97	0.96	-2.68	-3.29	-5.34
	(0.90)	(1.10)	(2.23)	(2.63)	(3.23)
Utilities	-0.91	-1.02	-0.04	-1.54	-0.96
	(0.89)	(1.14)	(1.86)	(2.09)	(2.53)
Telecom	-0.62	-0.80	-0.60	-2.07	-3.35
	(1.03)	(1.27)	(2.36)	(2.35)	(2.85)
Energy	-0.57	-0.61	-0.74	-1.26	1.73
	(1.27)	(1.47)	(3.77)	(4.58)	(5.50)
Durables	-0.32	-0.70	2.45	2.00	1.43
	(1.14)	(1.36)	(3.07)	(3.94)	(4.93)
Health Care	-0.18	0.03	-4.30	-4.81	-6.00
	(0.93)	(1.13)	(2.25)	(2.43)	(2.98)
Nondurables	0.16	0.20	-1.41	-2.97	-4.10
	(0.75)	(0.90)	(1.97)	(2.10)	(2.61)
Manufacturing	0.23	0.19	-0.50	-1.77	-1.33
	(0.94)	(1.15)	(2.42)	(2.80)	(3.51)
Observations	156	90	66	36	36
Avg. Adj. R-Square	0.00	0.00	0.02	0.01	-0.04

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The table reports the results from regressions of one-day returns of the total market and Fama-French industry portfolios (indicated in the row headings) on one-day FFR surprises, all expressed in percentage terms. Rather than use the equity returns from the same day as an FOMC event, returns from seven business days prior to the event are used. Using returns from fewer than three days prior to an FOMC event poses the risk of a significant relationship with FFR surprises, given the findings of Lucca and Moench (2015). Seven days is arbitrary. In theory, there should be no significant relationship between FFR surprises and equity returns seven days prior to the announcement. Aside from the equity returns, the regressions are identical to those in Table III. See notes under Table III for additional details.