The Economic Impact of ESG Ratings: A replication study estimating dynamic treatment effects in presence of treatment effect heterogeneity

Thesis

Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science (M.Sc.)
at the Department of Economics
Ludwig-Maximilians-Universität München

Submitted by:

John Doncy

Munich, 20 March 2023

1 Introduction

Researchers in empirical economics have been developing and improving quasi experimental research designs over the past many years. Among these achievements, the difference-in-differences (DiD) estimation approach has been a cornerstone of the "credibility revolution" (Angrist and Pischke, 2010, p.14). The canonical 2x2 DiD design estimates the difference in average outcomes before and after a single period treatment in two groups i.e. treatment versus control. But often times researchers also come across settings involving multiple groups and multiple time periods.[1] Hence, a generalized version of the standard DiD approach is used when the treatment timing is staggered, i.e. different units are treated at different times. Moreover, to estimate dynamic treatment effects (when the effect changes over time), researchers apply two way fixed effects (TWFE) regressions of the following form:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{l} \mu_l.1\{t - E_i = l\} + \nu_{it} \tag{1}$$

where, $1\{t - E_i = l\}$ is a relative-time indicator for unit i being l periods away from the initial treatment, coefficients μ_l estimate the treatment effect relative to an excluded base period (usually l = -1), Y_{it} is the outcome of interest for unit i at time t, E_i is the first time period when unit i receives the binary absorbing treatment[2], ν_{it} is the error term, α_i and λ_t are unit and time fixed effects respectively. Units are also categorized into cohorts based on the period in which they first receive treatment. The relative times l consists of periods which lead to the treatment (leads) and periods following the treatment (lags) (Sun and Abraham, 2021, p.2, De Chaisemartin and d'Haultfoeuille, 2020, p.11, Baker, Larcker, and Wang, 2022, p.18). These dynamic TWFE regressions, also known as event-study models have become popular in the applied literature over the last two decades (Schmidheiny and Siegloch, 2019, p.2, Baker, Larcker, and Wang, 2022, p.1). Due to its intuitive appeal and the fact that a 2x2 DiD estimate is equal to the treatment coefficient in a TWFE regression, researchers tend to extend the TWFE

[1] For example, Hoynes, Schanzenbach, and Almond (2016) uses variation in the roll-out of the U.S. Food Stamp Program to estimate the long-run effects of childhood access to the safety net.

[2] For an absorbing treatment, once the treatment status switches on, it stays on thereafter (Sun and Abraham, 2021, p.5).

design to settings with multiple periods and units. But, in such advanced settings, both the parallel trends assumption and treatment effect homogeneity must hold for TWFE to be an unbiased estimator for the average treatment effect (ATE). In most economic settings, even if parallel trends assumption holds it is highly likely that the treatment effect is heterogeneous across adoption cohorts. For example, while studying the effect of hospitalisation on out-of-pocket spending, Sun and Abraham (2021, p.27) argue that individuals in later treated cohorts are older and may have less insurance cover as compared to earlier cohorts, thereby causing heterogeneous effects. Oftentimes, researchers make this unconvincing and strong assumption of constant treatment effects across cohorts over time. Hence, it is highly important to address this issue for future research and to correct for potential bias in treatment estimates. This has resulted in a growing literature in econometric theory, studying the assumptions and developing alternative estimators needed for TWFE regressions to yield unbiased, causal estimates (Borusyak and Jaravel, 2017, Strezhnev, 2018,Callaway and Sant'Anna, 2021, Sun and Abraham, 2021, De Chaisemartin and d'Haultfoeuille, 2020, Imai and Kim, 2021, Goodman-Bacon, 2021, Athey and Imbens, 2022).

Most of these studies feature an empirical illustration, showing the effectiveness of these alternative estimators by replicating previous TWFE findings[3]. Motivated by their effectiveness and the need for remedying potentially biased estimates, this paper applies an alternative estimator within an interesting financial setting. To be precise, I apply the alternative regression based estimator suggested by Sun and Abraham (2021) to replicate and extend one of the findings of Berg, Heeb, and Kölbel (2022), wherein they explore how Environmental, Social, Governance (ESG) rating changes affect the holdings of dedicated ESG mutual funds. The goal of this study is to investigate whether two-way fixed effects (TWFE) estimates of ESG rating changes (upgrades and downgrades) on mutual fund holdings are biased under staggered treatment timing and treatment effect heterogeneity. The authors use a panel event study model, leveraging variation in the timing of rating changes, to show that the ESG ownership variable, i.e. the fraction of a firm's outstanding shares owned by ESG funds, react to

[3] For example, Goodman-Bacon (2021, p.20) replicate Stevenson and Wolfers (2006) while Sun and Abraham (2021, p.25) replicate Dobkin et al. (2018).

ESG rating upgrades and downgrades over time. I follow the original study in constructing my final data sample using CRSP U.S. Mutual Fund Holdings data and ESG ratings from MSCI. The final sample consists of a balanced panel of 135 U.S. listed companies with 270 ESG rating changes between December 2013 and December 2020. In December 2020, the funds jointly represented USD 30.88 billion in assets under management and on average owned 0.23% of a firm in my sample. I also use ESG data from Refinitiv for robustness check.

First, I follow Berg, Heeb, and Kölbel (2022) and estimate a two-way fixed effects regression to show that holdings of mutual funds react to ESG rating upgrades and downgrades. The results re-confirm that ESG mutual funds increase their ownership in firms which receive a ratings upgrade whereas they decrease ownership in firms which receive a ratings downgrade. This response occurs gradually over a period of two years.

Second, I show that the two-way fixed effects estimates are prone to contamination from other relative periods due to the variation in treatment timing and treatment effect heterogeneity. I follow Sun and Abraham (2021) and perform a decomposition of the even-study coefficient into a linear combination of cohort-specific average treatment effect on the treated ($CATT_{e,l}$) and corresponding weights. The results show that these weights are non-convex and non-zero in nature, preventing the event-study coefficient from accurately estimating the dynamic treatment effect for that period. I also discuss how even in presence of treatment effect homogeneity, the event-study estimates are prone to contamination from relative periods that are excluded from the event-study model.

Third, I implement an alternative estimation strategy to combat this issue of contamination. Accordingly, I use the 'Interaction-weighted' (IW) estimator proposed by Sun and Abraham (2021) to obtain estimates that are robust to treatment effect heterogeneity and by construction produce estimates that are equal to a convex combination of $CATT_{e,l}$ and sample shares of each cohort e as weights.

A fixed effects regression shows that two years after an upgrade, ESG ownership on average is 2.45% higher compared to it's level one month before the upgrade. Whereas, two years after a downgrade, ESG ownership on average is 3.07% lower compared to it's level one month before the downgrade. This is different from the authors' findings which had larger magnitudes of 17.1% and 13.1% for upgrades and downgrades respectively. On the other hand, the IW estimates show that two years after an upgrade, ESG ownership on average is 1.73% higher compared to it's level one month before the upgrade while it is 4.17% lower compared to it's level one month before the downgrade.

The results reveal the importance of re-examining the underlying assumptions of a two-way fixed effects regression in an event-study model. Very often researchers run an event-study analysis under the assumption of treatment effect homogeneity and this may lead to biased estimates in settings with variation in treatment timing and heterogeneous treatment effects. Moreover, there is evidence that ESG rating changes do influence the investment decisions of mutual fund managers despite the effect being smaller in magnitudes. It is important to note that I only replicate a part of the original study and a replication of the remaining findings is necessary to make any remarks about the real impact of ESG ratings. Nevertheless, my results provide further clarity on the impact of ESG ratings on ESG aware investors' holdings and investment decisions. To the best of my knowledge, this is the first study to apply latest econometric methods estimating dynamic treatment effects in a setting related to ESG ratings.

The rest of the paper is organised as follows. In Section 2, I briefly review the related literature. Section 3 introduces event-study design, it's building blocks and the identifying assumptions. Section 4 provides an overview of two-way fixed effects regression under staggered treatment design. Section 5 introduces the IW estimator. Section 6 explains the data used in the study while section 7 covers the empirical setting and identification strategy. Section 8 includes the results, followed by section 9 with robustness checks. Section 10 lays down certain caveats. In section 11, I provide a critical discussion of the results and section 12 concludes. The appendix includes additional robustness checks.

2 Related Literature

This paper draws inspiration from an expanding literature in econometric theory which explores the effectiveness of two-way fixed effects regressions when treatment timing is staggered (Athey and Imbens, 2022, Borusyak and Jaravel, 2017, Sun and Abraham, 2021, Callaway and Sant'Anna, 2021, Goodman-Bacon, 2021, De Chaisemartin and d'Haultfoeuille, 2020). A key element within the literature is to re-examine previous staggered DiD TWFE estimates using newer methods that are more robust.

Goodman-Bacon (2021, p.20-21) demonstrates the potential of the DD decomposition theorem by replicating the study of Stevenson and Wolfers (2006), analysing the effect of legal institutions on outcomes within families. The study shows that the coefficient on a single treatment indicator in a 'static' TWFE regression is different from the average of the post-treatment coefficients in an event-study model. The difference arises when treatment effects evolve over time and the TWFE estimate, a weighted average of several individual 2x2 DiD estimates, uses earlier-treated units as effective controls for later-treated units. While event-study models seem to offer better estimates, Sun and Abraham (2021) show that such models are free from contamination only when certain assumptions hold. They apply their 'interaction-weighted' estimator in re-examining the study of Dobkin et al. (2018), assessing the economic impact of hospitalization on adults. Likewise, I utilise their alternative estimator in this paper as it is robust to treatment effect heterogeneity and specifically focuses on the interpretation of dynamic treatment effects. Callaway and Sant'Anna (2021) is another closely related study featuring alternative estimators allowing for treatment effect heterogenity and dynamic effects. In their empirical application, they follow the literature studying the effect of minimum wage on teen employment(e.g.,Dube, Lester, and Reich, 2010, Meer and West, 2016). Besides, another prominent study by Borusyak and Jaravel (2017) examines the effect of tax rebate receipt on household spending by revisiting the event-study model of Broda and Parker (2014). Using their 'imputation estimator', they find that the treatment effects are upward biased, short-lived and the preferred estimates are only half as large as the original study. Their findings show that the impact of fiscal stimulus on boosting economic growth may be less than previously predicted.

We can see that event-study models are prevalent in numerous research settings. But this paper tries to find out if the above-mentioned methodological differences are also applicable to applied research in finance. Hence, this paper contributes to the existing literature by studying the empirical application of alternative estimators within finance. Berg, Heeb, and Kölbel (2022)'s study on the economic impact of ESG ratings offer an interesting and relevant setting to test the causal interpretability of dynamic estimates in event-study models with a staggered design. This paper also adds to the literature on the impact of ESG ratings by including data from an additional ESG ratings provider. Baker, Larcker, and Wang (2022) is a closely related paper wherein they test the validity of staggered TWFE estimates using alternative estimators. They reexamine two studies relying on staggered treatment identification strategy i.e., Beck, Levine, and Levkov (2010), assessing the impact of bank deregulation on income distribution in the U.S. and Fauver et al. (2017), analysing the impact of corporate board reforms on firm performance. Besides, there are other interesting studies that investigate the impact of ESG ratings in different settings. For example, Chen and Xie (2022) is an interesting paper that features a similar staggered DiD empirical strategy, analysing the impact of ESG disclosures on corporate financial performance. Although they check for robustness using Goodman-Bacon (2021, p.7) decomposition theorem, they do not test the validity of the results using an alternative estimator. A recent study by Lakkis (2022, p.15, 21) finds evidence that the level of holdings by ESG mutual funds affect environmental policies at their portfolio companies. But the author follows the common practice of using pre-period coefficients to test for pre-trends, which could be problematic as the test holds only under strong assumptions. These issues regarding pre-trends is an important finding of Sun and Abraham (2021, p.19-21), which I review in later sections of this paper.

3 A Review of Event Study Design

Within the applied literature an event study could refer to a simple difference-in-differences design where some units receive treatment at a particular time and the others are never-treated. Or, it could refer to a staggered treatment design where units receive initial treatment at different points in time and all units may get treated eventually or some may remain never-treated.[4] In this paper, I follow the second specification of event study models, focusing on staggered treatment design.

Following the setup in Sun and Abraham (2021, p.5), I consider an absorbing treatment whereby treatment "switches on" for some units (indicator D_{it} changes from 0 to 1) at various time periods and thereafter remains the same. The initial treatment therefore defines the treatment path for each unit. Likewise, the initial ESG rating upgrade or downgrade would define the dynamic treatment path for each firm in the sample. Moreover, the initial treatment period is also used to group units into distinct cohorts, i.e., a cohort in my setting would contain firms that receive their first ESG rating upgrade or downgrade in the same month. Hence, in a random sample of $i \in \{0, .., N\}$ and $t \in \{0, .., T\}$ over $T + 1$ time periods, the observed outcome Y_{it} for a unit i in period t, which received its first treatment in period $E_i = e$ would be as follows:

$$Y_{it} = Y_{it}^{E_i} = Y_{it}^{\infty} + \sum_{0 \leq e \leq T} (Y_{it}^{e} - Y_{it}^{\infty}).1\{E_i = e\} \tag{2}$$

For a unit treated in period e ($D_{it} = 1\{E_i = e\}$), the observed outcome includes the "baseline outcome" (Y_{it}^{∞}), i.e., the potential outcome if unit i never receives the treatment and an additional unit-level treatment effect (Y_{it}^{e} - Y_{it}^{∞}), i.e., the difference between unit i receiving treatment in period e and the counterfactual of never receiving treatment.[5] All units within a cohort $e \in \{0, .., T, \infty\}$ receive treatment at the same time. It is also assumed that the observations $\{Y_{it}, D_{it}\}_{t=0}^{T}$ are independent and identically distributed(i.i.d.).

In the next section, I explain how Sun and Abraham (2021, p.5-6) extend this basic framework to arrive at their definition of cohort-specific average treatment effects on the treated or (CATT).

[4] See (Roth, 2022, p.21 or Sun and Abraham, 2021, p.9-10) for examples of papers using these different specifications.

[5] This is similar to a simple 2x2 DiD setup where $Y_{it} = Y(0)_{it} + [Y(1)_{it} - Y(0)_{it}].D_{it}$.

3.1 Cohort-Specific Average Treatment Effect on the Treated (CATT)

In order to study the causal interpretation of dynamic treatment effect estimates, it is crucial to understand the constituents of the coefficients μ_l derived from a TWFE regression as in (1). The idea is to decompose the population regression coefficient μ_l into an average of unit-level treatment effects for a given relative period l across units in cohort e, who received their first treatment at the same time $\{E_i = e\}$. Sun and Abraham (2021, p.7) consider this cohort-specific ATT, l periods from initial treatment as their "building-block" and formally define it as follows:

$$CATT_{e,l} = E[Y_{i,e+l} - Y^{\infty}_{i,e+l} \mid \{E_i = e\}] \tag{3}$$

where, each $CATT_{e,l}$ equals the average treatment effect of being l periods from initial treatment e across all units within the cohort $\{E_i = e\}$. Also, ATT's are estimated instead of ATE's since they do not assume random treatment timing (Roth and Sant'Anna, 2021, p.10).

3.2 Identifying Assumptions

In this section, I recap the identifying assumptions mentioned in Sun and Abraham (2021, p.7-9) for their event study design.

So as to identify the ATT from an event study or dynamic TWFE regression, certain identifying assumptions need to be imposed. The three main assumptions considered are the following:

Assumption 1 - Parallel Trends in baseline outcomes

In the canonical DiD design this assumption states that in the absence of treatment, the baseline outcome for both the treated group and control group would be the same, i.e., they would follow a parallel path over time. For example in a 2x2 DiD model, let s be the group that receives treatment moving from period 1 to 2 and n be the control group which is untreated in both periods. Also, for unit $i \in \{s, n\}$ and $t \in \{1, 2\}$, let $Y_{i,t}(1)$ be the outcome when treated and $Y_{i,t}(0)$ be the outcome when untreated. Then, according to parallel trends assumption,

$$E[Y_{s,2}(0) - Y_{s,1}(0)] = E[Y_{n,2}(0) - Y_{n,1}(0)]$$

where, the counterfactual expected outcome for the treated group ($E[Y_{s,2}(0) - Y_{s,1}(0)]$) would evolve the same as the control group $E[Y_{n,2}(0) - Y_{n,1}(0)]$. Here, the outcome only depends on the current treatment and does not depend on the previous treatment effects (De Chaisemartin and d'Haultfoeuille, 2020, p.2).

But when dynamic treatment effects are included, the treatment for unit i in cohort e at period t is allowed to depend on effects from previous periods. Also, the parallel trends in outcome in comparison to units without treatment ($Y_{i,t}(0)$) is replaced by parallel trends in outcome in comparison to units without having ever been treated ($Y_{i,t}(\mathbf{0}_t)$), where $\mathbf{0}_t$ is a vector of t zeroes. (Refer Equation. 16 De Chaisemartin and d'Haultfoeuille, 2020, p.15). Hence, for all $s \neq t$,

$$E[Y_{i,t}^{\infty} - Y_{i,s}^{\infty} \mid E_i = e]$$

would be the same for all $e \in support(E_i)$, i.e., the expected outcome, in the absence of treatment, would follow the same path for all cohorts across all time periods.

If never-treated units are present in the sample ($\infty \in support(E_i)$) then the assumption would require that the expected outcomes for the units first treated in e would have evolved as that of the never-treated units in the absence of treatment.

$$E[Y_{i,t}^{\infty} - Y_{i,t-1}^{\infty} \mid E_i = e] = E[Y_{i,t}^{\infty} - Y_{i,t-1}^{\infty} \mid E_i = \infty]$$

In settings without a never-treated cohort, Sun and Abraham (2021, p.5) use last-treated cohort as the control group. The choice of control groups can vary based on the research setting and plausibility of assumptions.[6] [7] Very often never-treated units may differ from ever treated units. For example, in their empirical application Sun and

[6] For example, Callaway and Sant'Anna (2021, p.8) looks at both 'never-treated' and 'not-yet-treated' groups.

[7] See Marcus and Sant'Anna (2021) for a detailed explanation of the different parallel trends assumptions within the literature and their implications.

Abraham (2021, p.26-27) restrict the parallel trends assumption to units that were ever hospitalized instead of units that were never hospitalized.

Assumption 2 - No anticipatory behaviour prior to treatment
This assumption means that the treatment effect in pre-treatment periods is equal to zero, i.e., $E[Y^{e}_{i,t} - Y^{\infty}_{i,t} \mid E_i = e] = 0$ for all $e \in support(E_i)$ and relative periods $l < 0$.

Essentially, if units have access to any information regarding their treatment path or if they can choose their treatment status, then it is highly likely that they may adjust their behaviour in the pre-treatment periods in anticipation. Such changes in behaviour would lead to a pre-treatment outcome different from the baseline outcome. For example, in the case of unexpected hospitalization, it is highly unlikely that individuals anticipate such a treatment (Sun and Abraham, 2021, p.27).

Assumption 3 - Treatment effect homogeneity
This assumption means that the treatment effect for each relative period l must be the same across all cohorts e. In other words, for each l, $CATT_{e,l}$ does not depend on the cohort e and is equal to ATT_l (Sun and Abraham, 2021, p.8-9).

Regardless of being an earlier-treated or later-treated cohort, the treatment effect remains the same. In most cases, this assumption is implausible since it only takes the effect to differ across cohorts in any one relative period for it to be violated. For example, considering age as a covariate, individuals who receive their first hospitalization (i.e., treatment) in later cohorts are automatically older than the early cohorts. This difference in covariates can violate the homogeneity assumption (Sun and Abraham, 2021, p.27).

4 TWFE Regressions under Staggered Treatment

In this section, I give an overview of the two common specifications of TWFE reegressions, i.e., the static and dynamic specifications in settings with binary treatment and staggered design. Besides, I also discuss why the causal interpretation of TWFE estimates is problematic by reviewing recent findings on this topic.

4.1 Static Specification

Goodman-Bacon (2021, p.7) show that the 'static' TWFE estimate (μ^{DD}) obtained from a regression such as,

$$Y_{it} = \alpha_i + \lambda_t + \mu^{DD} D_{it} + \nu_{it} \tag{4}$$

is a variance-weighted average of all possible individual 2x2 DiD estimates (See theorem 1 in the paper). The main problem here is that among these 2x2 DiD estimates, there are pairs where already treated groups serve as controls for later-treated groups. When using the already treated group as control, one ends up subtracting the change in treatment effects over time for this group from the $\widehat{\mu}$ (late vs early group) comparison. This in turn leads to a biased estimate (See Equation 11(c) in Goodman-Bacon (2021, p.10)).

Moreover, Goodman-Bacon (2021, p.11) also provides a breakdown of the probability limit of the TWFE estimator, which includes three elements:

$$\underset{N \to \infty}{\text{plim}} \quad \widehat{\mu}^{DD} = VWATT + VWCT - \Delta ATT \tag{5}$$

where $VWATT$(variance-weighted average treatment effect on treated) is a positively weighted average of ATT's for treatment groups and post-periods covering all 2x2 DiD's that constitute $\widehat{\mu}^{DD}$. $VWCT$ (variance-weighted common trend) is the weighted average of the difference in untreated potential outcomes across groups and periods in 2x2 DiD's that constitute $\widehat{\mu}^{DD}$. This is an extension of the parallel trends outcome to a staggered treatment design. Hence, when parallel trends hold, $VWCT$ = 0 and when $VWCT \neq 0$, there exists differences in the counterfactual outcomes across groups. The third term ΔATT, is a weighted sum of the change in ATT within a treatment group's post-period window. It measures by how much the already-treated unit's outcome have changed when they are used as control in the later period and how much of this change in outcome is attributed to change in treatment effects over time.

With 'static' treatment effects (when the effects are allowed to vary across groups but not over time), ΔATT = 0 and together with $VWCT$ = 0, we get $VWATT$. But, when treatment effects vary across time $\Delta ATT \neq 0$ and $\widehat{\mu}^{DD}$ is biased even if parallel trends hold. Moreover, with treatment effect heterogeneity, OLS estimator applies different weights on the ATT for each group. This leads us to Sun and Abraham (2021) which can be seen as an extension of Goodman-Bacon (2021) to dynamic treatment effect estimates.

4.2 Dynamic Specification

Researchers use a variant of (4) to estimate a 'fully dynamic' specification of TWFE regressions or event-study regressions of the following form,

$$Y_{it} = \alpha_i + \lambda_t + \sum_{l=-K}^{-2} \mu_l D_{it}^l + \sum_{l=0}^{L} \mu_l D_{it}^l + \nu_{it} \tag{6}$$

where certain relative periods within and beyond the treatment window (i.e., -T,...-K-1,-1,L+1,...,T) are excluded to avoid multicollinearity[8]. $D_{it}^l = 1\{t - E_i = l\}$ is an indicator for unit i being l periods away from initial treatment at calendar-time t. Usually, the period prior to the treatment D_{it}^{-1} is excluded and the inference is made in reference to this indicator. Sun and Abraham (2021) uses relative-time instead of calendar-time in their specification. The coefficient μ_l estimates the difference in outcome differences between treated and untreated groups l periods from treatment relative to the outcome differences between treated and untreated groups in the excluded periods (Baker, Larcker, and Wang, 2022, p.18). For post-treatment periods ($l \geq 0$), μ_l provides the cumulative effect of $l + 1$ treatment periods while μ_l in pre-treatment periods ($l < 0$) is used as a check for the parallel trends assumption.

But, in the presence of treatment effect heterogeneity and staggered timing, Sun and Abraham (2021, p.15) show that the coefficients μ_l are contaminated by treatment effects from other periods (See Proposition 3 of their paper). According to Proposition 3, even when parallel trends and no anticipation assumptions hold, the population regression coefficient μ_l is a linear and non-convex combination of post-treatment CATT's from it's own relative period l and other relative periods.

[8] See (Borusyak and Jaravel, 2017) for more on collinearities or (Sun and Abraham, 2021, p.12) for a very brief discussion of the same.

$$\begin{aligned}\mu_g = &\sum_{l' \in\, g,\, l'>0} \sum_e \omega^g_{e,l} CATT_{e,l} \;+ \sum_{g' \neq\, g,\, g' \in G} \sum_{l' \in\, g',\, l'>0} \sum_e \omega^g_{e,l'} CATT_{e,l'} \;+ \\ &\sum_{l' \in\, g^{excluded},\, l'>0} \sum_e \omega^g_{e,l'} CATT_{e,l'}\end{aligned} \tag{7}$$

The first term in the above equation is a weighted average of CATT's across treatment cohorts in the post-treatment periods ($l' > 0$) within bin g, where G includes disjoint sets g of relative periods l. In this 'fully dynamic' specification, each g is a singleton where $g = \{l\}$. The last two terms show the contamination from other periods and excluded periods respectively. The contamination results from the interaction between the weights and corresponding $CATT_{e,l'}$. The weights are non-linear functions of the distribution of cohorts. More importantly, the weights[9] on the first term sum to 1, on the second term sum to 0 and on the third term sum to -1. So the second term cancels out only when we impose treatment effect homogeneity as the weights sum to zero. Also, no anticipation makes $CATT_{e,l} = 0$ for all $l < 0$ and they drop out from the equation for μ_g . Nevertheless, in a setting with both treatment effect homogeneity and parallel trends holding, we still get a bias due to contamination from the excluded periods. One way to avoid this contamination is to only exclude periods with zero treatment effect. Hence, we need to pay attention to what periods are being excluded and their effects. Finally, for each relative period l in the fully dynamic specification under parallel trends and treatment effect homogeneity, we get,

$$\mu_l = ATT_l + \sum_{l' \in\, g^{excluded}} \omega^g_{l'} ATT_{l'} \tag{8}$$

This decomposition also explains why using pre-period coefficients as evidence for parallel trends in event-study models is problematic. The pre-treatment coefficients may be zero when parallel trends do not hold and may be non-zero when parallel trends actually exist.[10]

[9] See Equation (12) in Sun and Abraham (2021, p.14) for the auxiliary regression estimating the weights. They regress relative period indicators on bin indicators along with two-way fixed effects.

[10] A common practice is to show that pre-treatment coefficients are close to 0 as evidence for parallel trends.

5 Alternative: Interaction-Weighted (IW) Estimator

After analysing the potential pitfalls associated with TWFE regressions in staggered settings, Sun and Abraham (2021) recommend an alternative estimation strategy that is robust to treatment effect heterogeneity. The idea is to obtain a similar weighted average of $CATT_{e,l}$ wherein the weights sum to one and are non-negative. Besides, the alternative method uses 'never-treated' or 'last-treated' (when never-treated units are absent) cohorts as control groups and hence overcomes the issue of improper comparisons using 'already treated' units as controls in later periods. The alternative regression method does not include covariates.[11]

The IW estimation procedure is completed in three steps. First, to estimate $CATT_{e,l}$, cohort indicators and relative-time indicators are used as an interaction term in the following linear two-way fixed effects regression:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l}(1\{E_i = e\}.D^l_{i,t}) + \epsilon_{i,t} \tag{9}$$

C is a set that includes cohorts used as control, the indicators for which are excluded from the regression. If there are never-treated units then they are used as control where $C = \{\infty\}$ and (9) is estimated using all observations. But when last-treated units act as control where $C = \{max\{E_i\}\}$, regression (9) is estimated on observations from $t = 0, ..., \{max\{E_i\}\} - 1$. Here, the last period is excluded since all units would be treated by then. If there is an always-treated cohort, it is excluded from the estimation as well. The coefficient $\widehat{\delta_{e,l}}$ is a DiD estimator for $CATT_{e,l}$. In the second step, the weights are calculated as sample shares of each cohort in the respective relative-period l. For example, if there are 4 cohorts for D^1_{it} with equal number of observations, then each has a weight of 0.25 i.e., the probability of being in cohort e conditional on the relevant period under consideration. We see that all weights are positive and the issue of negative weights is mitigated. In the final step, once again a weighted average of the $CATT_{e,l}$'s from step 1 and weights from step 2 is calculated to obtain the IW estimate.

[11] It is still possible to include covariates but it complicates the interpretation of the treatment estimate.(Sun and Abraham, 2021, p.14). However, it fits for the setting of my paper as the original study does not use any covariates in their identification strategy.

Hence, the IW estimate $\widehat{\nu_g}$ is given by,

$$\widehat{\nu_g} = \frac{1}{|g|} \sum_{l \in g} \sum_{e} \widehat{\delta_{e,l}} \widehat{Pr}\{E_i = e \mid E_i \in [-l, T-l]\} \tag{10}$$

where, $\widehat{\delta_{e,l}}$ and $\widehat{Pr}\{E_i = e \mid E_i \in [-l, T-l]\}$ are the estimates for $CATT_{e,l}$ from step 1 and estimated weights from step 2 respectively. In essence, under parallel trends and no anticipation, the estimate $\widehat{\delta_{e,l}}$ is a consistent estimator for $CATT_{e,l}$ and this holds regardless of treatment effect homogeneity or heterogeneity. Assuming a pre-period[12] $s < e$ and non-empty treatment cohort e and control cohort C exists, the estimate for $CATT_{e,l}$ is given by,

$$\widehat{\delta_{e,l}} = \frac{\frac{1}{N}\sum_{i=1}^{N} \ [(Y_{i,e+l} - Y_{i,s}) \ . \ 1\{E_i = e\}]}{\frac{1}{N}\sum_{i=1}^{N} \ . \ [1\{E_i = e\}]} - \frac{\frac{1}{N}\sum_{i=1}^{N} \ [(Y_{i,e+l} - Y_{i,s}) \ . \ 1\{E_i \in C\}]}{\frac{1}{N}\sum_{i=1}^{N} \ . \ [1\{E_i \in C\}]} \tag{11}$$

See section 4.2 of Sun and Abraham (2021, p.23-24) for a detailed overview of $\widehat{\delta_{e,l}}$ being an unbiased, consistent estimator of $CATT_{e,l}$.

6 Data

The primary data sources used in this study are Mutual Fund Holdings data from Center For Research in Security Prices (CRSP) and ESG data from MSCI and Refinitiv. In order to gauge the effect of ESG rating changes on the holdings of ESG funds, Berg, Heeb, and Kölbel (2022, p.10) rely on a measure called ESG ownership.

6.1 ESG ownership

ESG ownership is defined as the fraction of a company's outstanding shares that are owned by funds with an explicit ESG or sustainability strategy. Similar to the authors, I identify ESG mutual funds by screening fund names from the CRSP U.S. Mutual Fund

[12] A common practice is to use one period before treatment as pre-period, in that case $s = e - 1$ as we exclude $l = -1$.

Holdings Database using a set of ESG-related keywords[13]. The search covers domestic U.S. Equity funds following a capitalization, growth, growth and income or income based strategy. I identify 197 ESG-related funds along with their respective portfolio holdings.[14] The single unit of observation is each company within the portfolio of all identified ESG funds. The objective is to calculate the aggregate number of shares held by the funds for each company per month in the sample. Additionally, the total number of shares outstanding is also obtained from CRSP Monthly Stock Database.[15] The variable ESG ownership is then constructed by dividing the aggregate number of shares owned by the identified funds by the total shares outstanding for each company in each month. Therefore, ESG ownership gives us a measure of the joint ownership or holdings of ESG mutual funds in specific companies for each month in the sample. For example, the identified ESG mutual funds jointly owned 0.15 percent of Tesla's stocks in September 2020. This finding is identical to those of the authors' and provides evidence that my final dataset is constructed correctly. Similar to the original study, I trim the ESG ownership variable at the 1st and 99th percentiles for each month to account for any outliers.

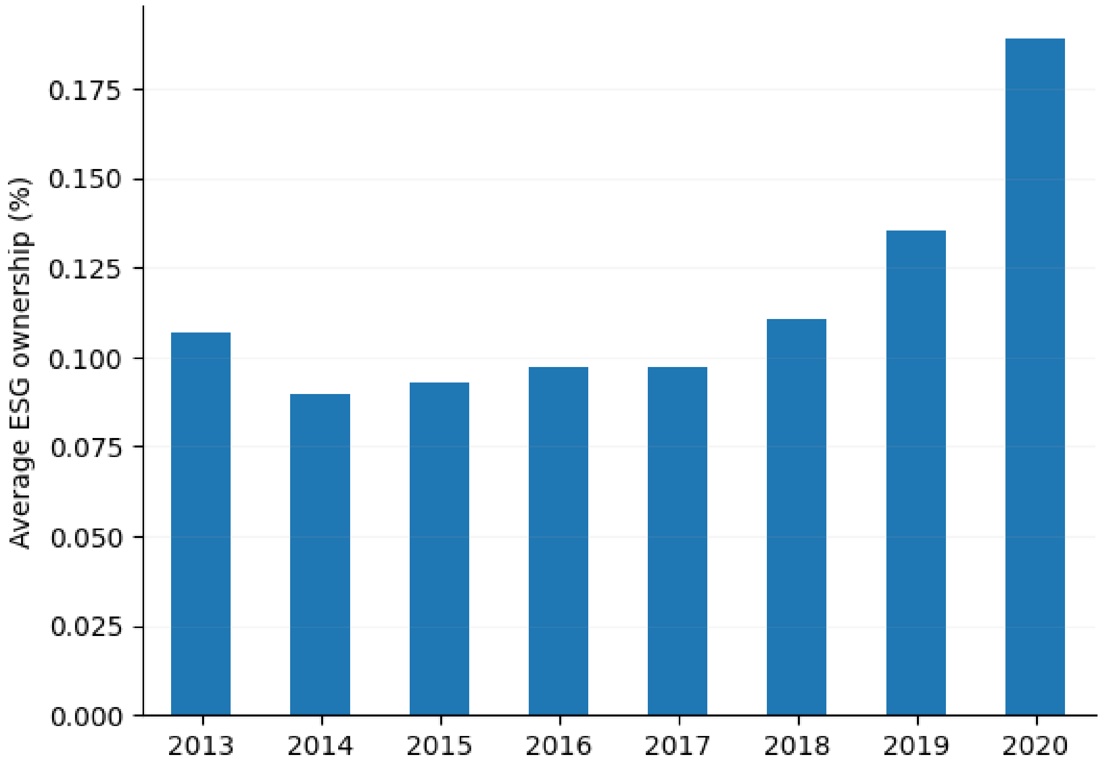
I find an increase in average ESG ownership over the 2013-2020 time period. Figure 1. shows the development in the proportion of companies' shares owned by ESG Mutual funds. Similar to Berg, Heeb, and Kölbel (2022), a key takeaway is that ESG ownership, while being more or less constant between 2013 and 2017, had a significant increase between 2018 and 2020. This highlights the increasing popularity of ESG-based mutual funds as an investment strategy in today's world.

But in terms of correlation between ESG ownership and numerical ESG scores, I obtain different results for the two ESG data sources. For September 2020, the ESG scores from MSCI show a significant positive correlation of 0.52 ($p < 0.001$) with the ownership variable while scores from Refinitiv, although significant, shows a much lower correlation of 0.13 ($p < 0.001$). Figure 2. plots the linear correlation between the two variables using both ESG sources. The diversion in results could be driven by the

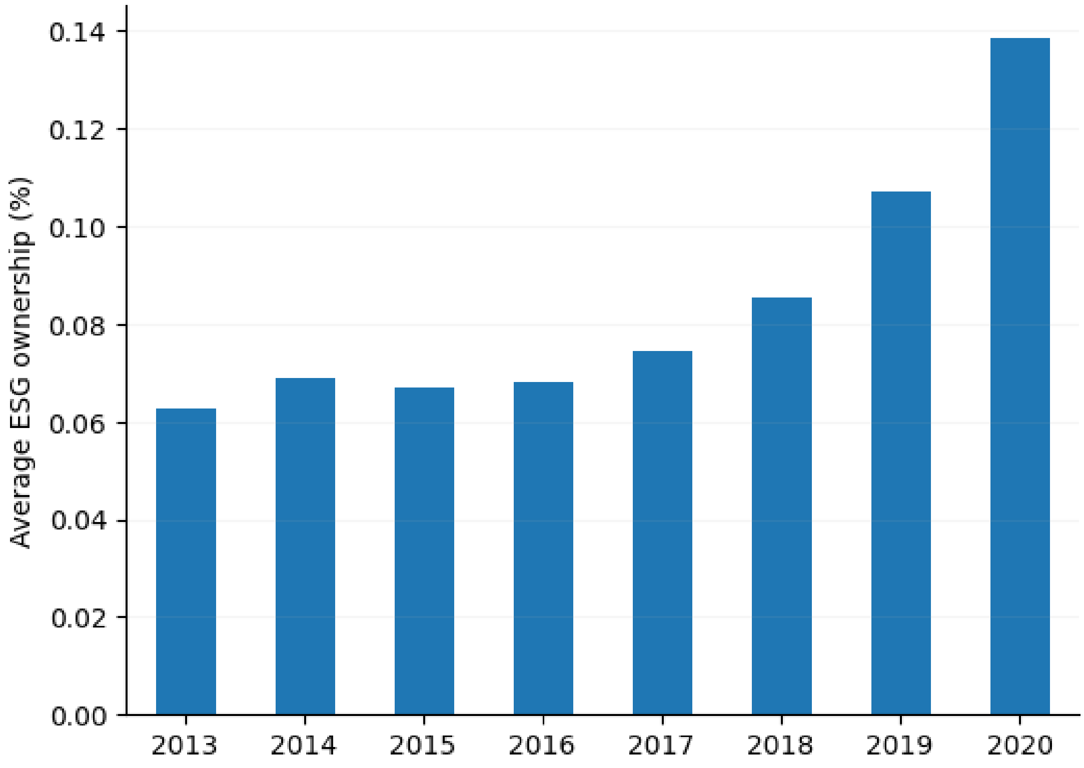
[13] Funds with the following keywords in their title were included in this study: SRI, social, ESG, green, sustain, environ, impact, responsible, clean, renewable.

[14] The original study did not mention the exact number or list of mutual funds used for the analysis.

[15] All CRSP data used in this replication is obtained from Wharton Research Data Services (WRDS) from the University of Pennsylvania https://wrds-www.wharton.upenn.edu/



(a) MSCI sample



(b) Refinitiv sample

Figure 1: **ESG ownership over time.**
This figure shows the increase in average ESG ownership (in percentage points) during the period 2013-2020. ESG ownership is calculated as the fraction of a company's outstanding shares owned by ESG mutual funds.

market share of different ESG data providers, wherein MSCI controls a higher share compared to Refinitiv.[16]

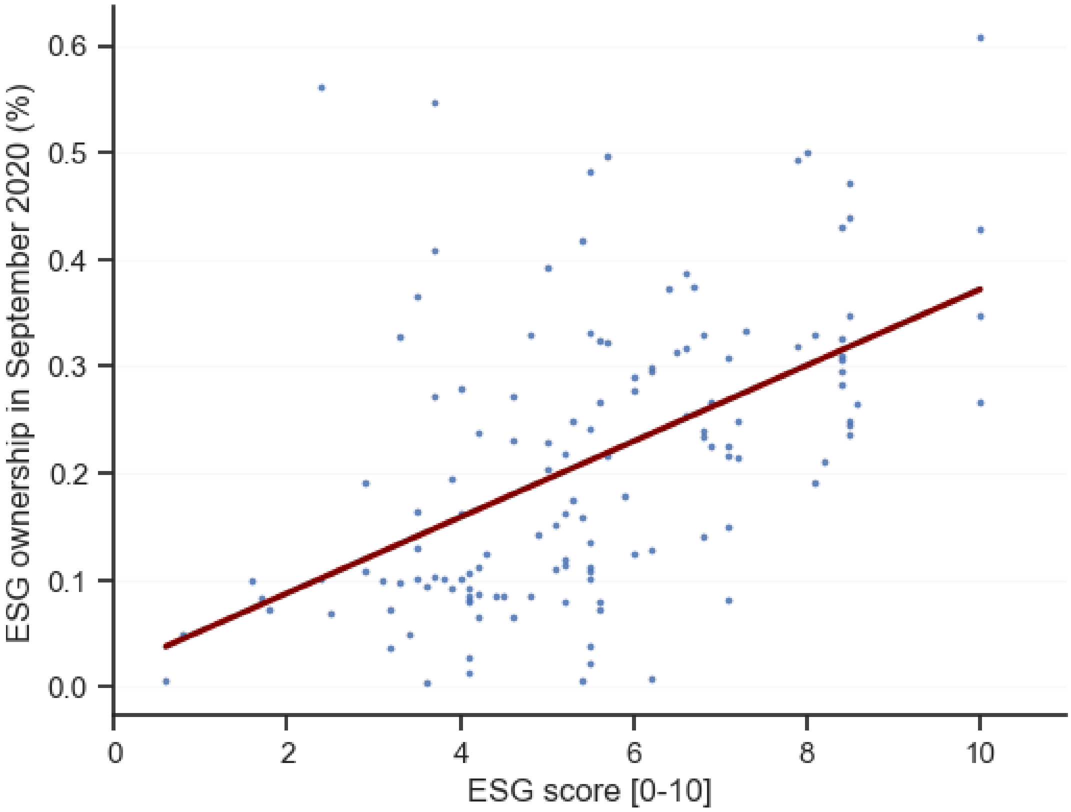
6.2 MSCI ESG rating changes

I use a balanced monthly panel data sample of ESG ratings from MSCI. It is slightly different to that of the original study such that my sample consists of 135 U.S. listed companies with 270 ESG rating changes between December 2013 and December 2020. I observe 135 upgrades and 135 downgrades in the sample. Figure 3. depicts the distribution of the underlying changes in the numerical ESG scores for both up and downgrades. Most changes lie in the range of (-1.5, -1) and (1, 1.5) for ESG numerical scores. Changes of extremely high magnitudes are not that common.

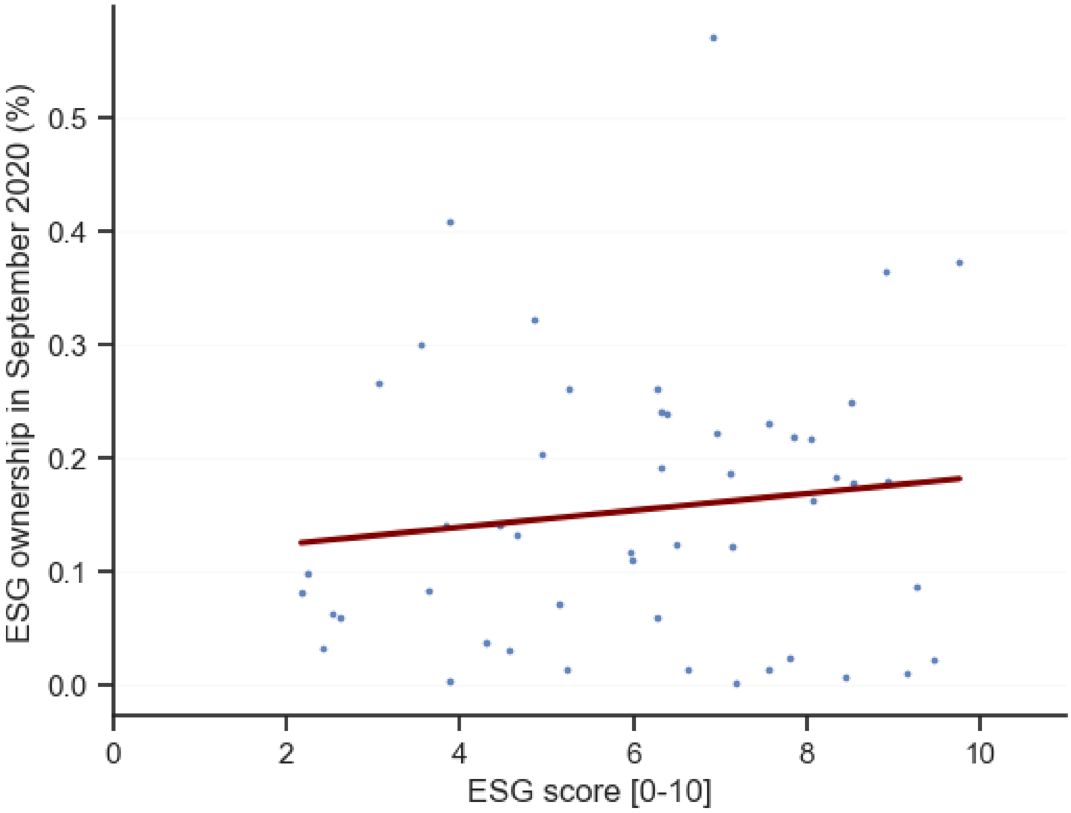
A set of 37 key-issues including (Carbon Emissions, Corruption, Health and Safety etc.) are used to formulate the three pillar scores covering environment (E), social (S) and corporate governance (G) aspects. Companies are evaluated on the basis of an exposure score, measuring companies' exposure to risks and a management score, measuring the ability of companies' to handle those risks. These two scores together constitute an issue score and a weighted sum across all the issues results in a raw ESG score. This score is then benchmarked against the raw scores at an industry level to obtain the final ESG score for the company. The numerical score ranges from 0-10, where 0 indicates poor ESG performance and 10 being the best.

Likewise, these scores are used to further sort the companies into seven equally spaced letter-based categories ranging from CCC to AAA, with AAA being the best ESG performance rating. When the numerical ESG score crosses certain thresholds based on updates from new company level data, it results in a change in the letter-based categories. For example, 0 - 1.428 corresponds to a CCC score and if the numerical score moves into the next bin i.e., 1.429 - 2.856, then the score changes to B. The score can change due to multiple reasons ranging from an improvement or deterioration in a company's ESG practices to MSCI's weighting of the financial relevance of the issues underlying the methodology or even due to changes in industry level

[16] https://www.opimas.com/research/742/detail/



(a) MSCI sample



(b) Refinitiv sample

Figure 2: **ESG ownership and ESG scores (A cross-sectional view)**
This figure shows the distribution of ESG ownership against numerical ESG scores for each firm in the sample for September 2020. The red line represents the linear relationship between the two variables for this date. The scores for Refinitiv are re-scaled to a range (0-10).

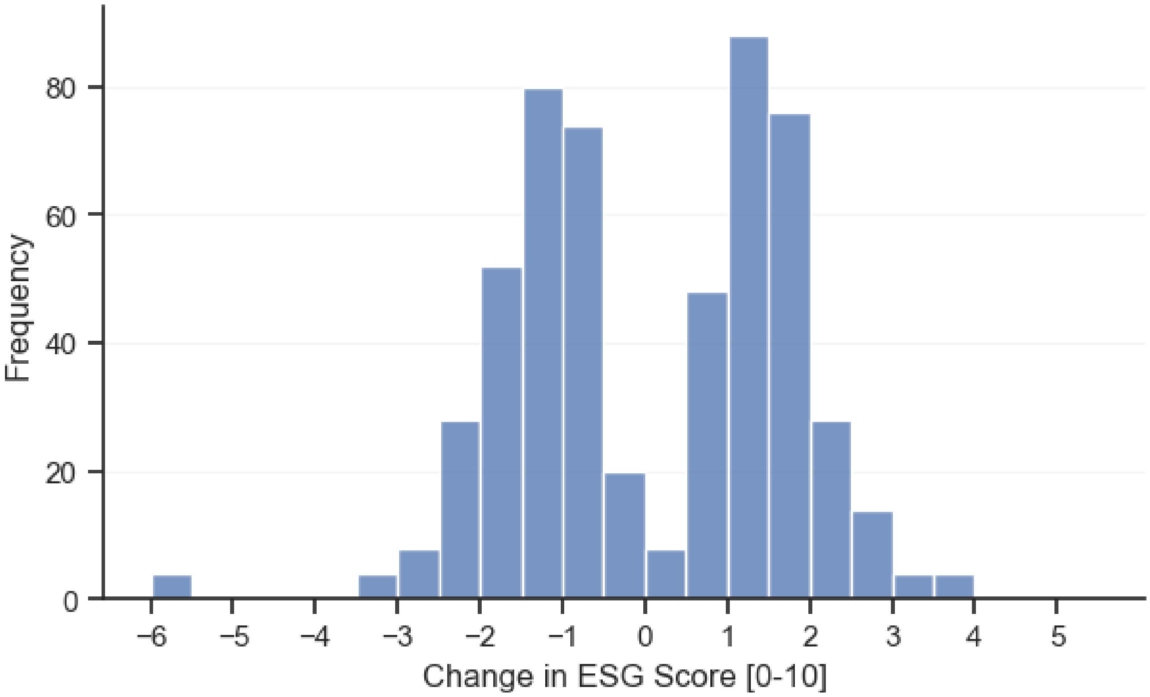


Figure 3: **The size of MSCI ESG rating changes.**
This figure shows the distribution of the size of changes in numerical ESG score for MSCI ESG rating up- and downgrades. Each bin has a width of 0.5 score points. I observe 135 rating upgrades and 135 rating downgrades during the period 2013-2020.

benchmarks (Berg, Heeb, and Kölbel, 2022, p.7).

6.3 Refinitiv ESG Rating Changes

The other ESG data source used in this study comes from Refinitiv. Their ESG scoring methodology uses 186 industry relevant and company-level metrics that are obtained from publicly available data, corporate disclosures and news reports. These metrics are grouped into 10 main categories such as emissions, innovation, human rights, CSR etc., which are reformulated to form the three pillar scores, environment (E), social (S) and corporate governance (G) and the final overall ESG score. The overall score reflects a company's performance along the ESG dimensions and it's commitment towards managing risks and opportunities in a sustainable and holistic manner. Depending on the industry group, each of the 10 categories receive certain weights based on the relevance of themes linked to that particular industry group. The company level values for each of these categories are then multiplied by the industry level category weights to form a weighted sum or the overall ESG score. The scores range from 0 - 100 and can also be

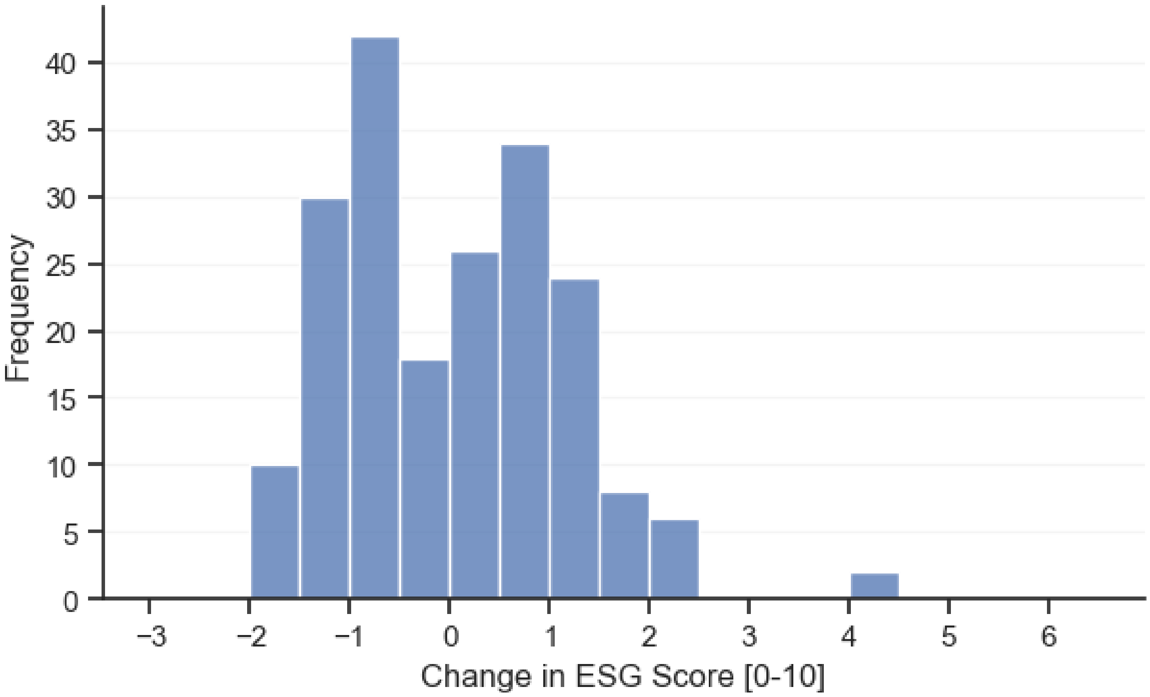


Figure 4: **The size of Refinitiv ESG rating changes.**
This figure shows the distribution of the size of changes in numerical ESG score for Refinitiv ESG rating up- and downgrades. Each bin has a width of 0.5 score points. The scores are re-scaled to a range (0-10). I observe 50 rating upgrades and 50 rating downgrades during the period 2013-2020.

converted to 12 groups ranging from A+ to D-, where a score of 91.6 - 100 and grade of A+ indicates excellent relative ESG performance and high degree of transparency in reporting ESG data publicly. [17] These letter based ratings are equally spaced and changes when the underlying numerical ESG score crosses a threshold, which puts it in a relatively better or worse category. These numerical scores get updated based on the availability of new company related information which feeds into the scoring methodology.

I use a balanced monthly panel data sample of 50 U.S. listed companies with 100 rating changes between February 2013 and September 2020. I observe 50 upgrades and 50 downgrades in my sample. Figure 4. shows the distribution of changes in the numerical ESG scores that lead to rating changes for both up and downgrades. Most changes are in a range of (-1, -0.5) and (0.5, 1) for the ESG numerical scores. Changes of extremely high magnitudes are not that common.

[17] Refer to the following material https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf for a detailed explanation of the scoring methodology.

Table 1 provides summary statistics on the balanced panel data samples used for primary analysis.

Table 1: Summary Statistics

	N	Mean	Std.Dev
Panel A. MSCI			
ESG ownership (%)	11475	0.116	0.103
ESG weight (%)	11475	0.284	0.456
ESG score [0-10]	11475	4.883	2.113
Panel B. Refinitiv			
ESG ownership (%)	4600	0.08	0.10
ESG score [0-100]	4600	52.25	18.74

Note: This table presents the summary statistics of firm-level characteristics from the primary balanced sample used for analysis. Observations for Panel A. MSCI (135 firms) ranges from December 2013-2020 while Panel B. Refinitiv (50 firms) ranges from February 2013-September 2020. ESG ownership is calculated as the fraction of a company's outstanding shares owned by ESG mutual funds. ESG weight is the fraction that a company's shares represent in the portfolio value of a synthetic ESG mutual fund that aggregates the holdings of all ESG mutual funds. All fund related data is obtained from CRSP mutual fund and monthly stock databases. ESG ratings are from respective providers.

7 Empirical Framework

Berg, Heeb, and Kölbel (2022) use a panel event study model to estimate the effect of ESG rating changes on ESG ownership. Essentially, they rely on a dynamic two-way fixed effects regression with a staggered treatment design to obtain non-parametric treatment effect estimates. Different companies receive the treatment i.e., an ESG rating upgrade or downgrade in different months within the sample. They utilise this variation or staggered approach in treatment timing within their identification strategy. Firms with a rating change are compared against the counterfactual of untreated firms which do not receive a rating change at the same time. The main identifying assumption is parallel trends in outcomes between treated and untreated units in the absence of treatment. Furthermore, in section 7.1, I discuss how the parallel trends assumption and no anticipation behaviour are plausible in this setting. I also argue why the treatment effect could be heterogeneous across companies receiving an upgrade or downgrade in different months. Hence, this study offers an ideal setting to test the interaction-weighted estimator and see if the resulting treatment estimates are different to that of the event-study model.

7.1 Setting

The event study model estimating the effect of ESG rating changes on ESG ownership follows a specification similar to that of (6). I define the treatment D_{it} as ever having received an ESG rating change. Following the approach mentioned in section 3, I categorise the firms into cohorts E_i based on their initial rating change. In the MSCI sample, I identify 59 cohorts for upgrades and 56 cohorts for downgrades. But, in the Refinitiv sample, I identify 17 cohorts for upgrades and 15 cohorts for downgrades. As the analysis focuses on initial rating change, the cohort-specific average treatment effect estimates the treatment path for each cohort $e \in E_i$ against the counterfactual of never receiving a rating change. As mentioned earlier, it is important to check if the necessary assumptions are likely to hold in this setting.

Parallel Trends - Assuming fund managers solely rely on external data providers to assess a firm's ESG performance, it is plausible that the parallel trends assumption holds in this case. Since ESG concerns play a central role in the fund's strategy and fund manager's decisions, it is likely that the investment decision is partly driven by the ratings. Hence, in the absence of a rating change, fund managers would have responded similarly between the treated and untreated unit. Fund managers could possibly base their decisions on other concurrent confounding events such as ESG related controversies or media scandals about firms and likewise adjust their holdings. Even in this case they are responding to ESG related novel information that would eventually be reflected in the ratings.

No anticipatory behaviour - Since ESG rating changes are driven by proprietary ESG scoring methodology of the rating provider, it is plausible that there is no anticipatory behaviour before the rating change. A violation of this assumption is possible if funds have access to private information that helps them to predict rating changes. But with complex in-house scoring models and proprietary weighting schemes, it is highly unlikely that a fund accurately predicts a rating change and adjusts its behaviour prior to the actual rating change.

Treatment effect heterogeneity - The demand for socially responsible or sustainable investing has gained popularity in recent years. Likewise, over time there have been regulations pressing for more transparency on ESG-related data and corporate disclosures. Rating providers are also continuously expanding the horizon of metrics and issues that go into the formulation of ESG ratings. Hence, the rating changes for later-treated cohorts could possibly have more impact on holdings as compared to earlier cohorts since funds can make better data driven investments. Moreover, due to their potential of reducing downside risks, the impact of a firms rating change could vary based on the current macroeconomic scenario. For example, a firm shifting to renewable energy sources (which leads to a rating upgrade) during an energy crisis could attract more attention from funds as opposed to a crisis free scenario. Also, the size of a rating change could possibly have different impacts on ownership. A bigger shift in ratings, signalling a significant improvement or deterioration in ESG performance could be viewed differently to a smaller shift in ratings. Depending on the fund's strategy, various industry sectors may be emphasised differently. For example, an upgrade in ESG performance for an oil company may grab more attention from funds as opposed to a tech company.

7.2 Identification Strategy

First, I follow the same strategy as Berg, Heeb, and Kölbel (2022) and jointly estimate the effect of ESG rating upgrades and downgrades on ESG ownership. Then I use the alternative interaction-weighted (IW) estimator to obtain coefficients which are robust to heterogeneity and interpretable as causal estimates. In all regressions, ESG ownership is the dependent variable with ESG rating upgrades and downgrades as the independent variables.

7.2.1 FE Estimates

Dummy variables are used to indicate the calendar-time periods in which ESG rating upgrades (u_{it}) and downgrades (d_{it}) occur for each company i at a month t,

$$u_{it} = 1[t \in \{\nu_{i,1}, \ldots, \nu_{i,n}\}] \tag{12}$$

$$d_{it} = 1[t \in \{\delta_{i,1}, \ldots, \delta_{i,n}\}] \tag{13}$$

where $\{\nu_{i,1}, \ldots, \nu_{i,n}\}$ is a set of months in which company i received upgrades and $\{\delta_{i,1}, \ldots, \delta_{i,n}\}$ is a set of months in which company i received downgrades.
The event-study model utilises a dynamic two-way fixed effects regression of the following form:

$$\begin{aligned} y_{it} = \beta \sum_{l<\underline{l}} b^l_{it} \;+\; \sum_{\underline{l}}^{\overline{l}} \beta_l b^l_{it} \;+\; \beta \sum_{l>\overline{l}} b^l_{it} \;+ \\ \gamma \sum_{l<\underline{l}} c^l_{it} \;+\; \sum_{\underline{l}}^{\overline{l}} \gamma_l c^l_{it} \;+\; \gamma \sum_{l>\overline{l}} c^l_{it} \;+\; \alpha_i + \lambda_t + \epsilon_{it} \end{aligned} \tag{14}$$

Here, y_{it} refers to the level of ESG ownership for firm i in month t and coefficients of interest are denoted by β_l and γ_l for upgrades and downgrades respectively. The unobserved error term is given by ϵ_{it}, while α_i and λ_t are individual and time fixed effects respectively. b^l_{it} (for upgrades) and c^l_{it} (for downgrades) are binary variables indicating leads and lags of the rating changes within the treatment window from $\underline{l}$ periods prior to treatment and $\overline{l}$ periods after treatment. More specifically, they refer to an individual firm i being l periods away from treatment in month t. For a binned specification, the "distant" relative periods that exceed the lower threshold $\underline{l}$ or the upper threshold $\overline{l}$ of the treatment window are collectively indicated by the end lead and lag respectively.[18] Rather than including sparsely populated distant relative periods individually, researchers combine them into a single dummy variable.[19] The treatment

[18] See equation (8) in Sun and Abraham (2021, p.12-13) or Schmidheiny and Siegloch (2019, p.5) for more details on 'binned' specification.

[19] Schmidheiny and Siegloch (2019, p.10) suggests binning to avoid underidentification issues.

effect is assumed to be constant for periods outside the observed window (Schmidheiny and Siegloch, 2019, p.6). Therefore b_{it}^{l} and c_{it}^{l} are defined as follows,

$$b_{it}^{l} = \begin{cases} \sum_{s=t-\underline{l}+1}^{\bar{t}} \; u_{i,s} & \text{if} \quad l = \underline{l} - 1 \\ u_{i,t-l} & \text{if} \quad \underline{l} \leq l \leq \bar{l} \\ \sum_{s=\underline{t}}^{t-\bar{l}-1} \; u_{i,s} & \text{if} \quad l = \bar{l} + 1 \end{cases} \tag{15}$$

$$c_{it}^{l} = \begin{cases} \sum_{s=t-\underline{l}+1}^{\bar{t}} \; d_{i,s} & \text{if} \quad l = \underline{l} - 1 \\ d_{i,t-l} & \text{if} \quad \underline{l} \leq l \leq \bar{l} \\ \sum_{s=\underline{t}}^{t-\bar{l}-1} \; d_{i,s} & \text{if} \quad l = \bar{l} + 1 \end{cases} \tag{16}$$

The periods $\underline{t}$ and $\bar{t}$ denote the first and last calendar-time periods respectively. I observe a treatment effect window of $\underline{l} = 12$ months prior to a rating change and $\bar{l} = 24$ months after a rating change for each firm in the sample. The indicators for being one time period before treatment, b_{it}^{-1} and c_{it}^{-1} are excluded from the regression and the remaining coefficients are expressed in relation to these baseline periods. Hence, β_0 to β_{24} and γ_0 to γ_{24} calculate dynamic effect of rating changes on ESG ownership level for each of the 24 months following the treatment, in reference to it's level one period before treatment. All periods that fall outside the treatment effect window are binned into the end-points, $l = -13$ and $l = 25$.

7.2.2 IW Estimates

Similar to (9), I estimate an interacted specification of two-way fixed effects regression[20] of the following form,

$$\begin{aligned} Y_{it} = \; & \beta \sum_{l<-12} D_{it}^{l} \; + \sum_{e=\underline{e}}^{\bar{e}-1} \; \sum_{l=-12,\neq-1}^{24} \beta_{e,l} 1\{E_i = e\} D_{it}^{l} \; + \; \beta \sum_{l>24} D_{it}^{l} \\ & + \alpha_i + \lambda_t + \epsilon_{it} \end{aligned} \tag{17}$$

[20] The STATA package *eventstudyinteract* from Sun (2021a) is used for the interaction-weighted (IW) estimation.

for all $t \in \{\underline{t}, \ldots, \bar{t} - 1\}$. A weighted-average of the $\widehat{\beta_{e,l}}$'s and sample share of each cohort e (used as weights) equals the IW estimates. I use the last-treated cohort as control group and hence need to drop the last time period since every firm gets a rating change by then and there exists no control cohort for estimating $CATT_{e,l}$ in the last period. Effects in the bins at end points are assumed to be constant for each cohort. Analogously, I estimate regression (17) for downgrades as well, wherein $\gamma_{e,l}$ replaces $\beta_{e,l}$.[21]

For MSCI ESG rating upgrades, cohorts range from $\underline{e} = 1$ to $\bar{e} = 59$, months range from $\underline{t}$ = December 2013 to $\bar{t}$ = December 2020. Control cohort $C = 59$, the cohort which received rating change in the last period. Likewise, for MSCI ESG rating downgrades, cohorts range from $\underline{e} = 1$ to $\bar{e} = 56$, months range from $\underline{t}$ = December 2013 to $\bar{t}$ = December 2020. Control cohort $C = 56$, the cohort which received rating change in the last period. The last month December 2020 is dropped from the estimation. All coefficients are expressed in reference to the outcome one month prior to treatment. Hence, $l = -1$ is dropped from all regressions. Standard errors are clustered both at the firm and month level.

Additionally, I also implement a decomposition[22] of the pre-treatment coefficient β_{-2} to check for any pre-trend of rating upgrades. Similar to (7), I decompose β_{-2} as

$$\sum_{e=\underline{e}}^{\bar{e}} \omega_{e,-2}^{-2} CATT_{e,-2} + \sum_{l=\underline{l}\ ,\ \neq\{-1,-2\}}^{\bar{l}} \sum_{e=\underline{e}}^{\bar{e}} \omega_{e,l}^{-2} CATT_{e,l} + \sum_{l'=-1} \sum_{e=\underline{e}}^{\bar{e}} \omega_{e,l'}^{-2} CATT_{e,l'} \tag{18}$$

The weights are estimated in the same manner as discussed in Section(4.2). The weights on each cohort from it's own relative period $l = -2$ would sum to 1, $\sum_{e=\underline{e}}^{\bar{e}} \omega_{e,-2}^{-2} = 1$. The weights on other included relative periods would sum to 0 for each period, $\sum_{l=\underline{l}\ ,\ \neq\{-1,-2\}}^{\bar{l}} \sum_{e=\underline{e}}^{\bar{e}} \omega_{e,l}^{-2} = 0$. Finally, the weights on excluded relative period $l = -1$ would sum to -1, $\sum_{l'=-1} \sum_{e=\underline{e}}^{\bar{e}} \omega_{e,l'}^{-2} = -1$. Similar decomposition is done for γ_{-2} in the case of downgrades.

[21] The STATA package *eventstudyinteract* for IW estimation does not allow for a joint specification. Hence, I run the regressions for upgrades and downgrades separately.

[22] The STATA package *eventstudyweights* from Sun (2021b) is used to estimate the implied weights on $CATT_{e,l}$ for the event-study model.

8 Results

Similar to Berg, Heeb, and Kölbel (2022), I also find a significant long-term effect of MSCI ESG rating upgrades and downgrades on ESG ownership. However, the magnitudes of the effects are significantly lower when compared to the authors' results. The results from my replication show that on average, ESG ownership is only 2.45% higher two years after an upgrade, relative to it's level one month before the upgrade. Similarly, on average, ESG ownership is only 3.07% lower two years after a downgrade, relative to it's level one month before the downgrade. This is in contrast to earlier findings where upgrades had a relatively higher effect on ownership than downgrades with much larger magnitudes (17.1% for upgrades and 13.1% for downgrades). Figure 5. shows how ESG ownership increases after a rating upgrade and decreases after a rating downgrade.

During the period in which the treatment occurs (month 0) both upgrades and downgrades have a positive effect on ESG ownership. But, one period after treatment (month 1), the ESG ownership level is positive for rating upgrades while it is negative for rating downgrades. The initial 11 months after treatment seem to show a fairly steady response in the case of upgrades while one notices a mixed response for downgrades, with 3 month-lag having a slightly positive coefficient. But ESG ownership levels consistently increases (respectively decreases) from month 12 following a rating upgrade (respectively downgrade). This is similar to the original study wherein the increase in ownership is visible from the first month after an upgrade while for downgrades the authors report a significant decrease in ownership only from the seventh month onwards. The core inference, however, remains the same. The adjustment of ownership by funds seems to be a slow and gradual process within the first 2 years after rating changes. Likewise, the adjustment seems to be persistent because the post-treatment windows bin $b_{it}^{l=25}$ (for upgrades) and $c_{it}^{l=25}$ (for downgrades), which contains all periods that exceed 24 months, is significantly positive and negative respectively. Another similarity is that the post-treatment dynamic path for both upgrades and downgrades are nearly symmetrical. This symmetry is used as an argument against any confounding effect from negative media exposure. While negative

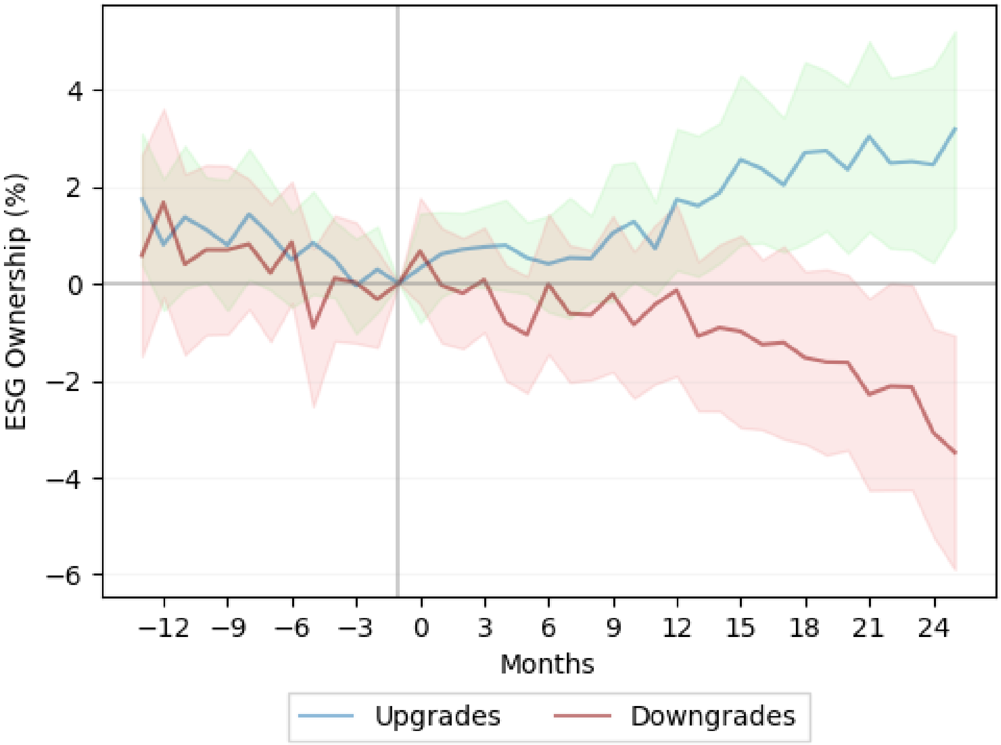
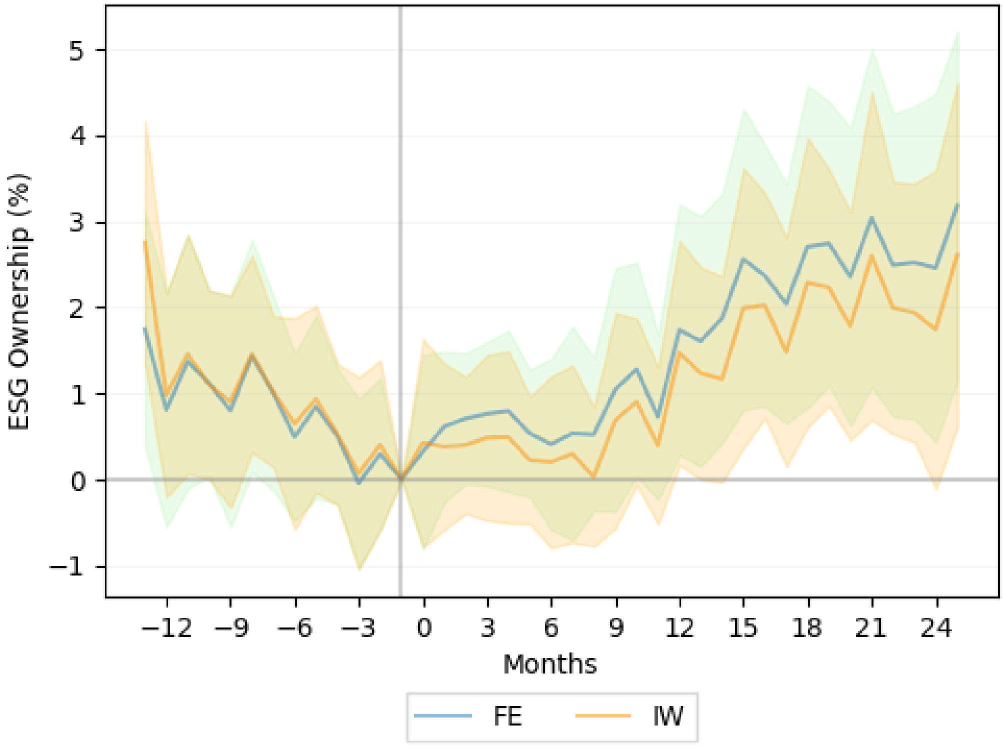


Figure 5: **The reaction of ESG ownership to ESG rating changes.**
This event study plot shows coefficients from a two-way fixed effects regression of ESG ownership (dependent variable) on up and downgrades of MSCI ESG ratings (treatments/events) as in (14). The observation period is from December 2013 to December 2020. Coefficients on relative period (months) dummy variables from 12 months prior to the event (leads) and 24 months after the event (lags) are shown here. The end lead (month = -13) and end lag (month = 25) bin all periods beyond -12 and 24 months respectively. Coefficients are normalized to the baseline level (i.e., the average level of ESG ownership 1 month before a rating change). The shaded region shows 95% confidence intervals which are based on standard errors clustered at firm and month level.

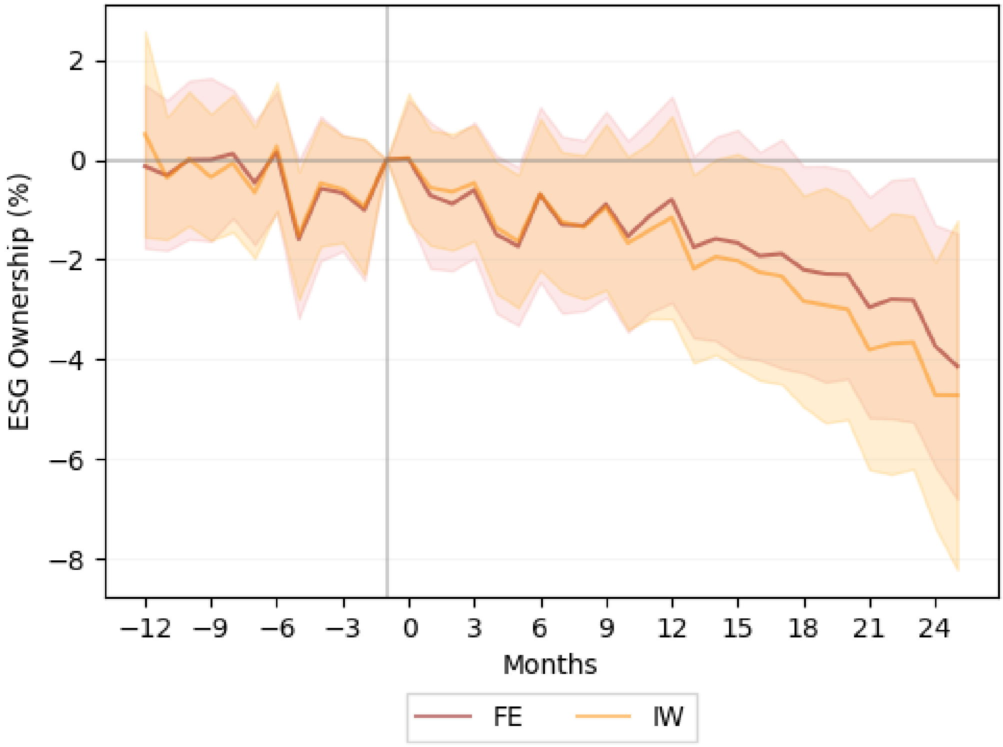
news may simultaneously influence ESG rating downgrades and ESG ownership levels, it is highly unlikely to increase ownership via upgrades (Berg, Heeb, and Kölbel, 2022, p.16).

However, it is difficult to rule out the existence of pre-treatment trends in ESG ownership, which is essential in establishing the relationship as a causal one. The ESG ownership levels are nearly zero upto 4 months before a rating upgrade or downgrade. Although there are deviations from zero in months 5-12 before a rating change, it could be attributed to the fact that these leads are far from the treatment period, making them prone to influences from omitted variables. It may also be that only firms which receive rating changes in later periods have pre-treatment periods extending up to 12 months.(Clarke and Tapia-Schythe, 2021, p.16-17). Such issues could arise since the panel dataset cannot be balanced both in calendar-time and relative-time simultaneously.

Pre-treatment coefficients for upgrades show a slightly downward sloping trend between leads -12 and -5. For downgrades, pre-treatment coefficients decreases from leads -12 to -5 while being more or less constant between leads -11 and -6. Commonly used pre-trend tests, although not ideal, would require these coefficients to be nearly zero. This may indicate that ESG ownership is influenced by variables other than ESG rating changes. For example, downgrades in particular could be influenced by negative news that can reduce the ownership of such stocks with poor ESG-related media exposure. It is possible that fund managers study characteristics other than ESG ratings before making an investment decision. Hence, they may divest from poorly managed firms in advance due to broader risk concerns. Although poor management could lead to a worse ESG rating, it would only be published with a time-lag and a discretionary fund manager may prefer to divest earlier. This behaviour may be reflected in the downward trend in ownership during the pre-treatment period for rating downgrades. Moreover, firms with high ESG ratings may showcase poor financial performance. This could be a plausible reason for funds to divest from firms that may be on track to receive rating upgrades in future. This behaviour may be reflected in the downward trend in ownership during pre-treatment periods for upgrades. Besides, it



(a) For MSCI rating upgrades



(b) For MSCI rating downgrades

Figure 6: **FE vs IW estimates for the effect of rating changes on ESG ownership**
Each figure plots the FE estimates $\widehat{\beta}_l$ from regression (14) and IW estimates $\widehat{\nu}_l$ from regression (17) for each relative period (month) l and shaded region showing 95% confidence interval. Both specifications estimate effect of rating changes on ESG ownership at l. The end lead (month = -13) and end lag (month = 25) bins all periods beyond -12 and 24 months respectively. Coefficients are normalized to the baseline level (i.e., the average level of ESG ownership 1 month before a rating change). Standard errors are clustered at firm and month level.

could also be that the coefficients are in fact contaminated by effects from other relative periods. Overall, this is different to previous findings wherein the authors did not find any evidence for pre-trends.

The interaction-weighted (IW) estimates, while tracing a dynamic path similar to the fixed effects (FE) model, shows a smaller effect of ESG rating upgrades on ESG ownership. Figure 6a. shows how the estimates differ in their magnitudes between the two regression specifications. For example, the 24 month-lag in the fixed effects model has an effect of 2.45% compared to baseline while the IW estimate is only 1.73% higher compared to baseline. This may plausibly indicate the problem of contamination of TWFE estimates by the effects from other relative periods that are included and excluded from the TWFE specification. It could be that the treatment effect in a particular month by itself is small but gets biased upwards by effects from other months. This depends on the weighting that is applied by the two-way fixed effects estimator on each cohort-average treatment effect on treated ($CATT_{e,l}$). For example, negative weights multiplied by a negative $CATT_{e,l}$ would result in a positive estimate, even when the underlying treatment effect could be negative.

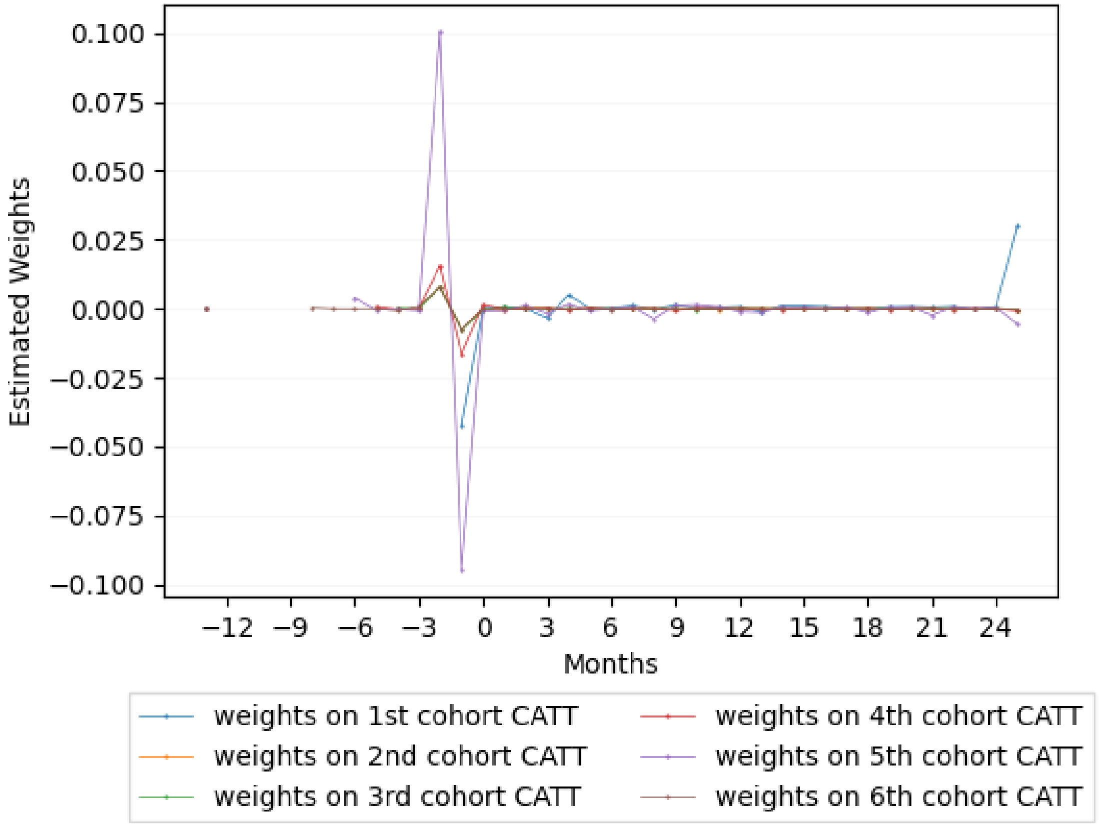
On the other hand, Figure 6b. shows how the IW estimates for rating downgrades are very similar in magnitude to the fixed effects estimates, with more negative estimates from 10 month-lag onwards. This may indicate an upward bias for fixed effects estimates in the case of downgrades as well. It is important to note that the goal of using IW estimates is not to look for deviations from the fixed effects estimates, but to produce estimates with reasonable weights, making them more interpretable as causal. Hence, even with similar estimates, it is important to check if the TWFE estimates suffer from non-zero and non-convex weighting. This would affect the interpretability of these estimates depending on the extent of treatment effect heterogeneity. To visualize the role of weights, I follow Sun and Abraham (2021) and focus on a single pre-treatment coefficient. Interestingly, a pre-trend test using decomposition of the pre-treatment estimate β_{-2} for upgrades and γ_{-2} for downgrades shows that it is sensitive to estimates from other periods. Figure 7. shows how certain post-treatment months have non-zero weights which could prevent β_{-2} (respectively γ_{-2}) from accurately isolating any pre-trend effect in the case of upgrades (respectively downgrades).

However, with non-negative weights that sum upto 1, the IW estimates are more interpretable as the average effect of treatment on the treated l periods from initial treatment. The IW estimates fall within the convex hull of it's underlying $CATT_{e,l}$ estimates and hence remain unaffected by estimates from other periods. Table 2 shows how the weights have properties as discussed in (18).

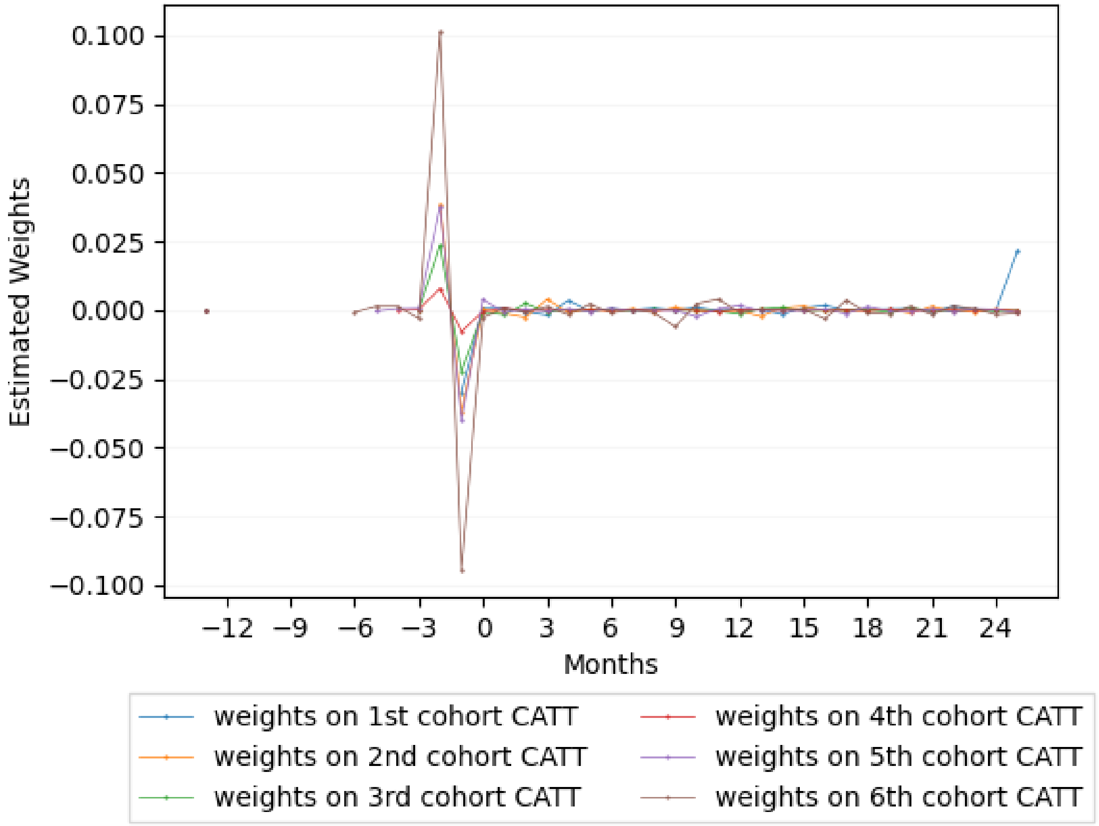
9 Robustness Checks

9.1 Excluding Index funds and ETF's

To confirm that the effect is not driven by certain funds within the sample, I re-estimate the regressions for a subgroup of the sample. The objective is to exclude passively managed index funds and ETF's which are specifically modeled to track an ESG related index. Hence, they are more likely to shift their holdings on the basis of ESG or sustainability metrics. Fund names containing keywords *'index'* and *'ETF'* are excluded from the overall list of funds. For rating upgrades, the response pattern from funds remain similar to the main results. But, there seems to be no consistent response for rating downgrades. The pre-treatment coefficients also do not showcase any noticeable trend. Figure 8. shows how MSCI rating upgrades lead to an increase in ESG ownership even after excluding passively managed index funds and ETF's. On average, two years after an upgrade, ESG ownership is 2.24% higher compared to baseline, which is only 0.21 percentage points smaller compared to results including all funds. For downgrades, ESG ownership on average is 0.86% lower compared to baseline, which is 2.21 percentage points smaller compared to results including all funds. This provides some evidence that passively managed funds are not entirely driving the results, at least in the case of rating upgrades. Berg, Heeb, and Kölbel (2022) do not analyse this effect of active against passive funds in their study. I also find that changing the length of the panel affects the ESG ownership estimates. It is possible that the length of the panel affects the weights applied to individual 2x2 estimates which in turn affects the TWFE staggered estimate (Baker, Larcker, and Wang, 2022, p.8).



(a) Estimated weights $\omega_{e,l}^{-2}$ underlying β_{-2} for MSCI rating upgrades



(b) Estimated weights $\omega_{e,l}^{-2}$ underlying γ_{-2} for MSCI rating downgrades

Figure 7: **Estimated weights for checking pre-trend of rating changes.** This figure plots the estimated weights $\omega_{e,l}^{-2}$ associated with each $CATT_{e,l}$ in the linear combination produced by a fixed effects estimator as in equation (18). Only the first 6 cohorts are shown here.

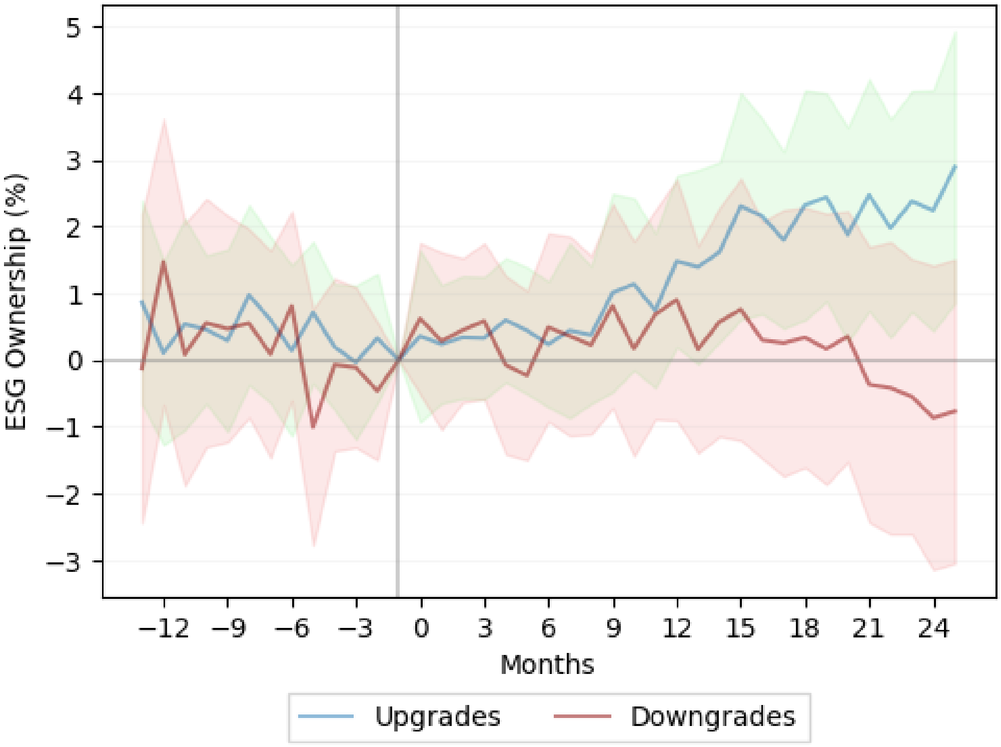


Figure 8: **The reaction of ESG ownership to ESG rating changes excluding index funds and ETF's.**

This event study plot shows coefficients from a two-way fixed effects regression of ESG ownership (dependent variable) on up and downgrades of MSCI ESG ratings (treatments/events) as in (14). The observation period is from December 2013 to December 2020. It uses a smaller sample of funds that does not include index funds and ETF's. Coefficients on relative period (months) dummy variables from 12 months prior to the event (leads) and 24 months after the event (lags) are shown here. The end lead (month = -13) and end lag (month = 25) bins all periods beyond -12 and 24 months respectively. Coefficients are normalized to the baseline level (i.e., the average level of ESG ownership 1 month before a rating change). The shaded region shows 95% confidence intervals which are based on standard errors clustered at firm and month level.

9.2 Using ESG ratings from Refinitiv

Additionally, I use ESG ratings data from Refinitiv, to check if similar results hold for other ESG data providers in the market. Estimating a joint specification as in (14) using Refinitiv data leads to a different set of results. Interestingly, Refinitiv ESG rating upgrades and downgrades do not have a consistent positive or negative effect on ESG ownership respectively. On average, two years after a rating upgrade, ESG ownership is only 1.47% higher compared to one month before the rating change. On the other hand, two years after a rating downgrade, ESG ownership is only 0.24% lower than its level one month before the rating change. Figure 9. shows how ESG ownership responds to Refinitiv ESG rating changes. Unlike MSCI ratings, the dynamic path for Refinitiv ratings does not consistently rise or fall following a rating change. Both upgrades and downgrades do not generate a consistent response pattern from funds.

9.3 ESG weight - An alternative measure of ownership

Similar to Berg, Heeb, and Kölbel (2022), I also check if my results are robust to using an alternative measure of ESG mutual funds' ownership of companies. For this, the dependent variable ESG ownership is replaced by a new measure called ESG weight. ESG weight is defined as "a company's market capitalization that is held by ESG mutual funds in a given month, divided by the total market capitalization held by ESG mutual funds in the same month." The rationale behind this measure is that the ownership levels are not directly affected by the increasing volume of assets owned by ESG mutual funds. (Berg, Heeb, and Kölbel, 2022, p.23). The share of market capitalization of companies owned by the ESG funds is calculated using stock price data from the CRSP Monthly Stock database. Figure 10. shows the response of ESG weight against MSCI ESG rating upgrades and downgrades. The results show that ESG rating upgrades lead to a consistent increase in ESG weight over 24 months following an upgrade. But surprisingly, ESG weight does not decrease consistently after a rating downgrade. It is almost close to zero and often positive during the 24 month period. However, for downgrades, ESG weight starts to decline from lag 17 onwards.

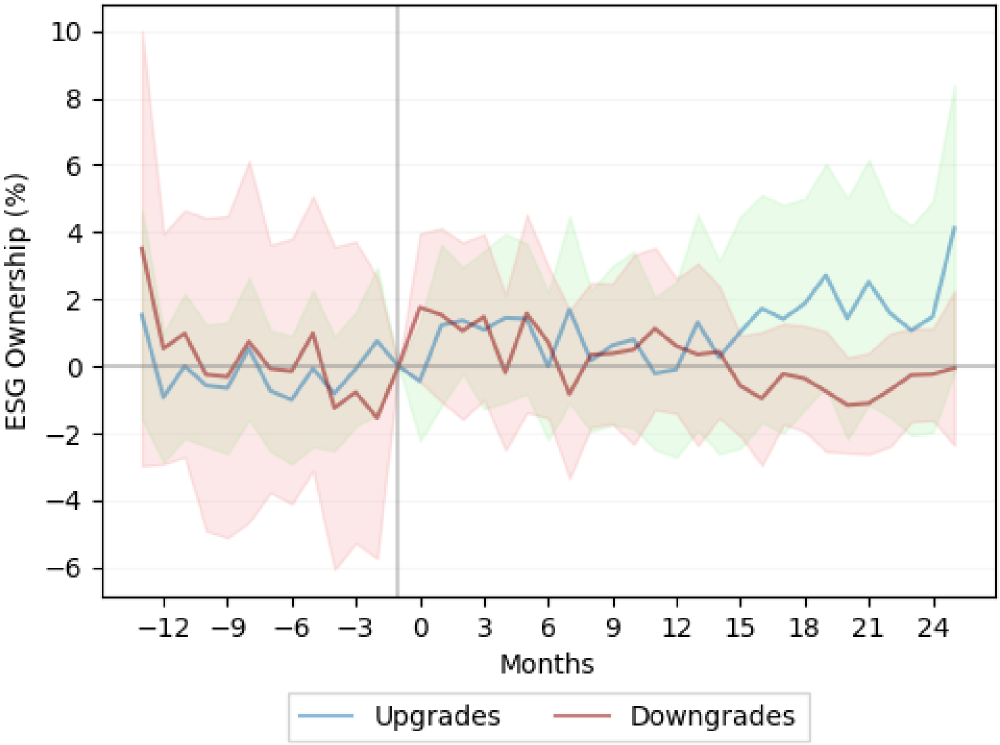


Figure 9: **The reaction of ESG ownership to Refinitiv ESG rating changes.**
This event study plot shows coefficients from a two-way fixed effects regression of ESG ownership (dependent variable) on up and downgrades of Refinitiv ESG ratings (treatments/events) as in (14). The observation period is from February 2013 to September 2020. Coefficients on relative period (months) dummy variables from 12 months prior to the event (leads) and 24 months after the event (lags) are shown here. The end lead (month = -13) and end lag (month = 25) bins all periods beyond -12 and 24 months respectively. Coefficients are normalized to the baseline level (i.e., the average level of ESG ownership 1 month before a rating change). The shaded region shows 95% confidence intervals which are based on standard errors clustered at firm and month level.

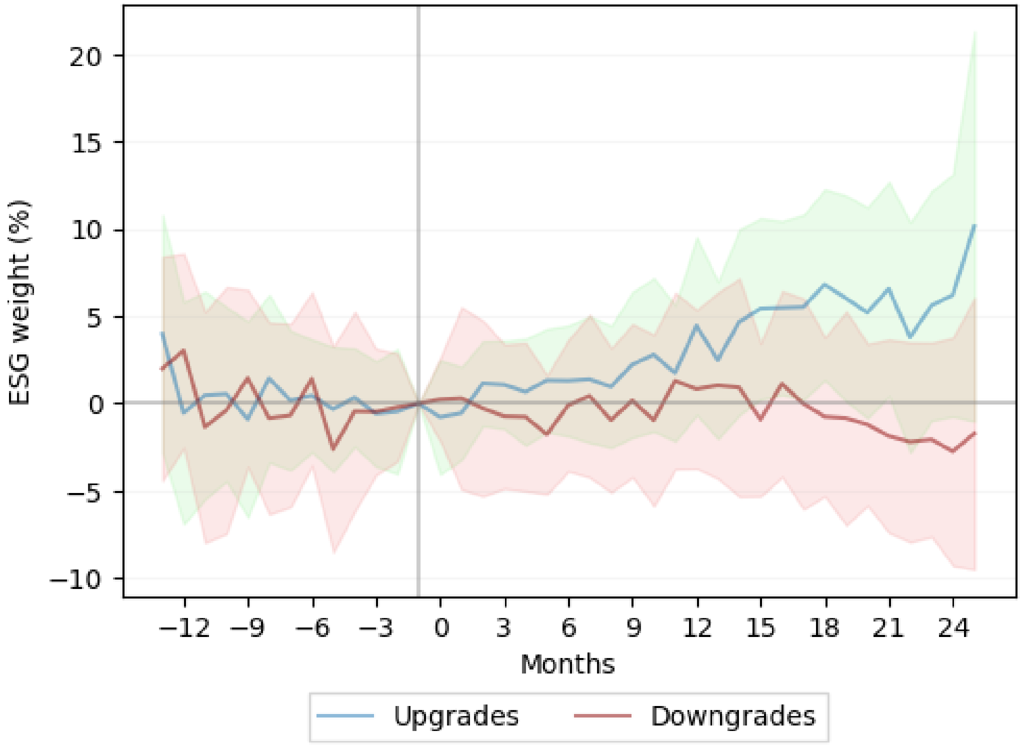


Figure 10: **The reaction of ESG weight to MSCI ESG rating changes.**
This event study plot shows coefficients from a two-way fixed effects regression of ESG weight (dependent variable) on up and downgrades of MSCI ESG ratings (treatments/events). The observation period is from December 2013 to December 2020. Coefficients on relative period (months) dummy variables from 12 months prior to the event (leads) and 24 months after the event (lags) are shown here. The end lead (month = -13) and end lag (month = 25) bins all periods beyond -12 and 24 months respectively. Coefficients are normalized to the baseline level (i.e., the average level of ESG ownership 1 month before a rating change). The shaded region shows 95% confidence intervals which are based on standard errors clustered at firm and month level.

This is in contrast to the original study wherein the authors found similar results using both *ESG ownership* and *ESG weight* as dependent variable.

10 Caveats

There are a few caveats associated with this replication study. First, this is not a one to one replication of Berg, Heeb, and Kölbel (2022). The reason being that the original study is not yet published and therefore the data samples and code are not available publicly for replication purposes. While I have followed the exact same steps as the authors in constructing the datasets, there may be substantial differences in the sample of ESG mutual funds and sample of firms with MSCI ESG ratings used for the study. The sample size of the final dataset I use for replication is much smaller compared to the one used in the original study. Also, the ESG ratings data used in this replication are legacy data. Hence, this may not account for updates in methodology and data corrections such as in the latest iteration of the rating providers. The difference in magnitudes of the event-study estimates between this replication and the original study could be attributed to differences in data samples.[23]

Second, the authors avoid ESG data from Refinitiv citing concerns regarding the consistency of historical data from the provider (Berg, Fabisik, and Sautner, 2020). While my results using Refinitiv data confirm that fund managers do not consistently respond to Refinitiv rating changes, it must be interpreted with caution. It is also important to note that all results only hold for U.S. domestic equity mutual funds and it may or may not produce similar results for other geographies.

Third, I also lose a substantial sample size when trying to construct a balanced panel dataset wherein each firm has an ESG rating for each month during the entire period from 2013 to 2020.

[23] An important reason for the difference in results between this replication and the original study could be due to differences in the way the estimates are visualized. While I plot the coefficients directly from the event-study models, the authors show them as percentage points of the baseline levels. Since their code is not published yet, it is difficult to understand any post-estimation calculations done by the authors. An explanation for calculating baseline estimates for upgrades and downgrades could not be found in the original study. Hence, I show the event-study coefficients as they are. Nevertheless, the main conclusions from both studies would roughly be the same since the difference is only in the way the estimates are expressed.

Most importantly, while the interaction-weighted (IW) estimator offers a robust alternative estimation strategy, this replication uses the last-treated cohort (rather than never-treated or not-yet-treated units) as comparison group. In this study, the last-treated cohorts represent 2.22% of the total sample for both upgrades and downgrades. For comparison, the largest treated cohorts represent 10.37% of the total sample for both upgrades and downgrades. Therefore, it is possible that the size of the last-treated cohort can influence the resultant estimates, leading to an efficiency loss in case of a smaller-sized last-treated cohort (Roth and Sant'Anna, 2021, p.27).

11 Discussion

The main result of this replication study reconfirms the authors' finding that ESG ratings do have a significant economic impact. It is clear that ESG ratings offer valuable signals which influence the investment decisions of mutual fund managers. While the general directions of the impact remain consistent (ownership rising after upgrades and falling after downgrades), the extent of this impact seems to vary across data samples and ESG data providers.[24] The robustness checks reveal that it is also important to distinguish the funds based on their management styles. This helps us understand whether the pattern of impact seen here is mainly driven by managers who actively base their investments on rating changes or if it is merely an effect of funds mirroring the holdings of an ESG based index. The effect of rating upgrades on ESG ownership holds for both actively and passively managed funds but the response pattern becomes inconsistent for rating downgrades. While the authors select MSCI ESG ratings based on it's high correlation to ESG ownership, they do not run the event-study model for other ESG data providers. Based on the results using Refinitiv data, I am able to provide some evidence that the event-study analysis using data from an alternate ESG source does not produce similar results.

The results also provide evidence that using a standard event-study model for this particular setting with staggered treatment timing and treatment effect heterogeneity

[24] The latest version of the authors' findings shows that MSCI ESG rating has the strongest relationship with the holdings of ESG funds in the U.S.

may not produce reliable causal estimates. The decomposition of the pre-treatment estimate shows that it is influenced by treatment effects from other relative periods via non-convex and non-zero weighting. However, this application uses the last-treated cohort as comparison group and hence one cannot infer how the estimates would be when using not-treated or not-yet-treated cohorts as comparison. The results show that the interaction-weighted estimates could differ in magnitudes depending on the event (upgrades or downgrades), but by construction produce causal estimates that are robust to treatment effect heterogeneity. It is also important to note that the results using the alternative interaction-weighted estimator holds under the assumptions of parallel trends and no anticipation behaviour.

Additionally, I find that using ESG weight instead of ESG ownership as dependent variable produces similar results for rating upgrades. This acts as a check against concerns that the ESG ownership variable could be merely capturing the recent growth in the market for ESG investments, especially in the case of increasing ownership following rating upgrades. Hence, the results using ESG weight reconfirm to a certain extent that funds increase a companies' weight in their portfolio after a rating upgrade while there is no consistent response after a rating downgrade.

12 Conclusion

This paper replicates Berg, Heeb, and Kölbel (2022), with a focus on studying the effect of ESG rating changes on ownership levels of ESG mutual funds in the U.S. Moreover, this study leverages the staggered difference-in-differences setting to test the efficiency of event-study models and more recent alternative econometric methods in estimating dynamic treatment effects. Following Sun and Abraham (2021), I implement a decomposition of two-way fixed effects estimates into a linear combination of cohort-specific average treatment effects on the treated ($CATT_{e,l}$) and their corresponding weights. I show that the non-convex and non-zero nature of these weights along with treatment effect heterogeneity across cohorts, lead to contamination of the coefficients with treatment effects from other relative periods.

As a solution to this, I analyse the event-study model using the Interaction-weighted (IW) estimator as proposed in Sun and Abraham (2021). With the IW estimator, each coefficient estimates a convex average of $CATT_{e,l}$ with the sample shares of each cohort e as weights. More importantly, I obtain estimates that are interpretable as the average effect of treatment on the treated l periods from initial rating upgrade or downgrade.

This study provides greater validity to the original findings of the authors' and reconfirms the response pattern of ESG mutual funds to ESG rating changes. The results also confirm that the extent of this impact is quite low and ESG rating changes are only of interest to a small group of mutual funds with an ESG mandate. Through this research, I am able to apply latest econometric methods within a relevant setting like assessing the economic impact of ESG ratings. This helps re-examine any violations of assumptions and it's effect on estimates underlying a two-way fixed effects regression in an event-study model. It also identifies potential contamination caused by a fixed effects model under the common assumption of treatment effect homogeneity. Besides the econometric relevance, this paper offers further evidence that ESG ratings can influence investment decisions and the extent of this influence could increase with the expansion of the market for ESG based funds.

The original study also covers other aspects in indetifying the impact of ESG ratings. A replication of these results is a promising avenue for future research which can strengthen the argument for ESG ratings and their impact. Implementing the analysis using other alternative econometric models such as in Callaway and Sant'Anna (2021) or Roth and Sant'Anna (2021) would also help in understanding how selection of comparison groups could affect the estimates.

While there is substantial evidence on the divergence of ESG ratings across different providers (for example Berg, Koelbel, and Rigobon (2022)), more research is needed to understand the variation in the impact of these ratings across providers. This is important in today's world where investors and asset managers are getting increasingly interested in integrating ESG metrics into their investment analysis and decisions.

References

Angrist, Joshua D, and Jörn-Steffen Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives* 24 (2): 3–30.

Athey, Susan, and Guido W Imbens. 2022. "Design-based analysis in difference-in-differences settings with staggered adoption." *Journal of Econometrics* 226 (1): 62–79.

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144 (2): 370–395.

Beck, Thorsten, Ross Levine, and Alexey Levkov. 2010. "Big bad banks? The winners and losers from bank deregulation in the United States." *The Journal of Finance* 65 (5): 1637–1667.

Berg, Florian, Kornelia Fabisik, and Zacharias Sautner. 2020. "Rewriting history II: The (un) predictable past of ESG ratings." *European Corporate Governance Institute–Finance Working Paper* 708 (2020): 10–2139.

Berg, Florian, Florian Heeb, and Julian F Kölbel. 2022. "The Economic Impact of ESG Rating Changes." *Available at SSRN 4088545.*

Berg, Florian, Julian F Koelbel, and Roberto Rigobon. 2022. "Aggregate confusion: The divergence of ESG ratings." *Review of Finance* 26 (6): 1315–1344.

Borusyak, Kirill, and Xavier Jaravel. 2017. "Revisiting event study designs." *Available at SSRN 2826228.*

Broda, Christian, and Jonathan A Parker. 2014. "The economic stimulus payments of 2008 and the aggregate demand for consumption." *Journal of Monetary Economics* 68:S20–S36.

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

Chen, Zhongfei, and Guanxia Xie. 2022. "ESG disclosure and financial performance: Moderating role of ESG investors." *International Review of Financial Analysis* 83:102291.

Clarke, Damian, and Kathya Tapia-Schythe. 2021. "Implementing the panel event study." *The Stata Journal* 21 (4): 853–884.

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

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J Notowidigdo. 2018. "The economic consequences of hospital admissions." *American Economic Review* 108 (2): 308–52.

Dube, Arindrajit, T William Lester, and Michael Reich. 2010. "Minimum wage effects across state borders: Estimates using contiguous counties." *The review of economics and statistics* 92 (4): 945–964.

Fauver, Larry, Mingyi Hung, Xi Li, and Alvaro G Taboada. 2017. "Board reforms and firm value: Worldwide evidence." *Journal of Financial Economics* 125 (1): 120–142.

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

Hoynes, Hilary, Diane Whitmore Schanzenbach, and Douglas Almond. 2016. "Long-run impacts of childhood access to the safety net." *American Economic Review* 106 (4): 903–34.

Imai, Kosuke, and In Song Kim. 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis* 29 (3): 405–415.

Lakkis, Emil. 2022. "Real Effects of ESG Investing." *Available at SSRN 4239243.*

Marcus, Michelle, and Pedro HC Sant'Anna. 2021. "The role of parallel trends in event study settings: an application to environmental economics." *Journal of the Association of Environmental and Resource Economists* 8 (2): 235–275.

Meer, Jonathan, and Jeremy West. 2016. "Effects of the minimum wage on employment dynamics." *Journal of Human Resources* 51 (2): 500–522.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights* 4 (3): 305–22.

Roth, Jonathan, and Pedro HC Sant'Anna. 2021. "Efficient estimation for staggered rollout designs." *arXiv preprint arXiv:2102.01291.*

Schmidheiny, Kurt, and Sebastian Siegloch. 2019. "On event study designs and distributed-lag models: Equivalence, generalization and practical implications."

Stevenson, Betsey, and Justin Wolfers. 2006. "Bargaining in the shadow of the law: Divorce laws and family distress." *The Quarterly Journal of Economics* 121 (1): 267–288.

Strezhnev, Anton. 2018. "Semiparametric weighting estimators for multi-period differencein-differences designs." In *Annual Conference of the American Political Science Association, August,* vol. 30.

Sun, Liyang. 2021a. "EVENTSTUDYINTERACT: Stata module to implement the interaction weighted estimator for an event study."

———. 2021b. "EVENTSTUDYWEIGHTS: Stata module to estimate the implied weights on the cohort-specific average treatment effects on the treated (CATTs)(event study specifications)."

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199.

13 Appendix

13.1 Robust to Multicollinearity

When using a sample balanced in calendar time, Sun and Abraham (2021, p.27) recommend excluding at least two relative period indicators to avoid multicollinearity. Hence, I redo the event-study analysis after excluding relative period -12, in addition to -1. The results are similar to those from the main analysis. Figure 11 shows the reaction of ESG ownership to rating changes without multicollinearity effect.

Now, on average, ESG ownership is 2.14% higher two years after an upgrade, relative to it's level one month before the upgrade. Similarly, ESG ownership is 3.73% lower two years after a downgrade, relative to it's level one month before the downgrade. Pre-treatment coefficients for both upgrades and downgrades are smaller. Moreover, ownership levels during the period of treatment (i.e., month 0) is nearly zero for upgrades and downgrades. This is in contrast to the main results where ownership levels for both upgrades and downgrades where positive during the treatment period and thereafter it diverges.

The IW estimates for rating upgrades and downgrades trace a similar treatment path to that of the FE estimates. For upgrades, the IW estimates are much more closer to the FE estimates while it is more negative (from lag 10 onwards) for rating downgrades.

13.2 Results using unbalanced panel

I analyse the event-study model using an unbalanced panel with a larger sample size. Table 3 presents the summary statistics on the unbalanced panel. The TWFE and IW estimates produce a similar pattern to that of the main results. But, the magnitudes of the TWFE estimates, two years after a rating change are smaller in comparison to main results. The pre-trends seem to differ from earlier findings with values now closer to zero. Hence, using a balanced or unbalanced sample does not seem to affect the results greatly.

13.3 ESG ownership and nature of rating changes

Table 4 partly replicates the regressions in Table 2 of Berg, Heeb, and Kölbel (2022). First, I check if the effect on ESG ownership changes over time. A dummy variable *Post 2016* indicates all rating changes that occur after the year 2016. Similar to the authors', I show that the interaction of *post 2016* and the 12 month-lag for downgrades does not correlate with ownership. However, I notice that the interaction of *post 2016* and 12 month-lag for upgrades has a significant effect on ownership at 10 percent level. Hence, it is plausible that the treatment effect changes over time.

Second, I check if the size of the numerical score changes have any effect on ESG ownership. The dummy variable *High ESG score change* indicates changes in underlying numerical score which are greater than or equal to the median of all numerical score changes. While the interaction of 12 month-lag for downgrades and *High ESG score change* does not correlate with ownership, I find a significant correlation at 1 percent level for upgrades. Hence, it is difficult to rule out that the size of numerical score changes does not affect ownership levels. These two findings provide more evidence that treatment effect heterogeneity is plausible in this particular setting.

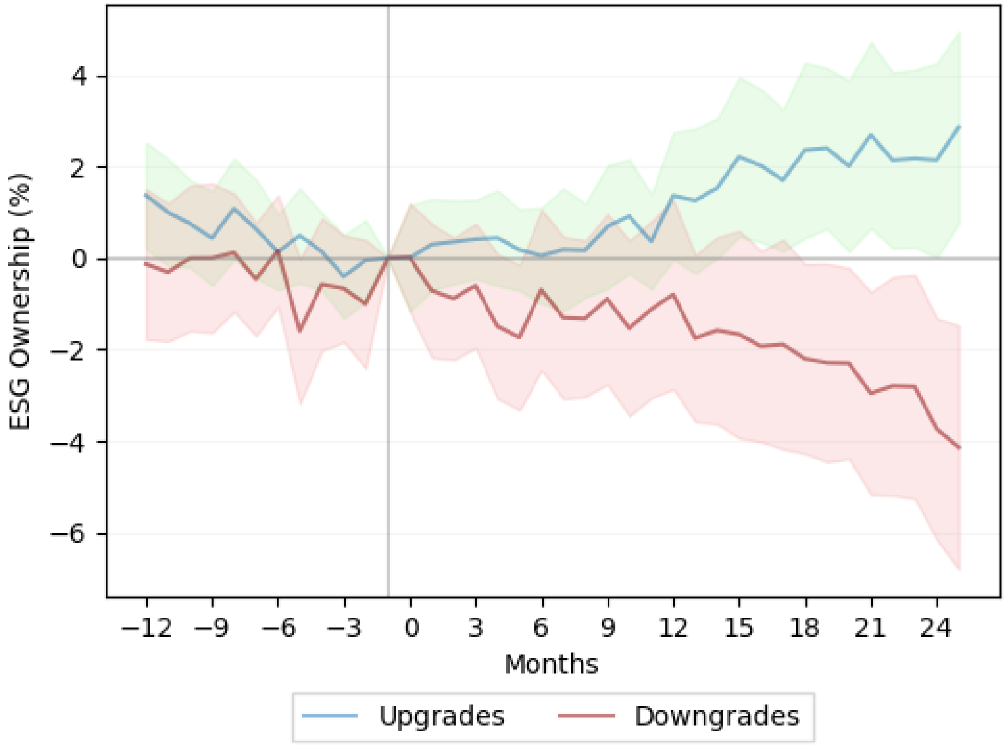
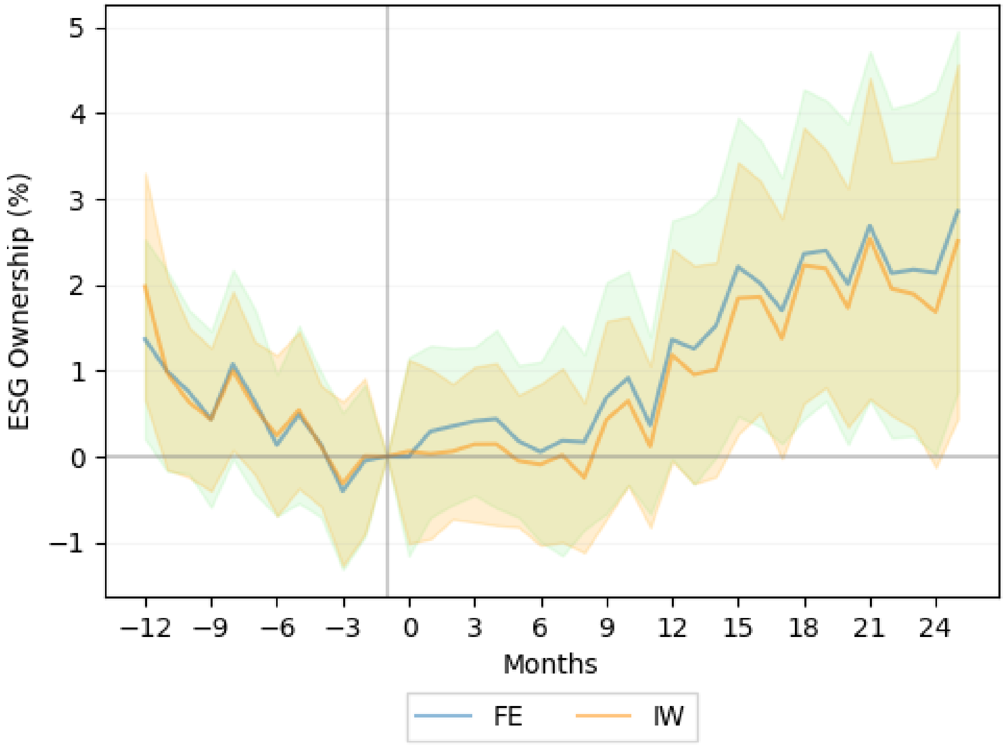
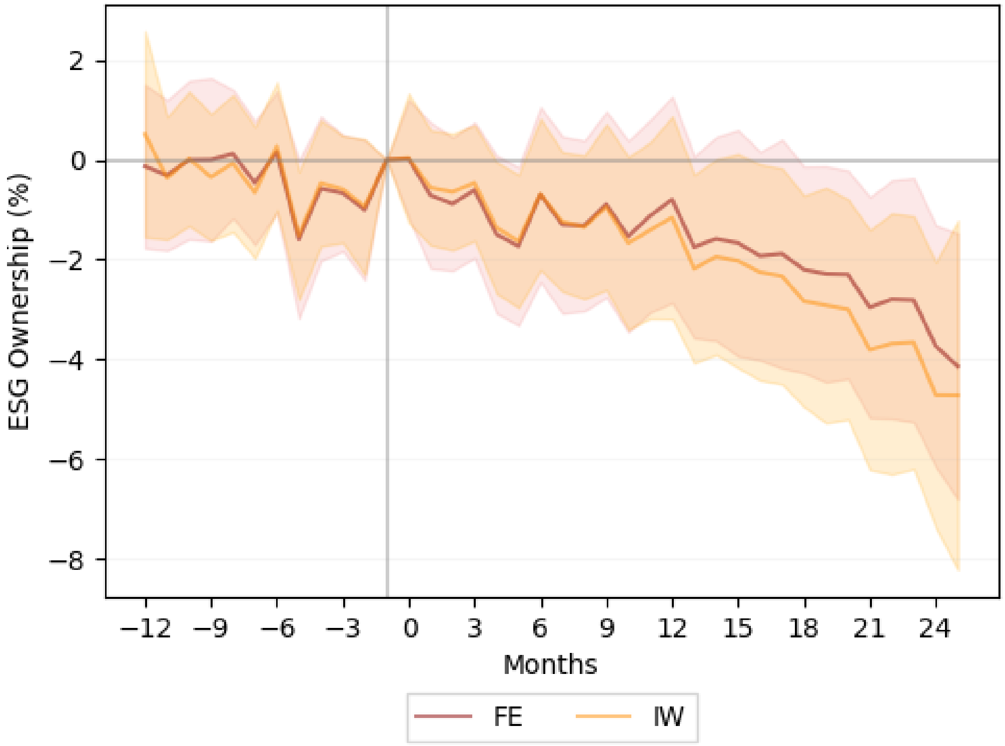


Figure 11: **Robust to Multicollinearity effect**
This figure shows the results from the event-study analysis after excluding two relative periods to avoid multicollinearity. I exclude periods -12 and -1. Here, the period -12 represents the bin which includes all periods before it. The observation period is from December 2013 to December 2020. The end lead (month = -12) and end lag (month = 25) bins all periods beyond -12 and 24 months respectively. Coefficients are normalized to the baseline level (i.e., the average level of ESG ownership 1 month before a rating change). The shaded region shows 95% confidence intervals which are based on standard errors clustered at firm and month level.



(a) For MSCI rating upgrades



(b) For MSCI rating downgrades

Figure 12: **Robust to Multicollinearity effect**
This figure shows the results from the event-study analysis after excluding two relative periods to avoid multicollinearity. I exclude periods -12 and -1. Here, the period -12 represents the bin which includes all periods before it. The end lead (month = -12) and end lag (month = 25) bins all periods beyond -12 and 24 months respectively. Coefficients are normalized to the baseline level (i.e., the average level of ESG ownership 1 month before a rating change). The shaded region shows 95% confidence intervals which are based on standard errors clustered at firm and month level.

Table 2: Properties of weights on fixed effects estimates

Relative Period	-3	-2	-1	0	1
Cohort 1	NaN	NaN	-0.0427	0.0002	0.0006
Cohort 2	0.0004	0.0078	-0.0076	-0.0001	-0.0007
Cohort 3	-0.0001	0.0079	-0.0076	-0.0007	0.0008
Cohort 4	0.0000	0.0157	-0.0165	0.0016	-0.0001
Cohort 5	-0.0007	0.1006	-0.0947	-0.0009	-0.0007
Cohort 6	0.0008	0.0078	-0.0076	0.0001	-0.0002
Cohort 7	-0.0001	0.0157	-0.0149	-0.0004	0.0004
Cohort 8	0.0004	0.0231	-0.0221	-0.0002	0.0000
Cohort 9	0.0002	0.0078	-0.0076	0.0001	-0.0003
Cohort 10	-0.0001	0.0079	-0.0075	-0.0003	0.0001
Cohort 11	0.0003	0.0380	-0.0372	0.0005	0.0003
Cohort 12	-0.0009	0.0240	-0.0223	0.0002	-0.0001
Cohort 13	0.0001	0.0080	-0.0075	-0.0001	-0.0001
Cohort 14	0.0001	0.0078	-0.0077	0.0000	0.0001
Cohort 15	-0.0001	0.0233	-0.0225	0.0001	0.0000
Cohort 16	-0.0002	0.0159	-0.0150	-0.0000	0.0001
Cohort 17	0.0001	0.0079	-0.0076	0.0001	-0.0001
Cohort 18	0.0001	0.0079	-0.0075	-0.0001	0.0001
Cohort 19	0.0001	0.0155	-0.0150	0.0001	-0.0003
Cohort 20	-0.0001	0.0080	-0.0075	-0.0002	0.0001
Cohort 21	0.0002	0.0232	-0.0225	-0.0000	0.0002
Cohort 22	-0.0003	0.0159	-0.0151	0.0001	-0.0003
Cohort 23	0.0001	0.0158	-0.0150	-0.0004	0.0002
Cohort 24	-0.0000	0.0080	-0.0077	0.0001	0.0001
Cohort 25	0.0002	0.0309	-0.0298	0.0002	-0.0007
Cohort 26	-0.0003	0.0160	-0.0150	-0.0004	0.0005
Cohort 27	0.0001	0.0080	-0.0077	0.0002	-0.0001
Cohort 28	0.0002	0.0309	-0.0293	-0.0005	0.0003
Cohort 29	0.0005	0.0156	-0.0150	-0.0001	0.0001
Cohort 30	-0.0001	0.0079	-0.0076	0.0000	-0.0002

Table 2: continued

Relative Period	-3	-2	-1	0	1
Cohort 31	0.0001	0.0157	-0.0150	-0.0004	0.0001
Cohort 32	-0.0001	0.0079	-0.0077	0.0000	0.0001
Cohort 33	0.0002	0.0309	-0.0300	0.0004	0.0003
Cohort 34	-0.0005	0.0238	-0.0223	0.0002	-0.0005
Cohort 35	0.0001	0.0080	-0.0075	-0.0002	0.0002
Cohort 36	0.0002	0.0232	-0.0221	-0.0000	-0.0005
Cohort 37	0.0000	0.0232	-0.0221	-0.0009	0.0000
Cohort 38	0.0009	0.0380	-0.0378	0.0011	-0.0005
Cohort 39	-0.0014	0.0394	-0.0367	-0.0006	0.0000
Cohort 40	-0.0000	0.0081	-0.0077	-0.0000	0.0001
Cohort 41	0.0007	0.0233	-0.0227	0.0003	0.0002
Cohort 42	-0.0003	0.0237	-0.0223	0.0002	0.0000
Cohort 43	-0.0000	0.0080	-0.0075	-0.0000	0.0000
Cohort 44	0.0000	0.0156	-0.0149	-0.0000	-0.0002
Cohort 45	0.0000	0.0156	-0.0149	-0.0000	-0.0001
Cohort 46	0.0000	0.0078	-0.0076	-0.0000	0.0000
Cohort 47	-0.0001	0.0079	-0.0076	-0.0000	-0.0000
Cohort 48	0.0000	0.0079	-0.0076	-0.0001	0.0001
Cohort 49	0.0000	0.0079	-0.0076	0.0001	0.0000
Cohort 50	0.0000	0.0157	-0.0150	-0.0000	0.0001
Cohort 51	-0.0001	0.0079	-0.0076	0.0001	0.0000
Cohort 52	0.0001	0.0079	-0.0075	-0.0000	-0.0000
Cohort 53	0.0000	0.0078	-0.0076	-0.0000	0.0000
Cohort 54	-0.0001	0.0157	-0.0151	-0.0000	0.0000
Cohort 55	-0.0001	0.0158	-0.0151	-0.0000	0.0003
Cohort 56	0.0000	0.0158	-0.0151	0.0002	0.0000
Cohort 57	0.0000	0.0158	-0.0149	-0.0000	0.0000
Cohort 58	0.0000	0.0156	-0.0152	0.0003	0.0000
Cohort 59	-0.0003	0.0235	-0.0221	-0.0000	NaN
Sum	0	1	-1	0	0

Note: This table shows the properties of weights associated with twoway fixed effects regressions as stated under (18). The remaining relative periods not shown in the table have a weight equal to zero (or extremely close to zero). The weights shown here are related to fixed effects estimates for rating upgrades. The properties hold for rating downgrades as well.

Table 3: Summary Statistics: Unbalanced Panel

	N	Mean	Std.Dev
MSCI			
ESG ownership (%)	16313	0.111	0.110
ESG weight (%)	16313	0.237	0.429
ESG score [0-10]	16313	4.791	2.097

Note: This table presents the summary statistics of firm-level characteristics from the unbalanced sample. Observations for Panel A. MSCI (212 firms) ranges from December 2013-2020. ESG ownership is calculated as the fraction of a company's outstanding shares owned by ESG mutual funds. ESG weight is the fraction that a company's shares represent in the portfolio value of a synthetic ESG mutual fund that aggregates the holdings of all ESG mutual funds. All fund related data is obtained from CRSP mutual fund and monthly stock databases. ESG ratings are from respective providers.

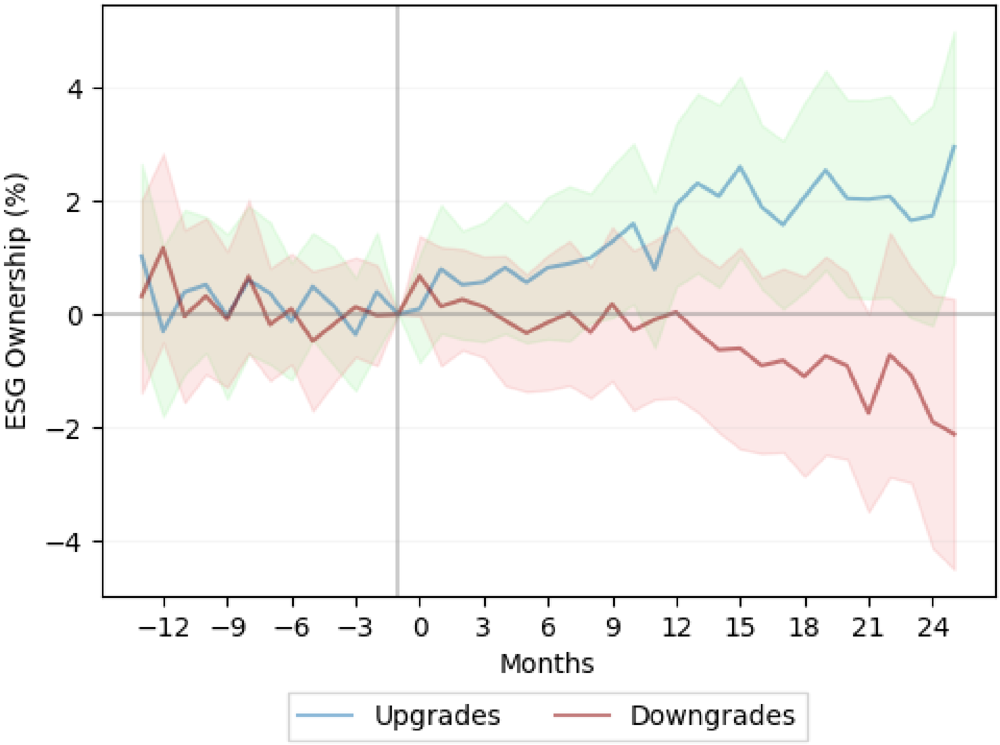
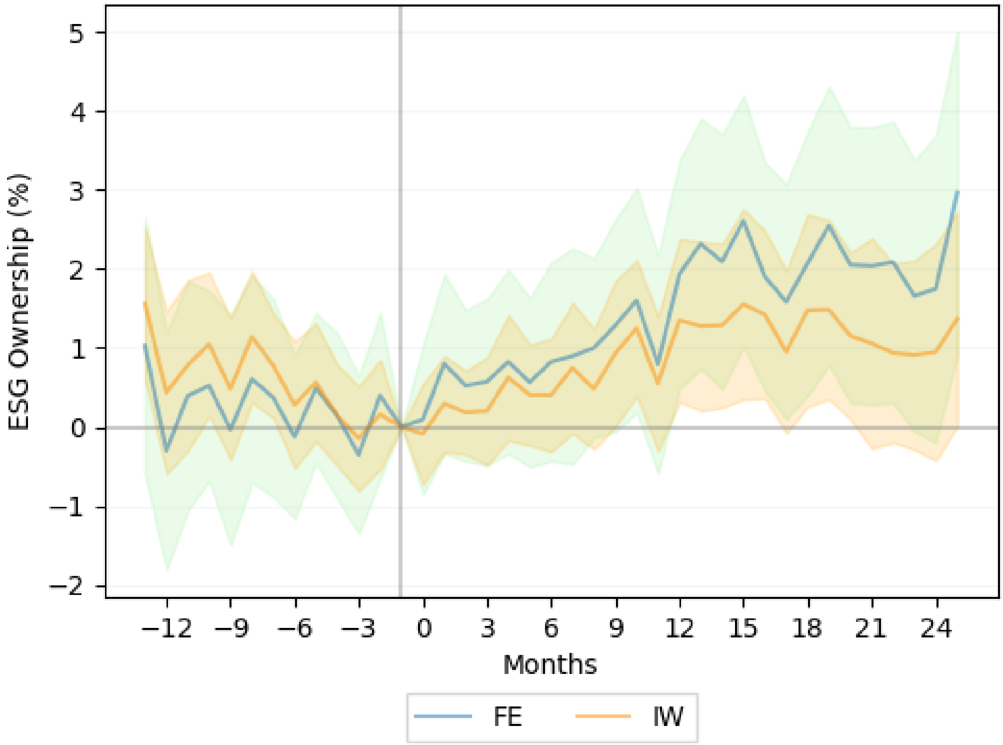
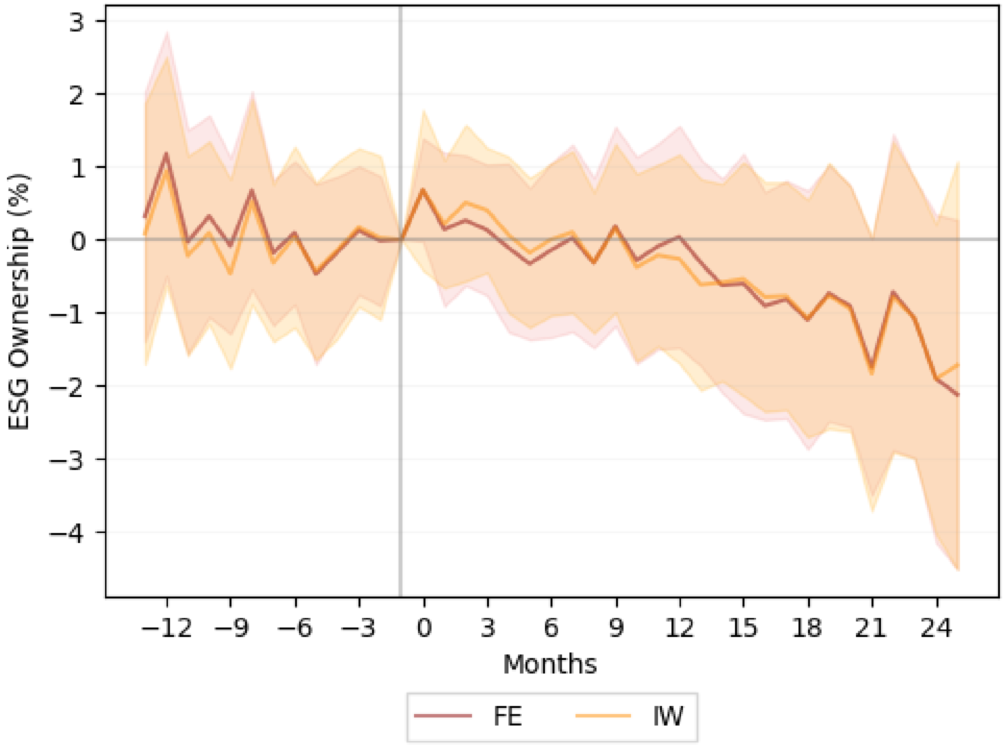


Figure 13: **The reaction of ESG ownership to ESG rating changes: Unbalanced panel** This figure shows the results from the event-study analysis using an unbalanced panel data sample.



(a) For MSCI rating upgrades



(b) For MSCI rating downgrades

Figure 14: **FE vs IW estimates: Unbalanced panel**
Each figure plots the FE estimates $\widehat{\beta}_l$ from regression (14) and IW estimates $\widehat{\nu}_l$ from regression (17) for each relative period (month) l and shaded region showing 95% confidence interval. Both specifications estimate effect of rating changes on ESG ownership at l.

Table 4: ESG ownership and rating change characteristics

	(1) ESG ownership	(2) ESG ownership
Upgrade (12 month-lag)	0.0270***	0.0178**
	(0.009)	(0.007)
Downgrade (12 month-lag)	-0.00639	-0.000882
	(0.009)	(0.009)
Upgrade (12 month-lag) x *Post 2016*	-0.0155*	
	(0.008)	
Downgrade (12 month-lag) x *Post 2016*	0.00821	
	(0.011)	
Upgrade (12 month-lag) x *High ESG score change*		-0.0334***
		(0.011)
Downgrade (12 month-lag) x *High ESG score change*		-0.0122
		(0.013)
Firm FE	Yes	Yes
Month FE	Yes	Yes
Pre-event leads	Yes	Yes
Post-event lags	Yes	Yes
Adjusted R-squared	0.652	0.652
N	11475	11475

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table shows regression results which check if the nature and timing of rating change has an effect on ownership. (1) represents interactions of 12 month-lag with dummy variable *Post 2016*, which indicates rating changes that occur after 2016. (2) includes interactions of 12 month-lag with dummy variable *High ESG score*, which indicates if the change in underlying numerical score was greater than or equal to the median of all score changes.