Finite-sample inference for small-N empirical designs: four diagnostics from the NGEU sovereign-spread debate*

Diogo Sampaio Lima
Maastricht University

May 2026

Abstract

This paper studies finite-sample inference in small-N applied empirical designs, using the debate on the effects of Next Generation EU (NGEU), the European Union's pandemic recovery programme, on euro-area sovereign spreads as a laboratory. The setting is empirically attractive but fragile: identification relies on a small number of countries, thin event-study cross-sections, continuous treatment intensity, and formula-based exposure measures. The paper shows that standard event-study inference rejects on 36–43% of clean pre-NGEU placebo trading days at a nominal 5% level, implying that conventional p-values cannot reliably distinguish NGEU news-day coefficients from ordinary pre-treatment cross-sectional spread movements in this design. The paper develops a practical diagnostic framework that links these failure modes to placebo-date calibration, few-cluster inference corrections, continuous-dose pre-trend diagnostics, and reduced-form and balance checks for formula-based exposure. Applied to NGEU, the framework substantially narrows the reduced-form evidence. Among four major NGEU news dates, only the 18 May 2020 Franco-German proposal lies outside the calibrated placebo distribution at short symmetric windows. Four-cluster auction specifications hit the wild-bootstrap p-value floor; monthly continuous-dose designs load on pre-2020 peripheral convergence; and formula-based exposure inherits the same pre-trend despite a strong first stage. The contribution is a portable decision framework and reporting standard for small-N empirical designs.

JEL codes: C12, C23, E43, E44, F33, G14, G15
Keywords: finite-sample inference; cluster-robust standard errors; wild cluster bootstrap; permutation inference; parallel trends; shift-share instruments; HC1 miscalibration; sovereign spreads; NextGenerationEU

*Email address: d.sampaiolima@maastrichtuniversity.nl.

1 Introduction

Next Generation EU (NGEU), the European Union's pandemic recovery programme, was one of the most important fiscal innovations in the euro area's response to COVID-19. By combining large country-level allocations with common EU borrowing, NGEU provided a natural setting in which to ask whether fiscal solidarity and safe-asset creation compressed euro-area sovereign spreads. A growing empirical literature studies this question using high-frequency event studies, panel regressions based on cross-country exposure to NGEU, and model-based decompositions of sovereign risk premia. This paper asks a narrower but prior question: are the reduced-form designs used, or naturally suggested, in this debate calibrated to the small effective sample sizes on which they rely?

The concern is not merely that the samples are small in a generic sense. The NGEU sovereign-spread setting combines several finite-sample problems that often arise together in macro-finance. Daily event studies use cross-sections of roughly a dozen sovereigns. Auction-panel specifications can collapse to four country clusters. Monthly continuous-dose designs assign larger exposure to peripheral countries whose spreads were already converging before NGEU. Formula-based exposure measures built from pre-pandemic allocation rules are predetermined, but their inputs also capture structural country characteristics that predict pre-treatment spread dynamics. Each issue is familiar in isolation: small-cluster inference (Cameron and Miller, 2015; MacKinnon and Webb, 2017; Canay et al., 2021), placebo calibration in small cross-sections (Hagemann, 2019), event-study pre-trend failures (Roth, 2022; Roth et al., 2023; Rambachan and Roth, 2023), and shift-share identifying assumptions (Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022, 2025). The contribution of this paper is to organize these concerns into a single diagnostic framework for applied small-N empirical designs.

The motivating result comes from the high-frequency event-study design. When the standard cross-sectional event-study regression is run on clean pre-NGEU placebo trading days, conventional HC1 inference rejects on 36–43% of placebo days at a nominal 5% level. This does not mean that the HC1 test has size 36–43% in a textbook structural-null experiment. It means that, in this empirical design, conventional event-study inference cannot reliably distinguish NGEU news-day coefficients from ordinary pre-treatment cross-sectional spread movements. The placebo exercise preserves the realized exposure vector and the dependence structure of sovereign spreads, and therefore measures the practical calibration problem faced by the event-study design actually used in this setting.

The paper develops four diagnostics matched to four failure modes. Placebo-date calibration addresses thin event-study cross-sections and common-factor dependence. Exact-enumeration wild cluster bootstrap and CR2/CR3 corrections address inference with very few country clusters. Binned event studies and Rambachan–Roth bounds address continuous-dose pre-trends. Reduced-form and balance checks, including an exact 10! permutation test on the ten-unit balance regression, address formula-based exposure measures whose inputs may predict pre-treatment outcome dynamics. The point is not to propose a new estimator for any one problem. It is to provide a practical map from empirical design to diagnostic: what is the

effective sample size, what identifying variation is being used, and which finite-sample failure mode is most likely to bind?

The paper makes three contributions. First, it integrates diagnostics that the econometrics literature has developed separately into a unified decision framework for small-N empirical applications. Second, it calibrates the framework in a concrete setting where the problems arise jointly: the NGEU sovereign-spread debate. The calibration is numerical rather than rhetorical. It reports clean 2018–2019 placebo distributions, exact-enumeration wild-bootstrap floors, binned pre-treatment coefficients, and formula-exposure balance tests with both asymptotic and exact-permutation p-values. Third, it translates the diagnostics into a reporting standard for researchers working with small cross-section event studies, small panels, continuous treatment intensity, or formula-based exposure.

The placebo-date calibration used here is related to, but distinct from, the placebo-permutation procedure of Hagemann (2019). Hagemann's procedure permutes treatment assignment across clusters and delivers formal size guarantees under exchangeability assumptions on the residuals across clusters. The calibration in this paper instead permutes the event date across a pre-treatment window while holding the cross-sectional structure fixed. The two approaches answer different questions. Cross-sectional permutation asks whether the realized assignment of exposure is unusual conditional on a given date. Date-placebo calibration asks whether the realized event-date coefficient is unusual relative to ordinary pre-treatment trading days generated by the same cross-section and exposure vector. In the NGEU event-study setting, where common factors move sovereign spreads strongly across days, the latter question is central for evaluating event-date extremeness.

The key message is that the relevant sample size is determined by the design, not by the number of rows in the dataset. A four-country auction panel does not become a large-cluster design because it contains hundreds of auctions. A twelve-sovereign event study does not inherit the calibration of large-N HC inference. A strong first stage does not make a formula-based exposure design credible if the formula's inputs also predict pre-treatment spread dynamics. The practical question is therefore not whether one can compute a conventional p-value, but whether that p-value is calibrated to the empirical design actually used.

Applied to NGEU sovereign spreads, four findings emerge. First, the high-frequency evidence is concentrated in one announcement. Among four major NGEU news dates examined against the pre-NGEU placebo distribution, only the 18 May 2020 Franco-German proposal lies outside the calibrated 95% interval, and only in short event windows. Second, the auction-panel evidence is limited by the effective number of country clusters. In the four-country specifications, exact wild-bootstrap inference has a minimum attainable two-sided p-value of 0.125, so asymptotically significant cluster-robust estimates cannot support conventional five-percent rejection. Third, the monthly continuous-dose design has a large negative post-2020 coefficient, but the same specification also shows large pre-2020 coefficients of the opposite sign. This indicates that the design loads on peripheral spread convergence already underway before NGEU. Fourth, the formula-based exposure design inherits the same pre-treatment pattern despite a strong

first stage. Its balance regression is suggestive under asymptotic inference but not under exact enumeration, so the diagnostic case against the design rests primarily on the reduced-form pre-trend evidence. Taken together, the diagnostics do not show that NGEU had no effect on sovereign spreads. They show that several reduced-form designs in this setting cannot credibly separate NGEU effects from finite-sample miscalibration and pre-existing cross-country dynamics.

The substantive claim is deliberately narrower than "NGEU did not matter." NGEU may well have affected euro-area sovereign risk premia through transfer, solidarity, and safe-asset channels. The question here is whether the reduced-form empirical tools used in this literature can establish those effects credibly at the sample sizes available. The placebo-date calibration speaks directly to the daily event-study design used in published sovereign-spread work. The monthly continuous-dose and formula-based exposure exercises evaluate natural extensions of the same empirical setting rather than literal replications of every published specification. The paper also does not solve the deeper macro-identification problems surrounding NGEU, including coincident policies, endogenous treatment intensity, and spillovers across sovereigns. Its claim is more limited but more general: even a researcher with a convincing answer to those broader identification problems still needs inference and design diagnostics calibrated to the small-sample environment.

The framework is portable beyond NGEU and sovereign debt. Its target is not a particular asset class, policy episode, or institutional setting, but a class of empirical designs in which the effective sample size is small relative to the inferential or identification claim being made. The same problems arise in event studies with few treated units, panels with few clusters, continuous-treatment designs where exposure may align with pre-existing trends, and formula-based allocation or exposure designs with limited cross-sectional support. Examples include studies of fiscal transfers, regional and industrial policy, climate and energy programmes, pandemic-era interventions, small-country panels, and shift-share designs built from a small number of exposure units. In such settings, conventional asymptotic inference is not self-validating. The diagnostic question should come first: what is the effective sample size, what identifying variation is being used, and which finite-sample failure mode is most likely to bind?

The remainder of the paper is organised as follows. Section 2 reviews the NGEU sovereign-spread literature and the econometric problems attached to its main empirical designs. Section 3 describes the data. Section 4 delimits what the diagnostics address and what they do not. Sections 5–8 present the event-study, small-cluster, monthly pre-trend, and formula-exposure diagnostics. Section 9 discusses the implications for published NGEU sovereign-spread findings. Section 10 packages the diagnostics as a decision framework. Section 11 discusses portability beyond NGEU and sovereign debt. Section 12 concludes.

2 Inference and identification problems in the NGEU sovereign-spread literature

2.1 Econometric problems in small-N designs

This paper connects four econometric concerns that are usually treated separately but often arise together in applied macro-finance. The first is the finite-sample calibration of heteroskedasticity-robust inference in small cross-sections: White-type standard errors are asymptotic objects (White, 1980; MacKinnon and White, 1985), and placebo or randomization procedures provide a way to assess test behavior in small samples (Hagemann, 2019; MacKinnon and Webb, 2020; Young, 2019). The second is inference with few clusters, where conventional cluster-robust approximations can be unreliable and wild-cluster procedures become sensitive to the effective number of clusters (Cameron and Miller, 2015; MacKinnon and Webb, 2017; Canay et al., 2021). The third is the credibility of parallel-trends assumptions in event-study and difference-in-differences designs, especially when treatment intensity is continuous and correlated with pre-existing outcome dynamics (Roth, 2022; Roth et al., 2023; Rambachan and Roth, 2023). The fourth is the identifying content of shift-share and exposure-based instruments, where predetermined exposure shares need not be exogenous to potential outcome trends (Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022, 2025).

The contribution of this paper is not a new estimator for any one of these problems. It is a unified diagnostic framework for settings in which several of them bind jointly. The NGEU sovereign-spread debate provides a useful laboratory because the same application contains very few country clusters, thin event-study cross-sections, continuous treatment intensity, and formula-based exposure measures built from structural country characteristics. Some of these problems arise directly in published reduced-form designs, while others arise in natural extensions that a researcher would plausibly run on the same data. The common feature is that the relevant empirical designs operate in a small-sample regime in which conventional inference and untested identifying assumptions are not self-validating. Table 11 in the appendix summarizes the mapping from empirical problem to diagnostic used in the paper.

2.2 NGEU sovereign-spread evidence

The empirical NGEU sovereign-spread literature is best organized by identification strategy. Three broad approaches account for most of the evidence: high-frequency event studies around major NGEU news dates, panel regressions using cross-country exposure to NGEU, and model-based decompositions of sovereign yields into underlying risk-premium components. Alongside these strands sits an auction-level literature on sovereign bid-to-cover ratios. That literature is not itself about NGEU, but it provides the primary-market template used in the small-cluster diagnostic below. Institutional and safe-asset studies provide background on the NGEU borrowing programme and its possible channels, but do not directly identify sovereign-spread effects.

Daily event-study papers provide the closest reduced-form sovereign-spread evidence. Havlik

et al. (2022) estimate event-study regressions for ten euro-area government-bond spreads and report that PEPP announcements had the largest spread-compressing effects while NGEU announcements had smaller but statistically significant effects. Related work studies reactions of yields, credit spreads, CDS spreads, bank stocks, and model-implied sovereign premia around PEPP, NGEU, and other pandemic policy news (Pancotto et al., 2023; Corradin and Schwaab, 2023). The common econometric issue is that these designs rely on a small number of countries observed around a small number of dates, so HC-type inference requires empirical calibration rather than large-sample presumption.

Panel designs use cross-country variation in RRF allocation intensity. A continuous-dose sovereign-spread design of exactly the form evaluated below is not the dominant published specification, but Alessi et al. (2025) provide a close neighbour by studying sovereign yields with a staggered design tied to Recovery and Resilience Plan approval and a continuous treatment based on the greenness share of each plan. The treatment differs from the RRF allocation-intensity measure used here, yet the identification concern is the same: if measured exposure is correlated with pre-existing spread dynamics, a panel coefficient can load on pre-treatment convergence rather than on the policy shock. The diagnostics below evaluate that concern for a natural continuous-dose extension of the NGEU spread setting; they do not re-estimate every published specification.

Model-based, institutional, safe-asset, and auction-level literatures provide the surrounding economic and empirical context. ECB and NIESR studies quantify possible macro-financial transmission of NGEU (Bańkowski et al., 2021, 2022, 2024; Millard, 2025), while work on EU safe assets and borrowing architecture explains why an NGEU effect on sovereign premia is economically plausible (Bletzinger et al., 2022; Christie et al., 2021; Temprano Arroyo, 2022). The auction diagnostic draws on the bid-to-cover literature, especially Beetsma et al. (2018) and Beetsma et al. (2020), because that literature supplies a primary-market outcome and a panel structure in which the effective number of country clusters can be very small. The monthly continuous-dose diagnostic is related to the broader TWFE and continuous-treatment DiD critique (de Chaisemartin and D'Haultfoeuille, 2020; Callaway et al., 2024), but its failure is logically prior to weighting concerns: the residual pre-trend restriction already fails in the data. These strands motivate the application, but the diagnostics target reduced-form inference and identification rather than structural transmission itself.

3 Data

The empirical analysis draws on three panels, each matched to a diagnostic: a daily bond benchmark-yield panel for the event-study calibration (Diagnostic 1, section 6), an bond auction-level panel for the small-cluster exercise (Diagnostic 2, section 7), and a Eurostat monthly sovereign-yield panel for the continuous-dose and formula exercises (Diagnostics 3, section 8 and 4, section 5). This section is the single reference for panel composition, spread construction, and the RRF allocation-intensity measures; the diagnostic sections refer back to it rather than restating panel details. Appendix A collects the dataset provenance, vendor identifiers and

ticker construction, comparison-unit caveats, and the sample-size arithmetic that is too detailed for the main text.

3.1 Empirical panels

All three panels measure sovereign yields relative to the German Bund. , the Bund spread is

$$\text{spread}_{c,t} = 100\,(y_{c,t} - y_{\text{DE},t})$$

in basis points, where $y_{c,t}$ is the maturity-matched benchmark yield. Germany is the reference throughout and never enters as a spread series, so a panel of K sovereigns that includes Germany contributes $K - 1$ non-German Bund-spread series.

The daily event-study panel uses ten-year benchmark sovereign yields from FactSet Workstation, downloaded in April 2026, covering January 2018 through December 2024. It contains thirteen sovereigns: eight euro-area countries (Austria, Belgium, Finland, France, Germany, Italy, Portugal, and Spain) and five non-euro advanced economies (Denmark, Norway, Sweden, Switzerland, and the United Kingdom). Netting out Germany leaves twelve non-German daily Bund spreads. The five non-euro units are insulated from direct Eurosystem sovereign-bond purchases and widen the cross-section, but they are not treated as clean counterfactuals for euro-area sovereigns; Diagnostic 1 in section 6 uses them only to discipline the placebo benchmark.

The auction panel extends the sovereign-auction design of Beetsma et al. (2020) through March 2023. Its main outcome is the bid-to-cover ratio, defined as total bids divided by competitive allotment. The full-control specifications use the Belgium–France–Italy–Portugal sample, the largest set of countries for which the Beetsma-style controls, lagged auction variables, volatility controls, and central-bank-purchase variables are jointly observed. The effective number of country clusters is therefore four, the binding feature examined in Diagnostic 2, section 7

The monthly country panel uses Eurostat's harmonised Maastricht long-term interest-rate series, `IRT_LT_MCBY_M`, covering January 2018 through December 2024. It contains eleven sovereigns: the same eight euro-area countries together with three non-euro comparison units (Denmark, Sweden, and the United Kingdom). Norway and Switzerland are not carried in this panel because they are not covered by the Eurostat harmonised series. Netting out Germany leaves ten non-German monthly Bund spreads. This panel supports the continuous-dose and formula-exposure diagnostics.

3.2 Treatment variables

The paper uses three cross-sectional RRF allocation-intensity measures. Throughout, RRF allocation intensity refers to grant-plus-loan or grant-only Recovery and Resilience Facility exposure scaled by 2019 GDP; formula-based RRF intensity refers to the reconstructed pre-pandemic grant-allocation formula measure. NGEU denotes the broader policy package, while NRRP is reserved for national recovery and resilience plans. Each measure is defined at the country level. In the panel specifications, the exposure measure is interacted with a post-

21 July 2020 indicator, where 21 July 2020 is the European Council agreement that finalized the core structure of Next Generation EU. Formally,

$$
\begin{aligned}
D_c^{\mathrm{BL}} &= \frac{\text{RRF grants}_c + \text{RRF loans}_c}{\text{GDP}_{c,2019}}, \\
D_c^{\mathrm{G}} &= \frac{\text{RRF grants}_c}{\text{GDP}_{c,2019}}, \\
Z_c^{\mathrm{F}} &= \text{formula-based pre-pandemic grant intensity.}
\end{aligned}
$$

The baseline intensity D_c^{BL} is the default exposure measure. The grants-only intensity D_c^{G} is used in robustness checks. The formula-based measure Z_c^{F} is used in the formula-exposure diagnostic.

The baseline measure is the sum of the country-specific Recovery and Resilience Facility grant and loan ceilings, divided by 2019 nominal GDP. Using ceilings rather than realized disbursements ties the exposure measure to the allocation design rather than to later implementation.

The second measure uses grants only, again scaled by 2019 nominal GDP. For countries that did not request RRF loans (Austria, Belgium, Finland, France, and Germany), this measure coincides with the baseline measure. The largest difference is for Italy, where the grants-only measure is much smaller than the combined grants-and-loans measure. The distinction matters because grants and loans have different fiscal implications: grants create no repayment obligation for the recipient government, while RRF loans do.

The third measure reconstructs the pre-pandemic component of the RRF grant-allocation formula from 2019 GDP per capita, average unemployment over 2015–2019, and 2019 population. The raw formula weights are normalised to the population of euro-area recipients and multiplied by the pre-pandemic share of the grants envelope. This measure is predetermined relative to the NGEU announcement, but its inputs are structural country characteristics that may also predict pre-treatment sovereign-spread dynamics. Diagnostic 4 tests exactly that concern.

Because each exposure measure is cross-sectional and time-invariant, identification in the panel specifications comes from interacting exposure with the post-21 July 2020 indicator, not from within-country variation in exposure. Country fixed effects then absorb persistent differences across sovereigns—average spread and bid-to-cover levels, auction institutions, investor bases, and the currency regimes of the non-euro units—while month or year fixed effects absorb common shocks such as the COVID liquidity episode, the euro-area policy responses, and the 2022 monetary-tightening cycle. These fixed effects remove level differences and common time variation. They do not remove treatment-correlated pre-trends, which is the identification problem examined in Section 7.

Table 12 in appendix B reports the three measures for the eight euro-area countries in the exposure contrast. Germany is included for completeness even though it serves only as the spread reference; its allocation intensity is among the lowest in the table. In the monthly and daily diagnostics the non-euro sovereigns are assigned zero exposure, a diagnostic convention for isolating the euro-area sovereign-spread channel rather than a statement of institutional

eligibility. The United Kingdom, Norway, and Switzerland are outside the European Union and mechanically ineligible for the Recovery and Resilience Facility; Denmark and Sweden are EU member states that received RRF allocations but lie outside the euro area, so their zero assignment reflects the channel under study, not ineligibility.

4 Scope: what the diagnostics address and what they do not

The diagnostics address two empirical layers. The first is inference calibration: whether reported standard errors and p-values reflect the sampling distribution of the relevant test statistic at the sample sizes actually used. The second is design-specific identification: whether the specification isolates the variation it claims to isolate. Diagnostics 6 and 7 primarily address inference. Diagnostics 8 and 5 primarily address identification. Both layers matter because well-calibrated inference cannot rescue a misidentified design, and a plausible identifying story still needs inference that is credible at the effective sample size.

First, NGEU was introduced during a dense euro-area crisis-policy episode that also included PEPP, SURE, the ESM pandemic credit line, and national fiscal packages. The diagnostics do not decompose NGEU from those coincident policies. A design that tries to separate those mechanisms would need additional variation, outcomes, or structure. The contribution here is narrower: even after setting the joint-policy problem aside, the reduced-form designs still require finite-sample and pre-treatment diagnostics.

Second, RRF exposure is not randomly assigned. The allocation formula directed more support toward countries with lower GDP per capita, higher unemployment, and larger populations, and those characteristics also predict sovereign-risk dynamics. Diagnostics 3 and 4 show the empirical trace of this problem: larger exposure aligns with pre-2020 spread convergence, and formula-based exposure inherits that pattern despite a strong first stage. The diagnostics reveal the selection problem; they do not eliminate it.

Third, Euro-area sovereign markets are tightly connected through portfolio rebalancing, common risk premia, redenomination-risk beliefs, and ECB policy expectations. Cross-country exposure contrasts may therefore combine direct effects on high-exposure countries with indirect equilibrium effects on other sovereigns. The paper does not solve this spillover problem. It shows that, even before imposing a no-spillover interpretation, the reduced-form designs face binding inference and pre-treatment identification problems.

The paper makes a narrow diagnostic claim. It does not deliver a structural decomposition of NGEU, PEPP, SURE, national fiscal policy, and sovereign-market spillovers. It shows that several reduced-form designs used, or naturally suggested, in the NGEU spread setting are fragile once their effective sample sizes and pre-treatment identifying variation are examined directly. The question is therefore not whether NGEU mattered in an economic sense. The question is what these small-N empirical designs can credibly establish.

5 Diagnostic 1: Event-study placebo-date calibration in a small sovereign-spread cross-section

This section turns first to the high-frequency event-study evidence on NGEU sovereign spreads. Event studies are attractive because narrow windows around policy announcements can isolate market reactions while avoiding the lower-frequency pre-trend problems documented later in Section 7. In this application, however, the main problem is not parallel trends. It is inference. When the cross-section contains only a small number of sovereigns and spread changes are strongly correlated across countries, conventional heteroskedasticity-robust inference can be badly miscalibrated.

5.1 The event-study regression benchmarked here

The high-frequency NGEU literature studies sovereign-yield, sovereign-spread, CDS, and bank-market reactions around major pandemic policy announcements (Havlik et al., 2022; Pancotto et al., 2023; Corradin and Schwaab, 2023). These papers differ in implementation and outcome, but the common econometric feature is that the identifying information comes from a small number of sovereigns observed around a small number of dates. This diagnostic isolates the small-cross-section sovereign-spread component of that evidence and asks whether a White-type heteroskedasticity-robust test is calibrated when roughly a dozen sovereigns are observed around one news date.

For each news day t^*, let $s_{c,t} = 100(y_{c,t} - y_{\mathrm{DE},t})$ denote country c's Bund spread in basis points. The main event-window object is a symmetric trading-day window, denoted $S \pm h$, with half-width h:

$$\Delta s_c^{S\pm h}(t^*) = s_{c,t^*+h} - s_{c,t^*-h},$$

where $h \in \{1, 2, 5\}$ in the benchmark calibration. Thus $S \pm 1$ is the spread change from $t^* - 1$ to $t^* + 1$; it is not a close-to-close one-day return. Appendix H.5 reports close-to-close and post-announcement timing checks separately. The cross-section event regression is

$$\Delta s_c^{S\pm h}(t^*) = \alpha + \beta\, \mathrm{intensity}_c + \varepsilon_c,$$

estimated with HC1 standard errors, the conventional finite-sample version of White-type heteroskedasticity-robust inference used in many small cross-section applications. The coefficient β measures the basis-point spread response per unit of RRF allocation intensity.

Table 1: Event-window notation used in Diagnostic 1

Label	Definition	Role in paper
$S \pm 1$	$s_{c,t^*+1} - s_{c,t^*-1}$	Main short symmetric benchmark
$S \pm 2$	$s_{c,t^*+2} - s_{c,t^*-2}$	Wider symmetric benchmark
$S \pm 5$	$s_{c,t^*+5} - s_{c,t^*-5}$	Long symmetric benchmark
CC	$s_{c,t^*} - s_{c,t^*-1}$	Close-to-close timing robustness
P1	$s_{c,t^*+1} - s_{c,t^*}$	Post-announcement timing robustness

The relevant sample size is the number of non-German sovereigns in the event-date cross-section. In the panel of Section 3.1, that number is twelve. At such N, heteroskedasticity-robust standard errors are not protected by large-cross-section asymptotics, and common-factor dependence across sovereign spreads is a first-order concern. The question is therefore empirical: how often would the same HC1 event-study regression reject on placebo dates when no NGEU effect is possible?

5.2 Cross-sectional dependence and common factors

Before benchmarking inference, it is useful to document the dependence structure of daily spread changes. Table 2 summarizes pairwise correlations of daily Bund-spread changes $\Delta s_{c,t}$ across the twelve non-German sovereigns. Correlations are estimated over the 474 trading days in 2018–2019 on which all twelve spread changes are observed, and then averaged within and between three groups: peripheral euro (IT, ES, PT), core euro (AT, BE, FI, FR), and non-euro (DK, SE, UK, NO, CH). Appendix H.1 reports the full correlation heatmap and a principal-components analysis.

Table 2: Group-level correlations of daily Bund-spread changes

Group pair	# pairs	Mean correlation	Median correlation
Peripheral–peripheral	3	0.64	0.65
Core–core	6	0.89	0.90
Non-euro–non-euro	10	0.49	0.50
Peripheral–core	12	0.53	0.51
Peripheral–non-euro	15	0.41	0.42
Core–non-euro	20	0.62	0.67

Notes: Correlations are computed using daily Bund-spread changes from 2018-01 to 2019-12 on the 474 trading days with complete observations for all twelve non-German sovereigns. Within-group entries use off-diagonal upper-triangle pairs only. The displayed correlations are raw Δs correlations, not residual correlations after partialling out intensity. Since intensity is country-fixed, partialling-out is mostly demeaning and the qualitative pattern is unchanged.

The dependence is strong. Within-group correlations average 0.64 for the peripheral sovereigns, 0.89 for the core euro sovereigns, and 0.49 for the non-euro sovereigns. Between-group correlations are also large: peripheral–core correlations average 0.53, and core–non-euro correlations average 0.62. The principal-components analysis in Appendix H.1 reaches the same conclusion: the first principal component explains 62% of the cross-sectional variance and loads positively on all twelve sovereigns.

This dependence matters for inference. HC-type standard errors are designed to handle heteroskedasticity, but they do not account for arbitrary cross-sectional covariance among the twelve sovereign residuals. In a large independent cross-section, that limitation may be less consequential. In the present small cross-section, where daily spread changes share strong common factors, the missing covariance terms can materially understate the variance of the slope on exposure. The placebo benchmark below measures the practical consequence of that problem directly.

5.3 The clean placebo benchmark

The daily event-study panel is the one described in Section 3.1: twelve non-German ten-year Bund spreads over 2018–2024, with the eight euro-area recipients carrying their baseline RRF allocation intensity from Table 12 and the five non-euro units assigned zero exposure as a diagnostic convention. The benchmark holds this cross-section and exposure vector fixed and varies only the event date, so it measures how often the same HC1 event-study regression rejects on pre-NGEU days when no NGEU effect is possible.

The benchmark uses placebo trading days from a strictly pre-NGEU, pre-PEPP, pre-COVID window. The main placebo set contains all eligible complete-cross-section trading days from 2018–2019, with a ten-trading-day buffer at each end so that the longest symmetric window, $S \pm 5$, always fits. This yields 488 placebo days.[12]

For each placebo date t^* and each window $S \pm h$, $h \in \{1, 2, 5\}$, the same cross-section regression as in Section 5.1 is estimated with HC1 robust standard errors, storing the estimated coefficient $\hat{\beta}$, the HC1 p-value, and the rejection indicator $\mathbf{1}\{p_{\text{HC1}} < 0.05\}$.

Because NGEU did not yet exist on these placebo dates, the structural NGEU effect is zero by construction. The placebo exercise deliberately keeps the realized exposure vector fixed. It does not require exposure to be uncorrelated with pre-existing risk factors; indeed, preserving that empirical correlation is the point. The benchmark asks how often the event-study regression would label ordinary pre-NGEU spread movements as significant when run with the same cross-section and the same exposure vector. The empirical 2.5 and 97.5 percentiles of the placebo coefficient distribution form the benchmark 95% interval used below.

5.4 Placebo rejection rates and what they measure

Table 3 reports the placebo rejection rates and coefficient distributions for the three event windows.

The result is stark. At a nominal five-percent level, the HC1 test rejects on 35.9% of placebo days at $S \pm 1$, 39.1% at $S \pm 2$, and 43.2% at $S \pm 5$; at a nominal one-percent level, the empirical placebo rejection rate is 23.8–30.7%. A moving-block bootstrap (block length 30 trading days, $B = 5{,}000$, seed 20260527; full intervals in the notes to Table 3) accounts for the day-to-day dependence induced by overlapping placebo windows, and the lower bound of every confidence interval still exceeds the nominal level by an order of magnitude. Even the most conservative reading of the $S \pm 1$ rate is several times nominal size. In this design, conventional HC1 inference is not mildly noisy; it is severely miscalibrated.

[1]The 2018–2019 window is the cleanest pre-treatment baseline available: it precedes the Pandemic Emergency Purchase Programme, the Franco-German proposal, and the broader COVID-19 macroeconomic shock. Appendix H.4 reports a broader 2018-01 to 2020-06 exercise as robustness. That broader window is useful for checking stability in a longer sample, but it is not the main benchmark because it includes the early COVID period and dates close to the NGEU policy sequence.

[2]The 474-day correlation denominator in Table 2 and the 488-day placebo denominator differ because they are built from different objects. The correlation table requires complete one-day spread changes observed simultaneously for all twelve non-German sovereigns, while the placebo benchmark requires complete event-window endpoint spreads after applying the symmetric-window buffer. The two completeness rules need not select the same days.

Table 3: Placebo benchmark of the HC1 cross-section event-study test

Window	Definition	N days	PRR @ 5%	PRR @ 1%	Mean $\hat{\beta}$	SD $\hat{\beta}$	95% interval
$S \pm 1$	$t^* - 1 \to t^* + 1$	488	35.9%	23.8%	-2.8	54.5	$[-93.3, 112.5]$
$S \pm 2$	$t^* - 2 \to t^* + 2$	488	39.1%	25.8%	-5.3	75.2	$[-127.2, 129.2]$
$S \pm 5$	$t^* - 5 \to t^* + 5$	488	43.2%	30.7%	-9.2	119.8	$[-185.1, 243.4]$

Notes: The placebo sample is the clean 2018–2019 pre-NGEU window, using all 488 complete-cross-section eligible trading days after applying the ten-trading-day buffer. $\hat{\beta}$ is the slope on RRF allocation intensity in the cross-section regression of spread changes on exposure. PRR is the empirical placebo rejection rate of the HC1 test under the pre-treatment placebo distribution. Under correct size, the 5% and 1% columns would be close to their nominal levels. The 95% interval is the empirical [2.5%, 97.5%] quantile range of the placebo $\hat{\beta}$ distribution. Moving-block bootstrap 95% CIs on the PRR @ 5% column (block length 30 trading days, $B = 5{,}000$, seed 20260527) are [29.1%, 41.6%] at $S \pm 1$, [32.4%, 45.3%] at $S \pm 2$, and [34.2%, 51.2%] at $S \pm 5$; the corresponding 1% CIs are [17.4%, 29.5%], [19.3%, 30.9%], and [22.5%, 37.3%]. The lower bound of every CI exceeds the nominal level by an order of magnitude.

The coefficient distribution is also wide. The $S \pm 1$ 95% interval runs from -93.3 to 112.5 basis points per unit exposure; the $S \pm 5$ interval runs from -185.1 to 243.4. A coefficient that looks statistically significant under HC1 may therefore be well within the range of ordinary pre-NGEU cross-section movements. This is the central inferential result of the section.

The likely mechanism is the dependence documented in Table 2. Daily sovereign spread changes contain common factors: a peripheral euro component, a core euro component, and a non-euro fixed-income component. The RRF allocation-intensity vector loads heavily on the peripheral group. On days when common factors move the peripheral group differently from the rest of the cross-section, the event regression can generate a large slope even though no NGEU news is present. HC1 does not account for the covariance induced by those common factors in a twelve-unit cross-section, so the test rejects more often than nominal size.

It is worth being precise about what this rate is and is not. The placebo rejection rate is not the size of the HC1 test in the conventional sense. Under the structural null $\beta = 0$ ("NGEU has no causal effect"), the population OLS slope of Δs on RRF intensity over a placebo day is not generically zero: it equals the cross-sectional covariance of intensity with whatever latent factor loads on the spread that day, divided by $\mathrm{Var}(\text{intensity})$. RRF intensity is correlated by construction with country loadings on the peripheral factor that drives most of the cross-sectional variation in Δs. The 36–43% rejection rate therefore reflects a combination of (i) HC1 miscalibration in thin cross-sections and (ii) the fact that on a typical day the population slope of Δs on intensity is materially different from zero in this design. What the rate cleanly diagnoses is the inability of conventional event-study inference to discriminate event-day coefficients from ordinary pre-NGEU coefficients in this design. That is the operationally relevant statement for evaluating the NGEU event-study literature, and the statement emphasized in the implications.

5.5 The four NGEU news dates under the benchmark null

Table 5 applies the clean placebo benchmark to four major dates in the NGEU policy sequence: the 18 May 2020 Franco-German recovery-fund proposal, the 21 July 2020 European Council agreement, the 14 December 2020 Council Regulation establishing the European Union Recovery

Instrument, and the 15 June 2021 first NextGenerationEU bond issuance. For each event-window pair, the table reports the cross-section coefficient, the conventional HC1 p-value, and whether the coefficient lies outside the benchmark 95% placebo interval for the corresponding symmetric horizon. Table 4 states the macro news content attached to each date. This is not a separate identification assumption; it clarifies why the four dates are grouped together as NGEU news while still allowing the placebo benchmark to decide whether the realized cross-section was unusual.

Table 4: News content of the four NGEU event dates

Date	News	Main channel	Diagnostic interpretation
18 May 2020	Franco-German proposal for a jointly financed recovery fund	Transfer and solidarity signal	Main surviving high-frequency spread response
21 July 2020	European Council agreement on NGEU package	Political agreement and allocation certainty	Exposure-aligned relative to permutation null, but not unusual relative to placebo dates
14 December 2020	Regulation establishing the European Union Recovery Instrument	Legal implementation	Conventional HC1 significance does not survive placebo benchmark
15 June 2021	First NextGenerationEU bond issuance	Funding execution and safe-asset supply	Small and sign-unstable spread response

Table 5: NGEU news dates under the benchmark placebo null

Event (t^*)	Window	Definition	$\hat{\beta}$ (bps)	HC1 p	Benchmark verdict
18 May 2020 Franco-German proposal	$S \pm 1$	$t^*-1 \to t^*+1$	-180.5	0.000	**outside** ($z = -3.26$)
	$S \pm 2$	$t^*-2 \to t^*+2$	-144.4	0.010	**outside** ($z = -1.85$)
	$S \pm 5$	$t^*-5 \to t^*+5$	-177.2	0.006	inside ($z = -1.40$)
21 July 2020 Council agreement	$S \pm 1$	$t^*-1 \to t^*+1$	-9.9	0.326	inside ($z = -0.13$)
	$S \pm 2$	$t^*-2 \to t^*+2$	-81.6	0.032	inside ($z = -1.02$)
	$S \pm 5$	$t^*-5 \to t^*+5$	-49.5	0.312	inside ($z = -0.34$)
14 December 2020 Recovery Instrument regulation	$S \pm 1$	$t^*-1 \to t^*+1$	-29.4	0.008	inside ($z = -0.49$)
	$S \pm 2$	$t^*-2 \to t^*+2$	-35.9	0.021	inside ($z = -0.41$)
	$S \pm 5$	$t^*-5 \to t^*+5$	8.5	0.574	inside ($z = 0.15$)
15 June 2021 First NGEU issuance	$S \pm 1$	$t^*-1 \to t^*+1$	-4.1	0.209	inside ($z = -0.02$)
	$S \pm 2$	$t^*-2 \to t^*+2$	15.9	0.035	inside ($z = 0.28$)
	$S \pm 5$	$t^*-5 \to t^*+5$	4.9	0.637	inside ($z = 0.12$)

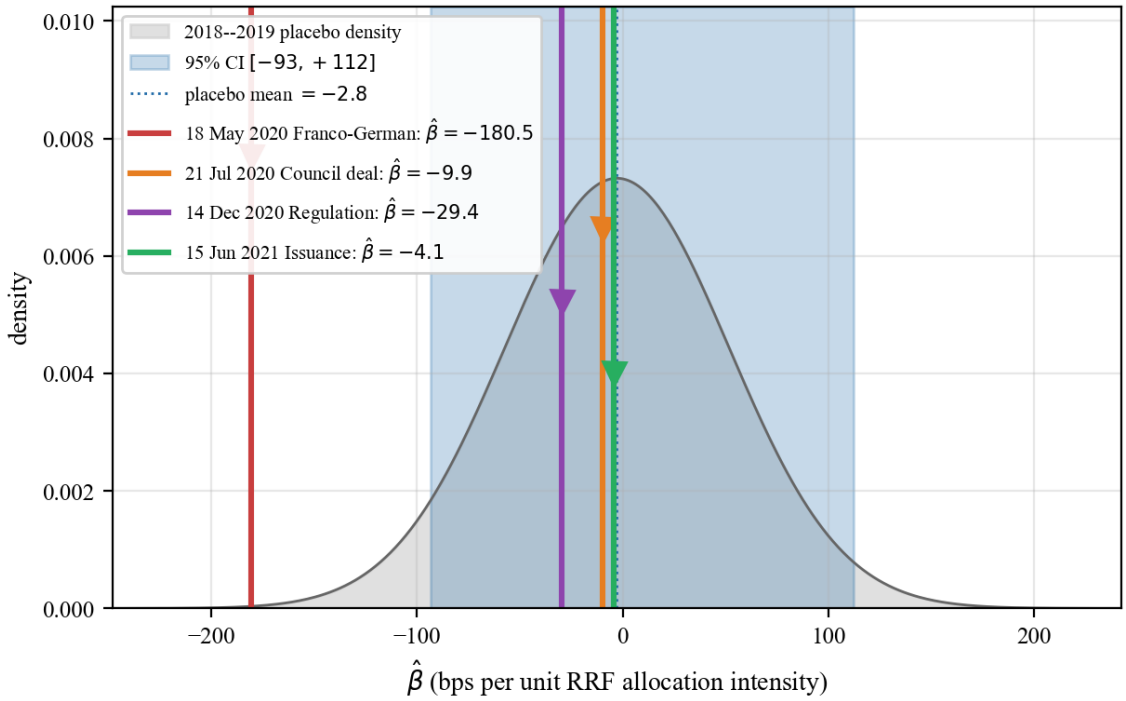
Notes: $\hat{\beta}$ is the cross-section estimate of the spread response per unit RRF allocation intensity, in basis points. "Outside" means the event coefficient lies below the 2.5% quantile or above the 97.5% quantile of the clean 488-day 2018–2019 placebo distribution for the same symmetric horizon. z is the location of the event coefficient relative to the placebo mean, in placebo-distribution standard deviations.

Only the 18 May 2020 Franco-German proposal lies outside the benchmark placebo interval at short symmetric windows. At $S \pm 1$, the coefficient is -180.5 basis points per unit exposure, more negative than the 2.5% placebo quantile of -93.3 basis points. The $S \pm 2$ coefficient also

lies outside the benchmark interval. At $S \pm 5$, however, the placebo interval widens and the coefficient falls inside it. The surviving event-study evidence is therefore concentrated in short symmetric windows around the Franco-German proposal.

The other three dates do not survive the placebo benchmark. The 21 July 2020 Council agreement is conventionally significant under HC1 at $S \pm 2$, but its coefficient lies comfortably inside the benchmark placebo interval. The 14 December 2020 Recovery Instrument regulation has HC1 p-values that would look significant in a conventional table at $S \pm 1$ and $S \pm 2$, but both coefficients sit close to the center of the placebo distribution. The 15 June 2021 first NGEU issuance is null by any standard: the estimates are small and switch sign across windows.

Figure 1: Symmetric $S \pm 1$ placebo distribution and NGEU news-day coefficients



Notes: The grey density is a normal approximation to the 488-day clean 2018–2019 placebo distribution; the blue shaded band is the empirical 95% interval from Table 3. The vertical lines mark the four NGEU news-day coefficients. Only the 18 May 2020 Franco-German proposal lies outside the benchmark 95% interval at $S \pm 1$.

Figure 1 visualizes the $S \pm 1$ result. The Franco-German proposal sits in the far left tail of the placebo distribution. The other three dates lie inside the range of ordinary pre-NGEU variation. The benchmark does not prove that NGEU had no effect on those dates. It shows that this small cross-section event-study design cannot distinguish their observed spread patterns from placebo realizations with sufficient confidence.

5.6 Robustness and alternative inference

The over-rejection is not specific to HC1 or to the placebo window. Appendix H.2 repeats the placebo benchmark using homoskedastic OLS standard errors and the HC0–HC3 family; HC3, the most conservative member, still rejects on roughly 22–27% of placebo days at the nominal 5% level in the broader robustness window. Appendix H.4 confirms that the over-rejection is stable

over a broader 2018–mid-2020 window, which it treats as robustness while the main calibration remains the clean 488-day 2018–2019 benchmark. Practitioners may prefer HC3 in cross-sections of this size, but even HC3 should not be treated as decisive without benchmarking.

Appendix H reports the remaining companion checks: exact permutation inference, event-window robustness across alternative window conventions, factor-residualized placebo benchmarks, and sample-composition robustness. Two conclusions matter for the main text. First, the 18 May 2020 result is also the most stable event-date result. Under alternative window conventions the close-to-close estimate is smaller, but windows that include the announcement day and the following day generate coefficients between roughly -130 and -180 basis points with very small HC1 p-values, and a leave-one-country-out check leaves the $S \pm 1$ Franco-German coefficient negative and conventionally significant across all twelve country drops. Second, factor residualization clarifies the mechanism: removing only the first principal-component loading leaves the Franco-German proposal outside the $S \pm 1$ and $S \pm 2$ placebo intervals, while removing the first two loading vectors absorbs most of the cross-sectional factor structure and makes the event no longer exceptional. This exercise should not be read as controlling for a structural NGEU factor; it is a diagnostic decomposition of how much placebo miscalibration is attributable to common-factor structure.

The exact permutation test is useful but answers a different null from the placebo benchmark. The permutation test holds the event-day spread-change vector fixed and permutes the exposure labels across sovereigns. The placebo benchmark instead asks whether the observed event-day coefficient is unusual relative to ordinary pre-treatment trading days generated by the same cross-section and exposure vector. In the presence of strong common-factor day-to-day variation, the placebo null is the relevant benchmark for event-date extremeness.

5.7 Interpretation

The conclusion is straightforward. In small sovereign-event cross-sections, conventional HC1 inference can be badly miscalibrated. In this application, a nominal five-percent HC1 test rejects on 36–43% of clean pre-NGEU placebo days. That is too much over-rejection to treat conventional event-study p-values as self-validating.

The benchmark event-study evidence is narrower than the conventional HC1 evidence. Among the four NGEU news dates examined here, only the 18 May 2020 Franco-German proposal produces a response that is clearly unusual relative to the pre-NGEU placebo distribution, and only at short symmetric windows. The other dates may still have mattered economically, but this design does not distinguish their observed cross-sectional patterns from placebo variation with sufficient confidence.

This is not a general indictment of HC1. In larger cross-sections with weak cross-sectional dependence, HC-type inference may be adequate. The point is specific to the data environment studied here: a twelve-sovereign cross-section with substantial common-factor dependence. Under those conditions, benchmarking against a placebo distribution is not optional. It is necessary for credible inference.

6 Diagnostic 2: Small-cluster inference and the wild cluster bootstrap p-value floor

The second diagnostic is the simplest and most mechanical. With very few clusters, exact-enumeration wild cluster bootstrap inference is constrained by the discreteness of the bootstrap distribution. In a panel with four clusters, a nonrandomized two-sided Rademacher wild cluster bootstrap cannot produce a positive p-value below $1/8 = 0.125$. Asymptotically significant results on such panels therefore do not, by themselves, support rejection at the conventional five-percent level under this exact-enumeration bootstrap reference distribution. The gap between the asymptotic and enumerated bootstrap p-values is not a feature of the NGEU application. It is a structural property of the Rademacher sign-vector support and is well known in the small-cluster inference literature (Webb, 2014; MacKinnon and Webb, 2017, 2020; Canay et al., 2021).

6.1 Why the cluster-count floor matters

The Rademacher implementation of the wild cluster bootstrap associated with Cameron et al. (2008) constructs a null distribution by re-signing restricted-null residuals at the cluster level and recomputing the test statistic. Under the null hypothesis $\beta = 0$, each cluster's residual contribution is multiplied by an independent Rademacher weight, $w_g \in \{-1, 1\}$. With G clusters, there are 2^G possible sign vectors. When G is small, these sign vectors can be enumerated exactly rather than sampled at random (MacKinnon and Webb, 2017; Webb, 2014).

Exact enumeration makes the discreteness of the test transparent. Because the Rademacher distribution is symmetric, every sign vector w has a mirror image $-w$. For a two-sided test based on $|t|$, those two sign vectors produce the same absolute statistic under the usual sign-symmetry condition. Consequently, the smallest non-zero two-sided p-value is

$$\frac{2}{2^G} = \frac{1}{2^{G-1}}.$$

For $G = 4$, this floor is $1/8 = 0.125$. For $G = 5$, it is $1/16 = 0.0625$. Only at $G = 6$ does the floor fall below the conventional five-percent threshold, since $1/32 \simeq 0.0312$.

The following lemma states the result formally. The result is not new. It is the basis for Webb (2014) introduction of the "6-point" auxiliary distribution and is developed further by MacKinnon and Webb (2017, 2020); see also Cameron et al. (2008) for the original discreteness observation. The result is restated here for self-containment because it is the structural basis for Diagnostic 2, section 7.

Lemma 6.1 (Two-sided exact WCB p-value floor; Webb, 2014; MacKinnon and Webb, 2017)**.** *Consider an exact Rademacher wild cluster bootstrap test with G clusters and a two-sided rejection rule based on $|t|$. If the bootstrap distribution is enumerated over all 2^G sign vectors and the bootstrap test statistic is sign-symmetric in the cluster-level residual contributions, then the smallest non-zero two-sided p-value is $2/2^G = 1/2^{G-1}$.*

Proof. For any sign vector $w \in \{-1, 1\}^G$, the vector $-w$ also belongs to the enumeration. Under sign symmetry, the bootstrap statistics $t^*(w)$ and $t^*(-w)$ have the same absolute value. The two-sided rejection set

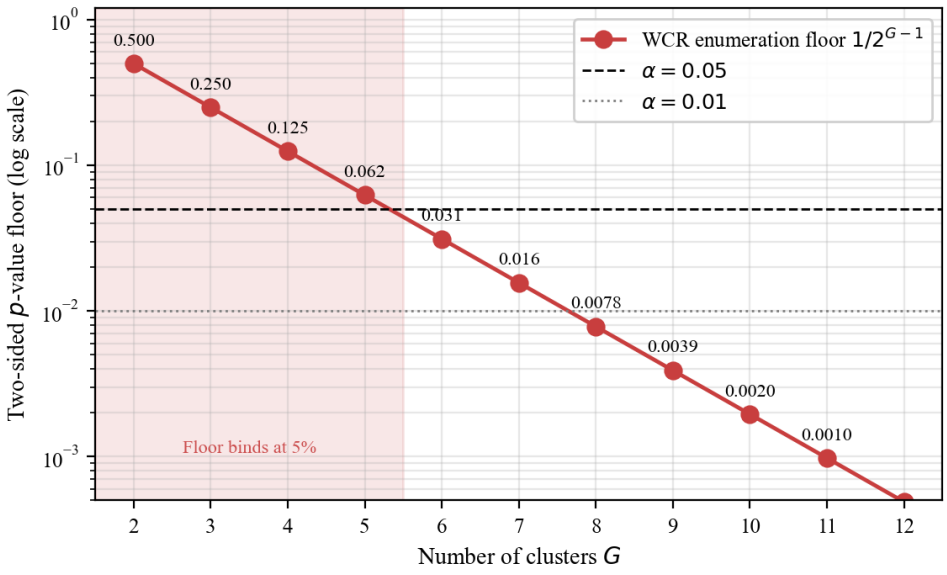
$$R = \{w : |t^*(w)| \geq |t^{\text{obs}}|\}$$

is therefore closed under negation. Any nonempty rejection set contains sign-vector pairs $\{w, -w\}$. The smallest nonempty rejection set contains exactly one such pair, giving probability $2/2^G = 1/2^{G-1}$. □

This matters directly for sovereign-debt panels with very few countries. In such settings, asymptotic cluster-robust inference may report very small p-values even when the enumerated Rademacher bootstrap reference distribution cannot reject at conventional levels. The issue is not that the coefficient is necessarily wrong in sign or magnitude. The issue is that the large-G cluster-robust approximation and the exact-enumeration bootstrap distribution are answering different inferential questions. Exact enumeration removes bootstrap simulation error; it should not be interpreted as making the test finite-sample exact under all clustered data-generating processes.

The floor is also a convention-specific statement. Randomized p-values, plus-one corrections, Webb six-point weights, and alternative bootstrap pivots can change the attainable grid. This paper reports the nonrandomized two-sided Rademacher enumeration because it matches the diagnostic implementation used below and because it makes the support limitation transparent.

Figure 2: Exact-enumeration wild cluster bootstrap p-value floor as a function of the number of clusters G



Notes: Each point is the smallest non-zero two-sided p-value achievable at G clusters, $1/2^{G-1}$. The dashed line marks the conventional $\alpha = 0.05$ threshold and the dotted line marks $\alpha = 0.01$. The shaded region indicates cluster counts for which five-percent rejection is unattainable under exact enumeration.

Figure 2 summarizes the logic graphically. Five-percent rejection is unattainable under exact two-sided enumeration when $G \leq 5$. The cluster count is therefore not a secondary implementation

detail in this application; it is a binding feature of the inferential environment.

6.2 Results on the auction-level panel

The auction panel and its Belgium–France–Italy–Portugal full-control sample are described in Section 3.1; the treatment is the baseline RRF allocation intensity of Section 3.2 interacted with the post-21 July 2020 indicator. The purpose of the table is diagnostic: it asks whether conventional clustered inference and exact finite-sample inference deliver the same conclusion in the small country panels used here. For inference on the dose-post coefficient, the effective cluster count is therefore four. A reader who sees hundreds of auction observations should not infer that asymptotic cluster-robust inference is well calibrated: what binds is $G = 4$, and the exact-enumeration p-value floor follows mechanically. Table 6 reports the resulting gap between conventional cluster-robust p-values and exact-enumeration wild-cluster bootstrap p-values for the auction-level specifications.

Table 6: Exact-enumeration wild cluster bootstrap results for the auction-level panel

ID	Specification	Outcome	G	N	$\hat{\beta}$	SE	p_{asymp}	p_{WCR}
A	4-country full controls	Bund spread	4	281	−6.196	0.389	< 0.001	0.1250
B	6-country reduced controls	Bund spread	6	413	−3.356	1.647	0.0424	0.0312
C	4-country full controls	log BCR	4	281	0.017	0.245	0.9435	0.8750
D	4-country + country trends	Bund spread	4	281	−7.567	1.850	0.0001	0.2500
E10	4-country 10Y interaction	Bund spread	4	396	−6.653	0.560	< 0.001	0.1250
E15	4-country 15Y interaction	Bund spread	4	396	−4.804	0.786	< 0.001	0.1250
E30	4-country 30Y interaction	Bund spread	4	396	−2.337	0.539	< 0.001	0.2500
F10	4-country 10Y interaction	log BCR	4	451	−0.030	0.322	0.9265	1.0000
F15	4-country 15Y interaction	log BCR	4	451	−1.152	0.173	< 0.001	0.1250
F30	4-country 30Y interaction	log BCR	4	451	−1.234	0.252	< 0.001	0.2500

Notes: The table reports conventional asymptotic cluster-robust p-values and exact-enumeration wild cluster bootstrap p-values. BCR denotes the bid-to-cover ratio, and log BCR its natural logarithm. N denotes the estimation sample after imposing the controls and lags required by each specification; it need not equal the raw auction-panel size reported in Section 3. Rows A, C, and D use the clean four-country 10-year full-controls sample (BE, FR, IT, PT), while row B uses the six-country reduced-controls 10-year sample. Rows E10–E30 and F10–F30 are all-maturity four-country specifications; the reported coefficient is the dose-post interaction for the indicated maturity bucket, not a separate tenor-only subsample. The E rows require spread controls and therefore use $N = 396$; the F rows use the raw all-maturity log-BCR panel and therefore use $N = 451$. The only specification with $p_{\text{WCR}} < 0.05$ is the six-country reduced-controls specification, and it does so exactly at the $G = 6$ floor of $1/32$.

Two patterns stand out. First, every asymptotically significant four-cluster specification lands at the four-cluster enumeration floor or one step above it, whether the outcome is the Bund spread or the log bid-to-cover ratio (BCR). The four-country headline spread specification illustrates the gap most clearly: asymptotic cluster-robust inference reports $p < 0.001$, while the exact-enumeration wild cluster bootstrap returns $p_{\text{WCR}} = 0.125$. The asymptotic approximation suggests rejection; the enumerated Rademacher bootstrap distribution does not. The two four-cluster specifications that are null under asymptotic inference—both with the bid-to-cover ratio as outcome—remain null under exact enumeration as well.

Second, the only specification that clears the five-percent threshold under exact enumeration is the six-country reduced-controls variant. That specification keeps Austria and Spain in the

sample by dropping the volatility control. Its exact p-value is 0.0312, exactly the floor implied by $G = 6$. The result therefore clears the conventional threshold, but only at the first attainable rejection point of the exact test. Because the six-country specification changes the control set to widen the country count, it should be read as a related diagnostic specification rather than as a direct replacement for the four-country full-controls design.

Companion corrections point the same way. Appendix E reports CR2 and CR3 companion corrections and leave-one-country-out sensitivity, which triangulate the same finite-sample problem: the six-country specification rejects under exact WCB but not under CR2 or CR3; the four-country headline spread result remains sensitive to the correction used; and the leave-one-country-out estimates are sign-stable but move above the five-percent threshold when Austria or France is dropped from the six-country specification. At very small G, improving the design by widening the cluster count is more valuable than treating any one small-sample correction as definitive.

6.3 Interpretation

The practical implication is straightforward. When the number of clusters is small, researchers should report the exact-enumeration wild cluster bootstrap p-value alongside the asymptotic cluster-robust p-value, and they should check the implied enumeration floor (Figure 2) before making significance claims at conventional thresholds.

This does not make small-cluster panels unusable. It means their inferential limits must be stated explicitly. An asymptotic rejection with four or five clusters is evidence under a large-G approximation, not rejection under the enumerated Rademacher bootstrap reference distribution. In the auction-panel application, that distinction is decisive: several specifications look overwhelming under asymptotic clustered inference but cannot reject under exact enumeration.

The diagnostic therefore shifts the interpretation of the auction evidence. The point estimates may still be informative about sign and magnitude, but the four-country designs do not deliver conventional five-percent rejection under the exact-enumeration Rademacher bootstrap. The six-country reduced-controls specification is suggestive, but it sits exactly at the $G = 6$ floor and changes the control set to gain clusters. The appropriate conclusion is not that the auction coefficients are false. It is that the auction-panel evidence is inferentially fragile at the available cluster counts.

7 Diagnostic 3: Pre-trend failure in monthly continuous-dose designs

This section turns from finite-sample inference to identification. A natural way to move beyond the four-country auction panel is to estimate a continuous-dose two-way fixed-effects specification on a monthly country panel that pools euro-area sovereigns with a small set of non-euro comparison units. On conventional criteria, this design initially looks attractive: it uses more countries, more periods, and a treatment variable that varies across countries in economically meaningful ways. In this application, however, the key identifying assumption does not hold.

The estimated post-2020 coefficient loads heavily on a pre-existing convergence pattern rather than on a clean treatment break.

7.1 A pooled monthly continuous-dose specification

The empirical design compares monthly Bund spreads after the July 2020 NGEU agreement across sovereigns with different RRF allocation intensity. The starting point is the standard two-way fixed-effects specification

$$\text{spread}_{c,t} = \alpha_c + \gamma_t + \beta\,(\text{intensity}_c \times \mathbf{1}\{t \geq 21\ \text{July}\ 2020\}) + \varepsilon_{c,t},$$

where $\text{spread}_{c,t}$ is the monthly Bund spread in basis points for country c in month t, α_c is a country fixed effect, γ_t is a month fixed effect, and intensity_c is the country-level RRF allocation intensity defined in Section 3.2. The panel is the monthly country panel of Section 3.1: ten non-German Bund spreads (eight euro-area recipients and three non-euro comparison units) over 2018–2024, for 840 country-month observations. Standard errors are clustered by country. In this setting the exact-enumeration Rademacher wild-bootstrap floor does not mechanically rule out five-percent rejection: with ten clusters, the smallest positive two-sided enumerated p-value is below conventional significance thresholds.

The TWFE and continuous-treatment DiD literature warns that such specifications can aggregate non-causal or non-convex comparisons when treatment effects are heterogeneous (de Chaisemartin and D'Haultfoeuille, 2020; Callaway et al., 2024). The diagnostic here is more basic: before any weighting critique can matter, the continuous-dose design must satisfy a residual parallel-trends condition.

Table 7 reports the headline results. On the conventional reading, higher RRF allocation intensity is associated with a large post-July 2020 compression of monthly Bund spreads. The estimate is statistically significant under both asymptotic cluster-robust inference and exact wild cluster bootstrap inference. It is also stable across simple sample variants: dropping the United Kingdom, restricting to euro-area countries, excluding the COVID peak, and replacing the combined grants-and-loans measure with grants-only intensity all leave a large negative coefficient. Taken at face value, Table 7 looks like a strong continuous-dose result.

Table 7: Pooled monthly TWFE continuous-dose results on Bund spreads

ID	Specification	$\hat{\beta}$ (bps)	SE	p_{asymp}	p_{WCR}	N	G
B1	Full pool (EZ + DK, SE, UK)	−426.6	181.4	0.019	0.0039	840	10
B2	Full pool, drop UK	−390.7	179.2	0.030	0.0039	756	9
B3	EZ only (no non-euro units)	−426.0	196.7	0.031	0.0156	588	7
B5	Full pool, drop 2020-02 to 2020-06	−437.5	193.6	0.024	0.0039	790	10
B7	Full pool, grants-only intensity	−805.7	259.5	0.002	0.0098	840	10

Notes: The dependent variable is the country-month 10-year spread over the German Bund in basis points. Treatment is the RRF allocation-intensity measure from Table 12 interacted with a post-21 July 2020 indicator. All specifications include country and month fixed effects; standard errors are clustered by country. Row B5 drops February–June 2020, removing five months across ten spread series. p_{WCR} is the exact-enumeration Rademacher wild cluster bootstrap p-value.

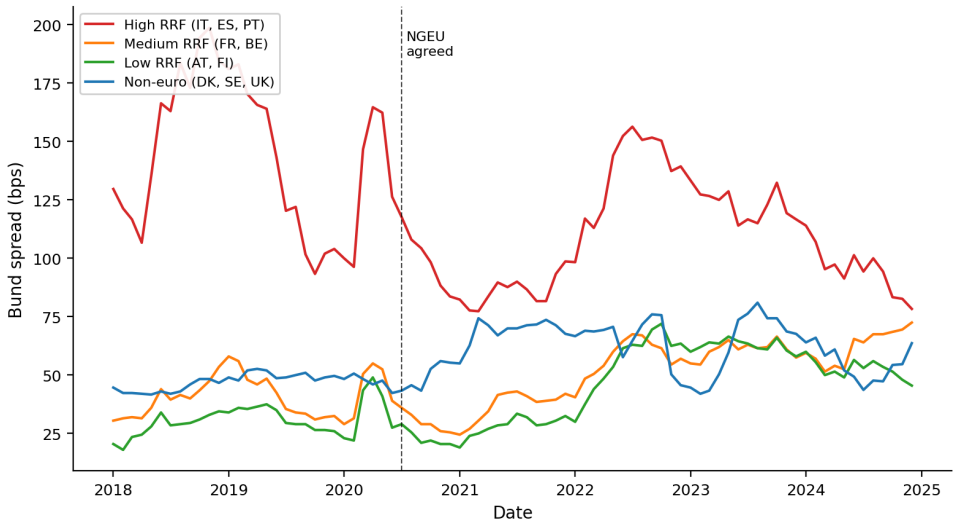
That reading is misleading. Robustness across sample variants is not enough in a continuous-dose design. Identification requires a credible parallel-trends assumption for a continuous country-level treatment: absent treatment, higher- and lower-intensity countries would have followed the same residual spread path, conditional on country and time fixed effects. The remainder of the section shows that this condition fails.

7.2 Pre-trend evidence

Because RRF allocation intensity is fixed across time, the parallel-trends requirement reduces to a restriction on residual spread paths: high- and low-intensity countries should not have systematically different pre-treatment trajectories. Two pieces of evidence, one informal and one formal, show that this restriction fails.

A first look at the raw paths already strains it. Figure 3 plots the average monthly Bund spread for four groups: high-exposure countries (Italy, Spain, Portugal), medium-exposure countries (France, Belgium), low-exposure euro-area countries closest to Germany (Austria, Finland), and the three non-euro comparison units (Denmark, Sweden, United Kingdom). These high/medium/low labels are RRF-allocation groups for the monthly diagnostic and differ from the peripheral/core/non-euro correlation groups used in Diagnostic 1. The high-exposure group rises sharply in 2018 and then declines steadily through 2019, while the other groups remain much flatter, so the convergence of the high-exposure group begins well before the July 2020 NGEU agreement. This pattern is not, by itself, a formal rejection, but it is exactly what should make a continuous-dose interpretation fragile: if higher-exposure countries are already on a different spread trajectory before treatment, the post-2020 coefficient may partly capture the continuation of that trajectory rather than the effect of NGEU.

Figure 3: Average monthly Bund spreads by RRF allocation-intensity group, 2018–2024



Notes: The dashed vertical line marks the 21 July 2020 European Council agreement. The high-intensity group (Italy, Spain, Portugal) is already on a downward trajectory before the treatment date. Source: Eurostat monthly sovereign yield data.

The formal check replaces the single dose-post interaction with event-time bins relative to

July 2020 and tests whether the pre-treatment bins are close to zero (see Athey and Imbens, 2017; Roth, 2022; Roth et al., 2023; Rambachan and Roth, 2023). Table 8 reports the binned event-study coefficients for the full-pool ten-cluster specification. The omitted reference bin is the twelve months immediately before treatment, July 2019 through June 2020; the specification includes two pre-event bins and four post-event bins.

Table 8: Binned event-study coefficients on RRF allocation intensity

Bin	Period relative to July 2020	$\hat{\beta}$ (bps)	p_{asymp}	p_{WCR}
pre_{2+}	$\geq$ 24 months before	294.8	0.048	0.023
pre_{2y}	12 to 24 months before	544.4	0.021	0.002
Reference	1 to 12 months before	0	—	—
post_{h1}	0 to 6 months after	−85.9	0.372	0.543
post_{1y}	6 to 18 months after	−334.8	0.037	0.031
post_{2y}	18 to 30 months after	42.7	0.337	0.354
post_{3y+}	$\geq$ 30 months after	−169.7	0.320	0.293

Notes: Coefficients are in basis points per unit RRF allocation intensity. The outcome is the monthly Bund spread in basis points. Standard errors are clustered by country; p_{WCR} is computed by exact Rademacher wild-cluster enumeration. The omitted reference window is the twelve months immediately before July 2020. Source: Eurostat monthly sovereign yield panel (`IRT_LT_MCBY_M`); author's calculations.

The central pattern is that the pre-event coefficients are large, positive, and significant under the exact wild cluster bootstrap, and economically large relative to the post-event coefficients. The bin covering twelve to twenty-four months before July 2020 delivers $\hat{\beta}_{\text{pre}_{2y}} = 544.4$ basis points per unit of RRF allocation intensity, with $p_{\text{WCR}} = 0.002$; the earlier pre-treatment bin is also positive and significant. These estimates mean that high-intensity sovereigns were already on a different residual spread path before NGEU.

The implied country-level magnitudes anchor the result. For Italy at intensity 0.107, the linear implication of the twelve-to-twenty-four-month pre-event coefficient (544.4 bps per unit) is approximately 58 basis points; for Spain at intensity 0.130, the implied deviation is approximately 71 basis points; for Portugal at intensity 0.077, approximately 42 basis points. The implied pre-bin magnitudes are consistent in order of magnitude with observed peripheral spread movements over 2018–2019. For Italy in particular, the 58 basis-point linear implication is roughly half of the actual 120 basis-point narrowing in the BTP–Bund spread between mid-2018 and mid-2019; the remaining difference reflects partialling-out by country and time fixed effects and the use of twelve-month averages rather than point-in-time values. In plain terms, high-exposure sovereigns were already converging toward the German Bund before the recovery fund existed.

The evidence makes the residual parallel-trends assumption untenable for this continuous-dose specification. The pre-treatment bins are not small deviations around zero. They are economically large, statistically meaningful under exact-enumeration Rademacher wild cluster bootstrap inference, and inconsistent with the residual parallel-trends condition needed for a causal interpretation of the post-treatment coefficient.

This is the central identification result of the section. The continuous-dose specification is not failing because of weak power or because asymptotic inference is unreliable. It is failing because

treatment intensity is aligned with pre-existing outcome dynamics. The identifying variation is contaminated before the treatment date.

7.3 Diagnostic triangulation: robustness and sensitivity

Four companion exercises triangulate the same failure mode. Appendix F reports first-difference estimates, alternative binning and reference-window choices, Rambachan–Roth sensitivity analysis, and leave-one-country-out sensitivity. Each asks a different question about the same pre-trend diagnostic.

First, the full-sample first-difference specification collapses the headline coefficient toward zero, with an estimate of roughly -12 basis points per unit intensity and an enumerated wild-bootstrap p-value above 0.4. The level estimate therefore does not survive a generic first-difference robustness check on the same panel. Second, eight of nine alternative binning and reference-window configurations produce at least one pre-event bin significant at the five-percent WCR level, so the pre-trend is not an artifact of a particular bin width or reference period. Third, the Rambachan and Roth (2023) sensitivity analysis shows that no practically relevant bounded relaxation of parallel trends leaves a post-treatment interval excluding zero: every post-treatment confidence interval already includes zero at the smallest feasible relative-magnitude bound, roughly $\bar{M} = 800$ basis points per unit intensity.[3] Fourth, the leave-one-country-out exercise shows that Italy is the only country whose removal flips the headline significance at five percent, while dropping Spain increases the magnitude of the level coefficient by roughly two-thirds.

Together, these checks leave little room for the monthly continuous-dose specification to be read as a calibrated test of the NGEU hypothesis. The result is not driven by one arbitrary binning choice, one standard-error correction, or one visual judgment about the raw paths. It is a robust pre-treatment imbalance in the identifying variation itself.

7.4 Interpretation and what survives

The conclusion is straightforward. The monthly continuous-dose specification does not identify the causal effect of RRF allocation intensity on sovereign spreads in this panel. Its headline coefficient comes from a design in which treatment intensity is systematically aligned with pre-treatment convergence. The estimate looks strong under clustered inference because the main problem is not standard-error understatement. It is identification.

The pre-trend pattern is consistent with the unwinding of earlier euro-area stress, the resolution of the 2018 Italian fiscal confrontation, and the broader re-rating of peripheral sovereign risk that was already visible before NGEU was announced. A specification that cannot distinguish these dynamics from a causal NGEU effect is not a calibrated test of the NGEU hypothesis; it

[3] Appendix F.3 reports the feasibility threshold for transparency. In this application it nearly coincides with the standard Rambachan and Roth (2023) breakdown value—the smallest $\bar{M}$ at which a post-treatment confidence interval first includes zero under $\Delta^{RM}(\bar{M})$—because at the first feasible $\bar{M}$ every post-treatment interval already includes zero; the standard breakdown value lies in the same range across the four post-treatment bins and never signals a non-zero post-treatment effect.

is partly re-labelling the pre-existing trend. A useful summary statistic is the ratio of pre- and post-period coefficient magnitudes within the same continuous-dose specification. In this panel, the pre-period magnitude is larger than the headline post-period magnitude, so the pre-existing drift is larger than the coefficient one would otherwise interpret as the treatment effect.

The first-difference evidence reinforces that interpretation. If the post-2020 coefficient reflected a discrete treatment break operating on otherwise parallel spread dynamics, one would expect some corresponding signal in monthly spread changes. Instead, the pre-event bins are large and the full-sample first-difference estimate moves close to zero. That pattern is more consistent with pre-existing peripheral convergence than with a clean continuous-dose NGEU effect.

This conclusion is specific to the design, not to the broader economic question. A continuous-dose specification can work in settings where treatment intensity is not correlated with outcome dynamics. In this panel, that condition is not satisfied. The monthly coefficient should therefore not be read as a causal estimate of NGEU-induced spread compression.

A further limitation remains in the background. Even if parallel trends held, the coefficient would still capture NGEU together with other euro-area-specific post-2020 forces, including PEPP, relative to non-euro comparison units. The identifying failure documented here comes before that broader interpretation problem. The pre-treatment side of the design already fails.

The pre-trend failure is not driven by the three non-euro comparison units, which—as Section 3.1 explains—serve only to widen the cross-section and carry zero exposure by diagnostic convention. Restricting to the euro-area sample leaves the binned pre-treatment coefficients large, positive, and WCR-significant, so the identification failure is intrinsic to the euro-area cross-section rather than a comparison-unit artifact.

The connection to the policy-endogeneity concern in Section 4 is direct. In the euro-area recipient sample, larger RRF allocation intensity is concentrated in countries whose pre-existing macroeconomic conditions, including lower GDP per capita, higher unemployment, and greater crisis exposure, also predict the pre-2020 convergence trajectory. The diagnostic captures the empirical trace of that endogeneity. Section 8 asks whether replacing observed exposure with a formula-based, predetermined RRF intensity measure can solve the problem. It cannot.

8 Diagnostic 4: Formula-based exposure inherits the pre-trend

Diagnostic 3 showed that the monthly continuous-dose specification fails because measured RRF allocation intensity is aligned with pre-treatment spread dynamics. A natural response is to replace observed exposure with a formula-based measure built from predetermined country characteristics. In principle, such a design could isolate variation in RRF allocation intensity that is fixed before the pandemic. In this application, however, the formula-based design does not solve the identification problem. The same country characteristics that generate the formula-based RRF intensity measure are closely related to the pre-treatment convergence pattern in sovereign spreads.

8.1 The formula-based exposure measure

The formula-based exposure measure used in the NGEU literature is not a multi-sector Bartik instrument in the sense of Adão et al. (2019) or Borusyak et al. (2022); it is a one-shot cross-sectional exposure index built from observable country characteristics. Borusyak et al. (2025) distinguish two routes to shift-share identification, one resting on many exogenous shifts and one on exogenous shares, and the formula-based measure falls in the second. There are no time-varying sectoral shocks here to argue are quasi-random, so identification, if available, must come from the assumption that the cross-sectional exposure index is unrelated to the potential trajectory of sovereign spreads—the exposure-based logic of Goldsmith-Pinkham et al. (2020), under which a Bartik instrument is a pooled exposure design whose validity turns on the exogeneity of the shares. That assumption bites here whenever the formula inputs predict pre-treatment outcome dynamics, so the balance and reduced-form checks below are framed as exogenous-shares diagnostics rather than as classical shift-share tests.

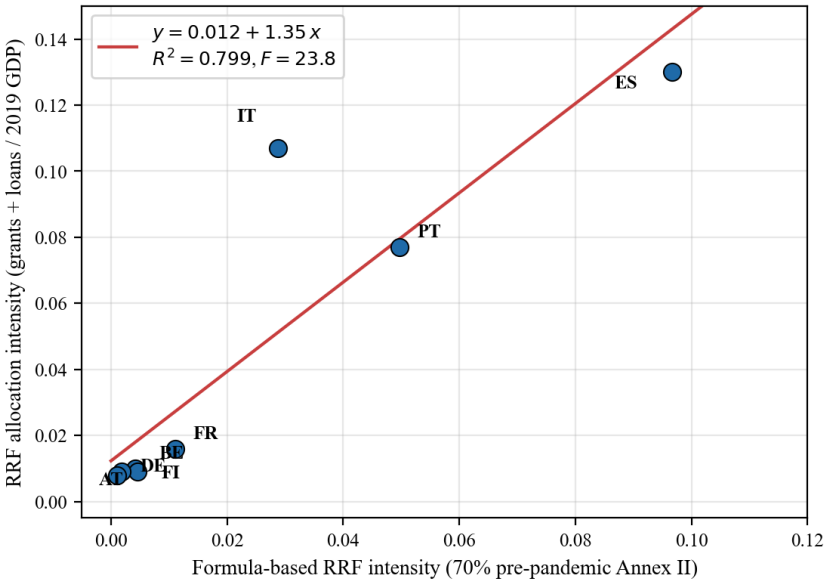
The measure is built from the pre-pandemic component of the formula used to allocate Recovery and Resilience Facility grants across Member States (European Parliament and Council, 2021). That component governs 70% of the RRF grant envelope and combines three pre-pandemic inputs—population, the inverse of 2019 GDP per capita, and the 2015–2019 average unemployment rate relative to the EU average—subject to caps and safeguards against excessive concentration of resources. Following the Annex II allocation logic and the implementation of Darvas (2020, 2021), the relevant pre-pandemic tranche is 70% of the EUR 312.5 billion (2018-price) grant envelope, or EUR 218.75 billion; the resulting country weights are converted into an exposure intensity by scaling by 2019 GDP. Figure 5 plots this formula-based RRF intensity against the baseline grant-plus-loan RRF allocation intensity for the eight euro-area recipients, with the underlying inputs and outputs tabulated in Appendix G.

The first stage is strong by conventional standards. A cross-sectional regression of baseline grant-plus-loan exposure on the formula-based RRF intensity and a constant produces $R^2 = 0.799$ and a homoskedastic F-statistic of 23.8. The measure also ranks countries in the expected economic order: Spain, Portugal, and Italy receive the largest formula intensities, while Germany, Austria, and Finland receive the smallest.

A strong first stage is not enough. In a formula-based exposure design, the central question is whether the exposure measure is orthogonal to unobserved determinants of the outcome trajectory. The rest of the section shows that, in this panel, it is not.

Inference here cannot lean on the standard shift-share correction. Adão et al. (2019) propose standard errors that account for correlation in shocks across units sharing similar exposures, but that AKM correction is valid under the assumption of many independent shocks across exposure cells, which fails outright in a cross-section of ten units. With $N = 10$, no asymptotic standard-error formula or resampling procedure is informative on its own. The diagnostic value of the reduced-form evidence (Section 8.2) therefore rests on the panel pre-bin coefficients in Table 9, which use the monthly cross-section of ten countries, combined with the exact 10! permutation test of the balance regression introduced in Section 8.3, rather than on any single

Figure 4: Formula-exposure first stage: baseline grant-plus-loan RRF allocation intensity plotted against the formula-based RRF grant-intensity measure for the eight euro-area recipients



Notes: The red line is the OLS fit through a constant. The fit is dominated by the three peripheral countries (Italy, Spain, Portugal), with Italy sitting below the fitted line because its baseline exposure includes EUR 122.6 billion in RRF loans that the 70% pre-pandemic grant-allocation formula does not weight.

asymptotic standard error.

8.2 The reduced form inherits the pre-trend

To test the design, this subsection reruns the binned event study from Diagnostic 3 using the formula-based RRF intensity measure in place of observed RRF allocation intensity. The panel, event-time bins, and omitted reference period are unchanged. If the formula-based measure corrected the identification problem, the pre-treatment bins would become small and uninformative.

Table 9: Reduced-form binned event-study coefficients with formula-based exposure

Bin	Period relative to July 2020	$\hat{\beta}^F$ (bps)	p_{asymp}	p_{WCR}
pre_{2+}	$\geq$ 24 months before	520.3	0.063	0.0020
pre_{2y}	12 to 24 months before	538.3	0.106	0.0020
Reference	1 to 12 months before	0	—	—
post_{h1}	0 to 6 months after	−7.5	0.904	0.9043
post_{1y}	6 to 18 months after	−287.1	0.059	0.0312
post_{2y}	18 to 30 months after	41.1	0.518	0.6426
post_{3y+}	$\geq$ 30 months after	−182.6	0.427	0.3418

Notes: The table reports reduced-form binned event-study coefficients using formula-based RRF intensity in place of observed RRF allocation intensity. The specification, sample, and cluster count match Table 8. Coefficients are in basis points per unit formula-based RRF intensity. Standard errors are clustered by country; p_{WCR} is computed by exact-enumeration Rademacher wild-cluster bootstrap.[4]

[4]The pattern that asymptotic p-values are larger than exact wild cluster bootstrap p-values may surprise readers familiar with small-cluster inference, where the WCR is usually more conservative than asymptotic. The reversal

They do not. Table 9 shows large positive pre-treatment bins. The bin covering more than 24 months before July 2020 is estimated at 520.3 basis points per unit formula-based exposure, and the bin covering 12 to 24 months before treatment is estimated at 538.3. The exact WCR p-values are reported for completeness, but the inference is interpreted together with the size and sign pattern of the pre-bins, the balance checks, and the Spain influence diagnostics rather than treated as decisive on their own.

The headline reduced-form coefficient on the full dose-post term is also large and negative: approximately -456 basis points per unit formula-based exposure, with an exact-enumeration Rademacher wild-cluster bootstrap p-value below 1%. Considered in isolation, that coefficient would look encouraging. The event-study decomposition shows why that reading is misleading. The formula-based measure does not purge the pre-trend; it inherits it.

This is the core result of the section. The formula-based design does not fail because the exposure measure is weakly related to baseline RRF allocation intensity. It fails because the formula inputs are correlated with the pre-treatment trajectory of sovereign spreads.

8.3 Balance and influence checks

A complementary way to assess the design is to test whether formula-based exposure is balanced with respect to pre-treatment spread outcomes. If the exposure measure were as-good-as-random conditional on the fixed effects used in the panel design, pre-treatment changes in spreads should not be systematically related to the formula-based RRF intensity.

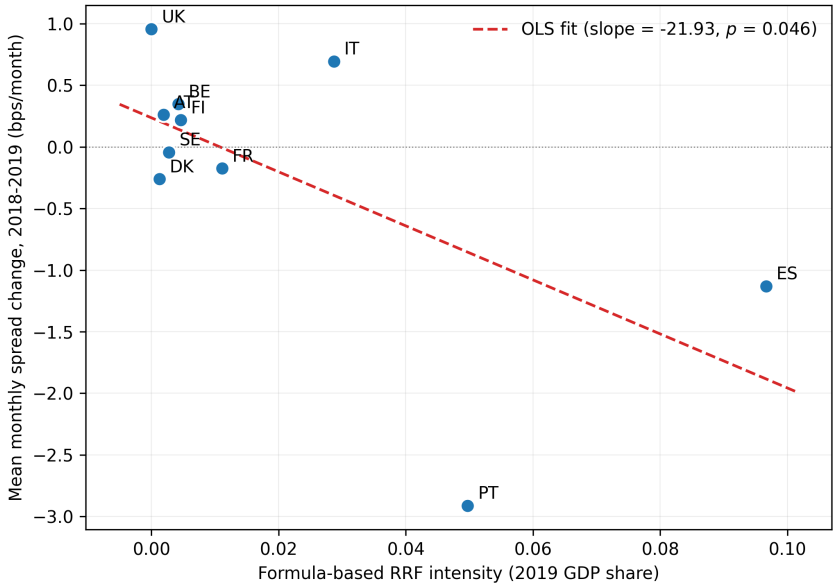
The balance regression is a cross-section of ten units. With this sample size, conventional standard errors are themselves unreliable, and the analysis therefore supplements the asymptotic p-value with an exact randomization test that enumerates all $10! = 3{,}628{,}800$ permutations of the formula-intensity vector across the ten cross-sectional units. The test asks how unusual the realized balance slope is relative to alternative alignments of formula intensity with the same pre-treatment outcome vector.

Figure 5 plots formula-based exposure intensity against the 2018–2019 mean monthly spread change, one point per country. The full balance regression and the exact 10! permutation test are reported in Appendix G. The point estimates indicate that countries with higher formula-based exposure experienced larger pre-treatment spread declines: the coefficient on mean monthly spread changes is -22 basis points, with an asymptotic HC1 p-value of 0.046 but an exact permutation p-value of 0.100. The level and trend balance regressions yield exact permutation p-values of 0.166 and 0.106 respectively. Under the exact test, no balance regression is significant at the 5% level. The qualitative pattern (negative-slope alignment of exposure with pre-trend tightening) remains, but the cross-section is too thin to deliver decisive inference: the suggestive HC1 significance of the monthly-change slope does not survive exact enumeration. This is itself

here arises because the t-statistic distribution is concentrated under the alternative on a small set of distinct exposure values: with $G = 10$ clusters and two clusters at near-zero formula intensity, the WCR statistic distribution becomes sharply bimodal, generating a small tail mass at the observed statistic and hence a smaller bootstrap p-value than the asymptotic approximation.

an illustration of the broader point of the paper, that asymptotic standard errors are not reliable in this regime.

Figure 5: Formula-based RRF intensity plotted against the 2018–2019 mean monthly spread change, one point per country



Notes: The fitted line summarizes the monthly-change balance regression reported in Appendix G: countries with larger formula-based exposure also have larger pre-treatment spread declines. This is the central concern behind the exogeneity-of-shares logic of Goldsmith-Pinkham et al. (2020). Sources: formula-based RRF intensity constructed from the Annex II inputs in Regulation 2021/241 following Darvas (2020, 2021); pre-treatment monthly spread changes estimated from daily 10-year benchmark yields aggregated to monthly Bund-spread means over 2018-01 to 2019-12; author's calculations.

Appendix G also reports leave-one-country-out sensitivity. The reduced form is sign-stable but highly sensitive to Spain: dropping Spain triples the magnitude of the coefficient and pushes the asymptotic p-value from 0.107 to 0.0002. No other country drop changes significance at five percent. The balance failure and the Spain-driven influence pattern point to the same problem: the formula-based design is contaminated by pre-trends in the aggregate and is unusually sensitive to one high-exposure country with a steep pre-treatment spread path.

8.4 Why the formula-based design fails

Predeterminedness is necessary but not sufficient. The formula inputs—2019 GDP per capita, 2015–2019 unemployment, and 2019 population—are mechanically fixed before the pandemic and before the July 2020 NGEU agreement, so they are predetermined relative to the policy date. But the identifying condition is stronger: the exposure index built from those inputs must be unrelated to the trajectory that spreads would have followed absent NGEU. Predetermined macroeconomic variables that predict pre-treatment sovereign-spread dynamics do not satisfy that condition.

The reason is economic rather than mechanical. The formula tends to assign larger grant exposure to countries with lower GDP per capita and higher pre-pandemic unemployment. Those are also the countries whose spreads were tightening in 2018 and 2019 for reasons

unrelated to NGEU: the resolution of the 2018 Italian fiscal confrontation, the ongoing unwinding of euro-area redenomination risk, and the broader re-rating of peripheral sovereign debt. The formula-based exposure measure therefore selects the same countries whose spreads were already on a convergence path before the policy.

The general lesson is familiar from the shift-share literature. A strong first stage does not validate an exposure design when the exposure shares are correlated with pre-treatment outcome dynamics (Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022). In such cases, the reduced form can inherit the same parallel-trends failure as the original OLS design. The contribution of this section is to show that this concern is binding in the NGEU sovereign-spread application.

This failure is an empirical manifestation of the policy-endogeneity problem discussed in Section 4. The formula's inputs are crisis-relevant macroeconomic variables. They predict RRF allocation intensity by construction and pre-treatment sovereign-spread dynamics empirically. The formula-based design therefore improves relevance but not exogeneity.

8.5 Interpretation

The formula-based design does not rescue identification. It has a strong first stage but a reduced form that reproduces the pre-treatment pattern invalidating the continuous-dose specification of Diagnostic 3: the pre-event bins remain large and positive, the balance slope aligns in the same direction, and the estimate is highly sensitive to Spain. Had the formula isolated an exogenous component of RRF allocation intensity, the pre-event bins would be small even with a large post-event coefficient.

The conclusion is narrow but clear. In this application, the formula-based exposure measure does not isolate exogenous variation in RRF allocation intensity; it repackages the same cross-country differences that were already driving spread convergence before treatment. Section 9 collects what the four diagnostics imply for the published NGEU sovereign-spread evidence.

9 Implications for published NGEU sovereign-spread findings

The four diagnostics developed in Sections 5, 6, 7, and 8 imply a more selective reading of the NGEU sovereign-spread evidence than conventional inference would suggest. One high-frequency result remains clearly persuasive under the benchmarked standards developed here. Several other reduced-form findings or natural extensions do not survive the relevant diagnostic. A further set of contributions is either exempt from these diagnostics or outside their direct scope because the empirical object differs from the sovereign-spread regressions studied in this paper. This section states that classification explicitly.

9.1 Classification framework

The classification uses four labels. A finding *survives* if it remains persuasive under the diagnostic appropriate to its design. For event-study cross-sections, this means that the estimated coefficient lies outside the benchmark placebo 95% interval. For small-cluster panel designs, it means that

the exact-enumeration Rademacher wild cluster bootstrap p-value can support rejection at the relevant threshold. A finding *does not survive* if it does not meet the corresponding benchmark standard. A finding is *exempt* if the statistic under discussion is model-implied rather than a direct reduced-form regression coefficient, so that the diagnostics developed here do not apply mechanically. A finding is *outside scope* if the outcome variable or empirical design differs sufficiently from the sovereign-spread regressions examined here that no direct re-evaluation is attempted.

Two clarifications are important. First, "does not survive" does not mean that the original estimate is necessarily wrong in sign, magnitude, or economic interpretation. It means that, given the sample sizes and inferential problems documented in this paper, the observed pattern cannot be distinguished with sufficient confidence from the relevant benchmark null, or that the identifying assumption fails in the diagnostic design. Second, survival is design-specific. A result may survive under the diagnostic that matches the design actually used in the literature without implying that it would survive under a different, hypothetical specification.

9.2 Findings that survive

One finding remains clearly persuasive under the benchmark event-study standard: the spread response to the 18 May 2020 Franco-German recovery-fund proposal, estimated here in the author's daily cross-section, with Havlik et al. (2022) as the closest published analog. In the cross-section benchmark of Section 5, the $S \pm 1$ and $S \pm 2$ symmetric-window coefficients lie outside the clean 2018–2019 placebo 95% interval; the $S \pm 1$ coefficient sits deep in the left tail of the placebo distribution, and only at $S \pm 5$ does the estimate fall back inside the wider placebo range.

This is the strongest event-study pattern in the paper, and it is also the most stable one. In the leave-one-country-out exercise reported in Section 5, the sign and conventional significance of the $S \pm 1$ Franco-German estimate are unchanged regardless of which country is removed. The close-to-close timing check is smaller, so the finding should be read as a short-window announcement response rather than as a literal one-day close-to-close effect. For the daily sovereign-spread cross-section, this is therefore the most defensible positive reduced-form finding in the NGEU event-study evidence.

9.3 Findings that do not survive the benchmark diagnostic

The classification that follows is not a statement about economic mechanisms. It is a statement about what a specific reduced-form design can distinguish from its benchmark null at the available sample size.

The remaining high-frequency NGEU dates do not survive the placebo-benchmarked event-study standard, though the 21 July 2020 Council agreement warrants more than a one-line verdict. Its coefficient lies inside the day-to-day placebo distribution at all windows but outside the cross-sectional permutation distribution at $S \pm 2$ ($p_{\text{perm}} = 0.0085$; see Appendix H). The two tests answer different nulls—the permutation test asks whether the realized alignment

of intensity and spread changes on that date is unusual relative to alternative alignments, while the placebo distribution asks whether the event-day coefficient is unusual relative to ordinary pre-NGEU common-factor movements—and the placebo result is the more relevant object here because it preserves the realized cross-sectional dependence structure that drives HC1 miscalibration. The Council agreement therefore produced an exposure-aligned spread movement, but one whose magnitude is not unusual against the calibrated placebo benchmark. The 14 December 2020 Recovery Instrument regulation likewise produces HC1 p-values that would look significant at short symmetric windows, yet its coefficients remain inside the placebo range, and the 15 June 2021 first NGEU issuance is null by any standard, with small estimates that switch sign across windows.

The monthly continuous-dose design studied in Section 7 does not survive the pre-trend diagnostic. Its headline coefficient is large, negative, and statistically strong under clustered inference, but the identifying assumption fails. The binned event-study decomposition shows large and significant pre-treatment coefficients of the opposite sign. The full-sample first-difference specification collapses toward zero. The pre-trend result is robust to alternative binning choices, and the Rambachan and Roth (2023) exercise shows that no practically relevant bounded relaxation of parallel trends leaves a post-treatment interval excluding zero. The problem is therefore not weak power. It is identification.

The formula-based exposure design of Section 8 fails for the same underlying reason. The exposure measure has a strong first stage for baseline RRF allocation intensity, but its reduced form inherits the pre-treatment convergence pattern rather than purging it. The balance regressions yield negative point estimates on pre-treatment monthly changes and on the daily trend slope, but the cross-section is too thin to deliver decisive inference: exact 10! permutation p-values are 0.100 on monthly changes and 0.106 on the trend slope, neither significant at 5%. The reduced-form pre-bins are large, positive, and significant under exact-enumeration wild cluster bootstrap, and the result is highly sensitive to Spain. In this application, the formula-based exposure design improves relevance but not exogeneity, even if the balance evidence on its own is suggestive rather than decisive.

Finally, the four-country auction specifications studied in Section 6 do not support five-percent rejection under exact-enumeration Rademacher wild cluster bootstrap inference. Their asymptotic cluster-robust p-values are extremely small, but with $G = 4$ the enumerated two-sided Rademacher bootstrap cannot produce a positive p-value below 0.125. Those estimates may still be suggestive under a large-G approximation, and the point estimates may remain informative about sign and magnitude. They do not, however, support conventional five-percent rejection under the enumerated bootstrap reference distribution used in this paper.

9.4 Findings exempt or outside scope

Some important contributions are not direct targets of the diagnostics developed here. The model-based sovereign-premium decomposition in Corradin and Schwaab (2023) is best treated as *exempt*. Its key objects are model-implied premia rather than direct cross-section regression

coefficients, so the placebo benchmark in Section 5 does not apply mechanically. The same is true of the structural and semi-structural macro analyses in Bańkowski et al. (2022), Bańkowski et al. (2024), and Millard (2025). These papers study economically important mechanisms and provide quantitative estimates of NGEU or RRF effects, but they are not reduced-form identification exercises of the type evaluated in this paper.

Other papers are better described as *outside scope*. The bank-stock, bank-CDS, and sovereign-CDS results in Pancotto et al. (2023) use different outcomes from the sovereign-spread regressions studied here. The staggered panel design in Alessi et al. (2025) uses a different treatment, based on the greenness of national recovery plans, and a different timing structure. The diagnostic logic of Section 7 is relevant by analogy, since treatment intensity and pre-existing dynamics can also interact in staggered or continuous-dose designs, but this paper does not re-estimate that specification.

9.5 What this classification does and does not imply

Three features of the classification are worth emphasizing. First, it is not a ranking of papers or authors, and it does not claim that any non-surviving estimate is wrong in sign, magnitude, or economic interpretation. It asks a narrower, design-specific question: given the sample sizes and empirical designs in this literature, which findings remain persuasive once inference is benchmarked and identifying assumptions are probed directly? The interpretive limits of the diagnostics bear directly on that question. Placebo-date calibration preserves the realized exposure vector and pre-treatment common-factor structure by design; the non-euro sovereigns serve as diagnostic calibration rather than as clean counterfactuals; and the wild cluster bootstrap floor is an inference limitation, not a statement about the true coefficient.

Second, the benchmarked evidence base is thinner than the conventional reading would suggest, but it is not empty. The Franco-German proposal remains the main positive reduced-form event-study finding; the other NGEU dates may still have mattered economically, but the small sovereign-spread cross-section studied here does not distinguish their observed patterns from placebo variation with sufficient confidence.

Third, the message is diagnostic rather than destructive. The paper proposes no new estimators and does not re-replicate every published specification; it benchmarks the designs actually used, and shows that several strategies which look convincing under standard practice become less persuasive once finite-sample inference and design-specific identification are taken seriously. That is also why the contribution is framed as design benchmarking rather than a verdict on NGEU itself. It points naturally to the next section, which develops a practical inference toolkit for small-sample macro-finance panels and event studies.

10 An inference toolkit for small-sample panels and event studies

The diagnostics above imply a practical reporting standard. The relevant check depends on the sample structure, the identification strategy, and the effective sample size. A daily event

study with a thin cross-section primarily needs placebo-date benchmarking; a panel with few countries primarily needs finite-cluster inference; continuous-dose and formula-based designs need explicit pre-trend and balance diagnostics. The common principle is that the diagnostic should match the failure mode of the design rather than the number of rows in the dataset.

10.1 Decision framework

Three questions should be asked before interpreting a small-sample result. First, what is the frequency and sample structure? Daily event studies raise different problems from monthly or quarterly panels, and mixed designs may require both checks. Second, what is the identification strategy? Cross-section event studies call for placebo-date and, when feasible, permutation benchmarks; continuous-dose panels call for pre-trend diagnostics; formula or shift-share exposure designs require balance and reduced-form checks aimed at the exposure measure itself. Third, how small is small? Here the effective sample size—the number of clusters, or the size of the event-study cross-section—determines which finite-sample tool is needed, not the number of rows in the dataset. Table 10 sets out the minimum diagnostic response for each design and cluster regime.

Table 10: Decision framework for small-sample inference and identification diagnostics

Design / sample regime	Main risk	Minimum diagnostic response
Panel with $G \leq 5$ country clusters	WCB enumeration floor binds	Report the exact WCB floor; avoid 5% claims under non-randomized two-sided Rademacher enumeration; add CR2/CR3 and leave-one-cluster-out checks
Panel with $6 \leq G < 30$ country clusters	Cluster-robust approximation may be fragile	Report WCB, CR2/CR3, and influence checks; do not rely on conventional CRVE alone
Panel with $G \geq 30$ country clusters	Standard asymptotics more plausible	Cluster-robust SEs are usually defensible as a rule of thumb, but influence and leverage checks remain useful
Continuous-dose or staggered panel	Treatment intensity correlated with pre-trends	Report binned pre-treatment event-study coefficients and sensitivity analysis such as Rambachan–Roth when pre-bins are large
Formula or shift-share exposure	Exposure inherits pre-treatment dynamics	State the exogenous-shares logic; test balance and reduced form on pre-outcomes; use exact $N!$ permutation when N is small
Cross-section event study with $N \leq 15$	HC-type inference miscalibrated by cross-sectional dependence	Benchmark the intended test on placebo dates; report the placebo rejection rate; use permutation inference when support and exchangeability allow

Notes: The $G \geq 30$ threshold is a rule of thumb consistent with Cameron and Miller (2015) and MacKinnon and Webb (2017), not a formal threshold in either source.

10.2 Minimum reporting standard

The toolkit can be summarized as a minimum reporting standard for small-N macro-finance designs:

1. Report the effective cluster count or event-study cross-section size before interpreting

standard errors.

2. Report the finite-sample diagnostic appropriate to that sample size. With $G \leq 5$, state the exact Rademacher WCB floor explicitly.

3. Report influence checks, such as leave-one-cluster-out or leave-one-unit-out estimates.

4. For continuous-dose or staggered panel designs, report pre-treatment event-study bins before interpreting the post-treatment coefficient.

5. For formula-based or shift-share exposure designs, report balance and reduced-form checks on pre-outcomes, including exact $N!$ permutation when N is small.

6. For small cross-section event studies, benchmark the intended test on placebo dates and report the placebo rejection rate for the same inference procedure used on the event date.

This reporting standard is intentionally modest. It does not make empirical work impossible; it aligns the strength of the inferential claim with the information content of the design. Appendix I lists software implementations for the main diagnostics.

11 Portability beyond NGEU and sovereign debt

The diagnostics are not specific to the NGEU sovereign-spread setting. They target empirical conditions that arise across applied work: treatment assigned or measured at a small number of aggregate units, high-frequency event windows evaluated in thin cross-sections, continuous treatment intensity aligned with pre-treatment dynamics, and exposure measures constructed from predetermined shares or formula inputs. The NGEU application is useful because all four conditions appear in one setting. In other literatures, usually only one or two bind at a time. The relevant lesson is therefore not that other literatures suffer from the same failures documented here. It is that the same diagnostic logic applies whenever the same design structure appears.

11.1 Shift-share and exposure designs

A first example is shift-share and exposure-based work, including the China-shock literature associated with Autor et al. (2013). In those designs, exposure is built from pre-period regional shares interacted with common shocks. The identifying question is whether the shares, or the formula inputs that play the same empirical role, are as-good-as-random with respect to potential outcome trends. The exogenous-shares interpretation in Goldsmith-Pinkham et al. (2020) and the shift-share inference results of Adão et al. (2019), Borusyak et al. (2022), and Borusyak et al. (2025) imply the same diagnostic discipline used here for formula-based RRF intensity. Researchers should state the source of identifying variation, test whether exposure predicts pre-treatment outcomes or trends, and use inference that accounts for the dependence structure induced by common shocks when the design requires it. The point is not that China-shock designs are equivalent to the NGEU formula design. The point is that predetermined exposure is not, by itself, an identification argument. A formula can be fixed before treatment and still sort units by structural characteristics that predict the outcome path.

This distinction is especially important in policy settings where allocation rules are designed to target need. A grant formula based on unemployment, income, population, or crisis exposure may be institutionally predetermined and economically sensible precisely because those variables identify vulnerable units. But the same variables may also predict pre-treatment outcomes. In that case, a strong first stage confirms relevance but does not establish exogeneity. The portable diagnostic is therefore to run the reduced form and the balance checks on pre-outcomes before treating the post-treatment exposure coefficient as causal. When the cross-section is small, exact permutation or randomization logic can be more informative than another asymptotic standard error.

11.2 Policy evaluation with few treated clusters

A second example is policy evaluation with few treated clusters. In development, education, health, public economics, and political economy, treatment often varies at the level of districts, schools, hospitals, municipalities, provinces, regions, or countries. Administrative microdata can contain thousands or millions of individual observations, but if treatment varies only across a handful of aggregate units, the relevant sample size for treatment inference is the number of clusters supporting the contrast, not the number of rows inside those clusters. The small-cluster diagnostic in this paper is therefore a special case of a broader problem: conventional cluster-robust inference can look precise even when the effective cluster count is too small for large-G approximations to be credible.

Several tools travel to that setting: exact-enumeration wild-cluster bootstrap to expose the discreteness of the reference distribution at small cluster counts (Cameron et al., 2008; MacKinnon and Webb, 2017, 2020), CR2 and CR3 corrections to reveal sensitivity to leverage and degrees-of-freedom adjustments (Pustejovsky and Tipton, 2018), randomization or permutation procedures where the assignment mechanism supports them (Conley and Taber, 2011), and leave-one-cluster-out diagnostics to show whether a result rests on a single influential treated or control cluster. None converts a weak design into a strong one; their purpose is to make the information content of the design visible before a conventional significance claim.

11.3 Event studies with small cross-sections

A third example is macro-finance event studies with small cross-sections of assets, banks, firms, or countries. High-frequency windows are attractive because they can isolate market reactions around news. Monetary-policy announcement studies, for example, use narrow windows to separate policy surprises from slower-moving macroeconomic confounders (Gürkaynak et al., 2005; Altavilla et al., 2019). But a narrow event window does not make a twelve-unit cross-section large. If the units share strong common factors, heteroskedasticity-robust p-values can be badly miscalibrated even when the event date is economically well defined.

The placebo-date diagnostic used in this paper travels directly to that setting. Draw or enumerate placebo dates from the same pre-event panel, re-estimate the event-study regression under the same window definition and inference procedure, and report the empirical placebo rejection rate and the placebo coefficient distribution. If nominal five-percent rejection corresponds to a much

larger placebo rejection rate, conventional HC-type p-values should be treated as descriptive rather than decisive. The diagnostic does not require the placebo dates to be economically identical to the event date. It asks a narrower operational question: how unusual is the event-day coefficient relative to coefficients generated by the same cross-section, exposure vector, window construction, and common-factor structure when no event of interest is present?

12 Conclusion

Empirical studies often works at sample sizes where standard asymptotic inference is a poor guide. Country panels may have only a handful of clusters, event-study cross-sections only a dozen units, continuous treatment intensity may align with pre-treatment outcome dynamics, and formula-based exposure measures may inherit those same dynamics. This paper develops four diagnostics for that environment and applies them to the NGEU sovereign-spread debate.

The results narrow the reduced-form evidence. In the daily event-study setting, HC1 inference is severely miscalibrated in a twelve-sovereign cross-section with strong common-factor dependence. Under the clean 2018–2019 placebo-date calibration, only the 18 May 2020 Franco-German proposal lies outside the calibrated 95% interval, and only at short symmetric event windows. Four-cluster auction specifications hit the exact wild-bootstrap p-value floor before they can support a 5% rejection. The monthly continuous-dose design loads on pre-existing spread convergence. A formula-based exposure design has a strong first stage but inherits the same pre-treatment pattern in reduced form.

The contribution is methodological rather than estimator-based. No individual tool used here is new; the contribution is to match existing tools to specific small-sample failure modes—exact-enumeration wild-cluster bootstrap, CR2 and CR3 companion corrections, leave-one-cluster-out diagnostics, permutation inference, placebo-date calibration, and formal pre-trend sensitivity analysis. The point is not to multiply robustness checks but to choose the one that targets the design's binding weakness.

The substantive claim about NGEU is deliberately limited. NGEU may well have affected sovereign risk premia through transfer, solidarity, and safe-asset channels. The question addressed here is whether the reduced-form designs used, or naturally suggested, in this literature can establish those effects credibly at the available sample sizes. In several prominent or natural specifications, the answer is no. The paper also does not solve the broader macro-identification problems created by coincident policies, endogenous treatment intensity, or spillovers across sovereigns. Those remain deeper design challenges.

The broader implication is simple: identify the effective sample size created by the research design, then report the finite-sample or pre-treatment diagnostic that matches the binding failure mode. In small-sample macro-finance and related policy settings, that discipline is a precondition for taking reduced-form estimates seriously.

During the preparation of this work the author(s) used ChatGPT and Claude code in order to to assist with language editing, organization of manuscript revisions, code review, and robustness-

check documentation. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

Rodrigo Adão, Michal Kolesár, and Eduardo Morales. Shift-share designs: Theory and inference. *Quarterly Journal of Economics*, 134(4):1949–2010, 2019.

Lucia Alessi, Amela Duranovic, Virmantas Kvedaras, and Irene Monasterolo. From risks to opportunities: the impact of green public investment programs on sovereign yields. *Research in International Business and Finance*, 2025.

Carlo Altavilla, Luca Brugnolini, Refet S. Gürkaynak, Roberto Motto, and Giuseppe Ragusa. Measuring euro area monetary policy. *Journal of Monetary Economics*, 108:162–179, 2019.

Susan Athey and Guido W. Imbens. The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

David H. Autor, David Dorn, and Gordon H. Hanson. The china syndrome: Local labor market effects of import competition in the United States. *American Economic Review*, 103(6):2121–2168, 2013.

Krzysztof Bańkowski, Marien Ferdinandusse, Sebastian Hauptmeier, Pascal Jacquinot, and Vilém Valenta. The macroeconomic impact of the Next Generation EU instrument on the euro area. ECB Occasional Paper 255, European Central Bank, 2021.

Krzysztof Bańkowski, Othman Bouabdallah, João Domingues Semeano, Ettore Dorrucci, Maximilian Freier, Pascal Jacquinot, Wolfgang Modery, Marta Rodriguez-Vives, Vilém Valenta, and Nico Zorell. The economic impact of Next Generation EU: a euro area perspective. ECB Occasional Paper 291, European Central Bank, 2022.

Krzysztof Bańkowski et al. Four years into the Next Generation EU programme: an updated preliminary evaluation of its economic impact. ECB Occasional Paper 362, European Central Bank, 2024.

Roel Beetsma, Massimo Giuliodori, Jesper Hanson, and Frank de Jong. Bid-to-cover and yield changes around public debt auctions in the euro area. *Journal of Banking & Finance*, 87:118–134, 2018.

Roel Beetsma, Massimo Giuliodori, Jesper Hanson, and Frank de Jong. Determinants of the bid-to-cover ratio in Eurozone sovereign debt auctions. *Journal of Empirical Finance*, 58:96–120, 2020.

Tilman Bletzinger, William Greif, and Bernd Schwaab. Can eu bonds serve as euro-denominated safe assets? *Journal of Risk and Financial Management*, 15(11):530, 2022.

Kirill Borusyak, Peter Hull, and Xavier Jaravel. Quasi-experimental shift-share research designs. *Review of Economic Studies*, 89(1):181–213, 2022.

Kirill Borusyak, Peter Hull, and Xavier Jaravel. A practical guide to shift-share instruments. *Journal of Economic Perspectives*, 39(1):181–204, 2025.

Brantly Callaway, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. Difference-in-differences with a continuous treatment. Working Paper 32117, National Bureau of Economic Research, 2024.

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

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.

Ivan A. Canay, Andres Santos, and Azeem M. Shaikh. The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics*, 103(2):346–363, 2021.

Rebecca Christie, Grégory Claeys, and Pauline Weil. Next Generation EU borrowing: a first assessment. Bruegel Policy Contribution 22/2021, Bruegel, 2021.

Timothy G. Conley and Christopher R. Taber. Inference with "difference in differences" with a small number of policy changes. *Review of Economics and Statistics*, 93(1):113–125, 2011.

Stefano Corradin and Bernd Schwaab. Euro area sovereign bond risk premia during the Covid-19 pandemic. *European Economic Review*, 153:104402, 2023.

Zsolt Darvas. Next Generation EU: 75% of grants will have to wait until 2023, 2020. Bruegel Blog Post, 10 June 2020.

Zsolt Darvas. The nonsense of Next Generation EU net balance calculations, 2021. Bruegel Blog Post.

Clément de Chaisemartin and Xavier D'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

European Parliament and Council. Regulation (EU) 2021/241 of the European Parliament and of the Council of 12 February 2021 establishing the Recovery and Resilience Facility, 2021. Official Journal of the European Union L57/17, including Annex II allocation formula.

Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: what, when, why, and how. *American Economic Review*, 110(8):2586–2624, 2020.

Refet S. Gürkaynak, Brian Sack, and Eric T. Swanson. Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 1(1):55–93, 2005.

Andreas Hagemann. Placebo inference on treatment effects when the number of clusters is small. *Journal of Econometrics*, 213(1):190–209, 2019.

Annika Havlik, Friedrich Heinemann, Samuel Helbig, and Justus Nover. Dispelling the shadow

of fiscal dominance? Fiscal and monetary announcement effects for euro area sovereign spreads in the corona pandemic. *Journal of International Money and Finance*, 122:102578, 2022.

James G. MacKinnon and Matthew D. Webb. Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254, 2017.

James G. MacKinnon and Matthew D. Webb. Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, 218(2):435–450, 2020.

James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985.

Stephen Millard. A macroeconomic analysis of the impact of the eu recovery and resilience facility. NIESR Discussion Paper 564, National Institute of Economic and Social Research, 2025.

Livia Pancotto, Owain ap Gwilym, and Philip Molyneux. Deal! Market reactions to the agreement on the European Union Covid-19 recovery fund. *Journal of Financial Stability*, page 101157, 2023.

James E. Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, 2018.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591, 2023.

Jonathan Roth. Pretest with caution: event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322, 2022.

Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? *Journal of Econometrics*, 235(2):2218–2244, 2023.

Heliodoro Temprano Arroyo. The EU's response to the COVID-19 crisis: a game changer for the international role of the euro? European Economy Discussion Paper 164, European Commission, 2022.

Matthew D. Webb. Reworking wild bootstrap based inference for clustered errors. Queen's Economics Department Working Paper 1315, Queen's University, Department of Economics, Nov 2014.

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

Alwyn Young. Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557–598, 2019.

A Dataset details and summary statistics

This appendix collects dataset provenance and summary statistics that are too detailed for §3.

Auction-level panel. The auction panel extends the sovereign-auction design of Beetsma et al. (2020). The underlying data were collected from national debt-management-office releases and extended through March 2023. Each observation is a sovereign auction indexed by country and date. The raw four-country panel contains 451 auctions for Belgium (70), France (198), Italy (113), and Portugal (70) between January 2018 and March 2023.

The main primary-market outcome is the bid-to-cover ratio, defined as total bids divided by competitive allotment. This is the standard auction-demand measure in the sovereign-auction literature (Beetsma et al., 2018, 2020). Controls follow the Beetsma-style design: allotment, secondary-market yields, maturity buckets, lagged own and foreign bid-to-cover ratios at the same maturity, volatility controls, central-bank purchase variables, and an NGEU issuance indicator. The competitive allotment is itself chosen by the issuer and may respond to market conditions; it is included with that caveat in mind.

The full-controls estimation sample used in the headline auction specifications is smaller than the raw panel because lagged controls, maturity-specific controls, volatility variables, and central-bank-purchase variables must be observed; this produces the 281-observation full-controls sample reported in Table 6. The expanded 10-year subset contains 431 auctions before imposing the reduced-control estimation restrictions and adds Austria (56), Finland (18), and Spain (76). The six-country reduced-controls specification in Table 6 uses 413 observations after the necessary merges and lags. Volatility controls are the one-day lagged five-day rolling standard deviation of daily yield changes for bond-market volatility, and the same rolling construction on the VSTOXX index for equity volatility. The central-bank flow and stock variables are the country's one-day lagged daily change in national central bank holdings of euro-area general government debt, and the one-day lagged absolute stock in billions of euro. The four-cluster problem studied in Diagnostic 2 is therefore a consequence of data availability, not an arbitrary restriction.

Monthly country panel. The monthly country panel uses Eurostat `IRT_LT_MCBY_M`, which reports the harmonised Maastricht convergence long-term interest rate. It covers eleven sovereigns over 84 months from January 2018 to December 2024: eight euro-area countries (Austria, Belgium, Germany, Spain, Finland, France, Italy, and Portugal) and three non-euro comparison units (Denmark, Sweden, and the United Kingdom). The full panel has 924 country-month observations. Germany is the reference country for the spread calculation, leaving ten non-German spread series and 840 observations on the outcome variable $\text{spread}_{c,t} = 100 \cdot (y_{c,t} - y_{\text{DE},t})$.

The monthly panel is used for the continuous-dose and formula-exposure diagnostics. Monthly frequency is sufficient for those diagnostics because the questions are whether a panel continuous-dose design, or a formula-based exposure design, can identify NGEU-related effects on sovereign spreads. The three non-euro sovereigns take zero RRF allocation intensity by construction and

are insulated from ECB sovereign-bond purchases because PEPP and PSPP operate only in euro-area secondary markets. They are not clean counterfactuals for euro-area countries: currency regime, monetary-policy framework, investor base, and redenomination-risk exposure differ. Differential post-2020 movements in euro-area spreads relative to Denmark, Sweden, and the United Kingdom therefore reflect the euro-area exposure contrast together with other euro-area-specific post-2020 forces, including PEPP. For that reason, the monthly dose-post coefficient should be interpreted as NGEU-plus-PEPP relative to no euro-area programme, not as NGEU in isolation. The pre-trend failure documented in Section 7 arises before that interpretation problem becomes decisive.

FactSet daily panel. The daily panel is built from vendor benchmark-yield identifiers for 10-year sovereign bonds downloaded from FactSet Workstation in April 2026. It contains eight euro-area countries (Austria, Belgium, Finland, France, Germany, Italy, Portugal, and Spain) and five non-euro countries (Denmark, Norway, Sweden, Switzerland, and the United Kingdom). The raw series run from 14 April 2006 to 15 April 2026; the working sample runs from January 2018 through December 2024.

As in the monthly panel, Germany is the reference country. This leaves twelve non-German daily spread series. The five non-euro sovereigns serve as PEPP-insulated comparison units because ECB purchase programmes do not operate in those secondary markets. Their role is to widen the cross-section for the small-sample event-study diagnostic, not to provide clean counterfactuals for treated euro-area sovereigns. Norway and Switzerland appear in the daily panel but not in the monthly panel because FactSet daily coverage is broader than the Eurostat monthly panel. Missing values are forward-filled up to three trading days; such fills account for less than 0.3% of observations and do not affect the diagnostics. The replication README reports the vendor identifiers needed by licensed users to regenerate the raw yield panel.

RRF source data. Grant and loan ceilings are from Annex I of Regulation 2021/241 (European Parliament and Council, 2021). 2019 nominal GDP is from Eurostat `nama_10_gdp`. 2019 GDP per capita is from Eurostat `nama_10_pc`. Average 2015–2019 unemployment is the simple mean of the Eurostat harmonised unemployment rate `une_rt_a` over the five calendar years. Population shares are from Eurostat `demo_pjan` on 1 January 2019. The formula-based exposure reconstruction follows the Annex II weights documented in Darvas (2020) and Darvas (2021).

B Tables

Table 11: Mapping of econometric problems to diagnostics

Empirical problem	Econometric concern	Diagnostic used here	NGEU application
Small event-study cross-section	HC-type miscalibration under cross-sectional dependence	Placebo-date calibration, permutation checks	Daily spread events
Few country clusters	Cluster-robust asymptotics fail; WCB discreteness	Exact WCB, CR2/CR3, LOCO	Auction panel
Continuous treatment intensity	Treatment intensity correlated with pre-trends	Binned event study, first differences, Rambachan–Roth	Monthly panel
Formula-based exposure	Exposure predicts pre-treatment dynamics	Reduced-form pre-bins, balance tests, exact 10! permutation	RRF formula exposure

Table 12: Country-level RRF allocation-intensity measures

Country	Grants (EUR bn)	Loans (EUR bn)	2019 GDP (EUR bn)	Baseline intensity	Grants-only intensity	Formula-based intensity
Austria	3.5	0.0	397.6	0.009	0.009	0.0019
Belgium	5.0	0.0	480.0	0.010	0.010	0.0042
Finland	2.1	0.0	240.6	0.009	0.009	0.0046
France	39.4	0.0	2437.6	0.016	0.016	0.0111
Germany	28.0	0.0	3473.3	0.008	0.008	0.0010
Italy	68.9	122.6	1796.1	0.107	0.038	0.0287
Portugal	13.9	2.7	214.4	0.077	0.065	0.0497
Spain	77.2	84.0	1244.8	0.130	0.062	0.0967

Notes: The baseline measure equals grants plus loans divided by 2019 GDP; the grants-only measure equals grants divided by 2019 GDP. The formula-based measure reconstructs the pre-pandemic component of the RRF grant-allocation formula from 2019 GDP per capita, average unemployment over 2015–2019, and 2019 population. The table reports the euro-area exposure vector used in the dose and formula diagnostics, with Germany included for completeness even though it serves only as the spread reference and never enters as a treated unit. Non-euro sovereigns enter the monthly and daily diagnostics with zero exposure as a convention for the euro-area sovereign-spread channel; this is not a statement that Denmark and Sweden were RRF-ineligible.

C Exact-enumeration wild cluster bootstrap

This appendix gives pseudocode for the exact-enumeration wild cluster bootstrap of §6.1 and §10.2. The procedure follows Cameron et al. (2008) with the exact enumeration refinement of MacKinnon and Webb (2017). The input is a panel dataset $\{(y_i, x_i, z_i, g_i)\}_{i=1}^{N}$ with outcome y_i, coefficient-of-interest regressor x_i, other regressors z_i, and cluster indicator $g_i \in \{1, \ldots, G\}$. The output is the exact-enumeration two-sided wild cluster bootstrap p-value for the null $\beta = 0$ on the coefficient of interest.

1. Fit the unrestricted model by OLS and store the cluster-robust t-statistic on x, call it t^{obs}.
2. Fit the restricted model (with β set to zero) by OLS. Store the restricted residuals $\hat{\varepsilon}_i^r$ and the restricted fitted values $\hat{y}_i^r$.

3. Partial the coefficient-of-interest regressor out of the other regressors: regress x on z, store the residuals $\tilde{x}_i$.

4. For each of the 2^G Rademacher sign patterns $w \in \{-1, +1\}^G$:

   (a) Form the bootstrap outcome $y_i^* = \hat{y}_i^r + w_{g_i} \cdot \hat{\varepsilon}_i^r$.

   (b) Compute the bootstrap coefficient via the Frisch-Waugh-Lovell partial-out: $\hat{\beta}^* = (\tilde{x}'\tilde{x})^{-1}\tilde{x}'(y^* - P_z y^*)$, where P_z projects onto the z-columns.

   (c) Compute the cluster-robust t-statistic t_w^* using the same cluster-sandwich formula as step 1.

5. Return the two-sided p-value $p_{\text{WCR}} = \#\{w : |t_w^*| \geq |t^{\text{obs}}|\}/2^G$.

At $G = 12$ the loop has 4,096 iterations; at $G = 16$, 65,536; at $G = 20$, slightly over a million. On modern hardware, step 4 vectorises across all sign patterns and runs in milliseconds for $G \leq 16$ and in seconds for $G \leq 20$. The smallest two-sided p-value achievable is $1/2^{G-1}$. The mechanism is the one given in Lemma 6.1: each Rademacher sign pattern w yields the same $|t^*(w)|$ as its sign-reverse $-w$ by sign-anti-symmetry of the t-statistic, so the two-sided rejection set is closed under negation and the smallest nonempty rejection set contains exactly one pair $\{w_{\text{obs}}, -w_{\text{obs}}\}$, which has probability $2/2^G = 1/2^{G-1}$.

D Placebo-date calibration procedure

This appendix describes the placebo-date calibration procedure used in 5.3 and recommended in 10.2. The input is a balanced panel of daily sovereign spreads, a treatment-intensity vector, and a list of news dates. The output is a calibrated 95% reference interval for the cross-section $\hat{\beta}$ under the pre-treatment benchmark, together with the base-rate HC1 placebo rejection rate (PRR).

The procedure can be implemented either by drawing B placebo dates from a pre-treatment window or by enumerating all eligible placebo dates when the window is small enough. The implementation in this paper enumerates all 488 complete-cross-section trading days in the clean January 2018 through December 2019 window, with a ten-trading-day buffer at each end. Enumeration removes simulation noise. Random subsampling remains acceptable when enumeration is infeasible, for instance in larger pre-treatment windows.

1. Choose a pre-treatment window long enough to accommodate the widest symmetric event window on both sides. In 5.3 the window is January 2018 through December 2019 with a ten-trading-day buffer.

2. Form the placebo-date set $\mathcal{P}$. Either select all eligible trading days in the window for which a complete cross-section is observed at every $t^* \pm h$, or draw B dates at random from that set if enumeration is infeasible. The main calibration in 5.3 uses $|\mathcal{P}| = 488$ eligible dates.

3. For each placebo date $t^* \in \mathcal{P}$ and each symmetric event window $S \pm h$, with $h \in \{1, 2, 5\}$:

   (a) Compute $\Delta s_c(t^*, h)$ for every sovereign c as $100 \cdot [(y_{c,t^*+h} - y_{\text{DE},t^*+h}) - (y_{c,t^*-h} - y_{\text{DE},t^*-h})]$.

   (b) Regress $\Delta s_c(t^*, h)$ on the treatment-intensity vector using OLS with HC1 robust standard errors.

   (c) Store the estimated $\hat{\beta}_{t^*,h}$, the HC1 p-value $p^{\text{HC1}}_{t^*,h}$, and the indicator $\mathbf{1}\{p^{\text{HC1}}_{t^*,h} < 0.05\}$.

4. For each window $S \pm h$, compute the empirical 2.5% and 97.5% quantiles of the $\{\hat{\beta}_{t^*,h}\}_{t^* \in \mathcal{P}}$ distribution. These form the calibrated 95% confidence interval for β under the pre-treatment null.

5. For each window $S \pm h$, compute the base-rate HC1 placebo rejection rate $\#\{t^* : p^{\text{HC1}}_{t^*,h} < 0.05\}/|\mathcal{P}|$. Under correct size the base rate equals the nominal 5%. Departures from 5% diagnose miscalibration of the HC1 test in the sample at hand.

6. For each event-day specification, compute the observed $\hat{\beta}$ and compare it to the calibrated confidence interval. Reject the null only if the observed coefficient falls outside the calibrated interval.

The procedure is agnostic to the choice of treatment-intensity vector: it calibrates the *test*, not the *treatment*. The same placebo dates can be reused across alternative intensity measures by applying step 3(b) under each.

E Small-cluster inference: companion corrections

This appendix collects the CR2 and CR3 companion corrections and the leave-one-country-out (LOCO) sensitivity analysis for the auction-level specifications of Section 6. The main text reports the headline exact wild cluster bootstrap evidence and a short main-text summary; the full tables are reported here.

E.1 CR2/CR3 companion table

Table 13: CR1, CR2 (Satterthwaite-df), and CR3 (jackknife) cluster-robust inference alongside the exact wild cluster bootstrap

Specification	G	$\hat{\beta}$	SE_{CR1}	SE_{CR2}	SE_{CR3}	p_{CR1}	p_{CR2}	p_{CR3}	ν_{Satt}	p_{WCR}
4-country panel, spread outcome	4	−6.196	0.389	0.604	1.383	<0.0001	0.0028	0.0207	2.77	0.1250
6-country panel, spread outcome	6	−3.356	1.647	2.247	3.752	0.0424	0.2487	0.4121	2.52	0.0312
4-country panel, bid-to-cover outcome	4	0.017	0.245	0.469	1.317	0.9435	0.9728	0.9903	2.95	0.8750
4-country panel + linear trends, spread	4	−7.567	1.850	3.013	7.148	0.0001	0.1221	0.3675	2.11	0.2500

Notes: Standard errors reported to three significant figures; p-values to four decimal places. The CR3 p-value uses the conservative cluster df $G - 1$. Specifications $G \in \{4, 6\}$ correspond to rows A, B, C, D of Table 6. Source: auction-level panel; CR2 and CR3 implemented following Pustejovsky and Tipton (2018) and the `clubSandwich` R conventions; exact WCB via Rademacher enumeration; author's calculations.

Table 13 shows that CR2 and the exact wild cluster bootstrap point in the same broad direction

but do not deliver identical verdicts. That disagreement is itself informative. In the six-country specification, exact wild cluster bootstrap rejects at the five-percent level while CR2 does not. In the four-country headline spread specification, CR2 remains more favorable than exact enumeration. At the same time, the Satterthwaite degrees of freedom are extremely low, ranging from about two to three across the table.

Adding CR3 sharpens the picture in the direction that the small-cluster theory predicts. The CR3 jackknife-style adjustment inflates each cluster's residual contribution by the inverse leverage block $(I - H_{gg})^{-1}$ rather than its square root, so it is mechanically more conservative than CR2. In this application the gap is large rather than incremental: the CR3 standard error is between 1.7 and 2.8 times the CR2 standard error across the four specifications. For the four-country headline spread, $\mathrm{SE}_{\mathrm{CR3}} = 1.38$ against $\mathrm{SE}_{\mathrm{CR2}} = 0.60$, a factor of 2.3. For the six-country specification, the CR3 p-value is 0.41 where CR2 was 0.25 and CR1 was 0.04. The bid-to-cover null is robust to all three corrections, but every other specification moves substantially further from rejection as the correction grows more conservative. The qualitative reading is clear: the magnitude of the CR1→CR2→CR3 inflation is itself a diagnostic that the design sits in an extreme small-sample region, exactly where Pustejovsky and Tipton (2018) warn that alternative cluster-robust variance estimators can disagree by factors of two or more.

E.2 Leave-one-country-out sensitivity

A natural influence check is to drop one country at a time and re-estimate the surviving six-country specification. The sign of the coefficient is stable, but significance is not. Dropping Austria moves the asymptotic p-value above five percent, and dropping France does the same. By contrast, dropping Spain increases the magnitude of the coefficient sharply while preserving significance. The result is therefore not driven by sign reversal, but it is sensitive to which countries remain in the sample.

Table 14: Leave-one-country-out sensitivity for the headline six-country reduced-controls auction-level spread specification

Dropped country	$\hat{\beta}$	p_{asymp}	Sign flip?	5% sig. flip?	G	N
Full sample: $\hat{\beta}_{\mathrm{full}} = -3.356$, $p_{\mathrm{full}} = 0.0424$, $G = 6$, $N = 413$						
Austria	−3.212	0.0645	No	Yes	5	357
Belgium	−3.438	0.0443	No	No	5	370
Finland	−3.356	0.0424	No	No	6	413
France	−2.750	0.0935	No	Yes	5	292
Italy	−1.939	0.0113	No	No	5	342
Portugal	−3.267	0.0476	No	No	5	367
Spain	−6.409	<0.0001	No	No	5	337

Notes: The result is sign-stable across drops, but Austria and France are the influential clusters whose removal moves the asymptotic p-value above the five-percent threshold. Finland appears with the full-sample numbers because it is not in the six-country sample to begin with. Source: auction-level panel; OLS with cluster-robust standard errors on the indicated leave-one-out subsamples; author's calculations.

Table 14 confirms the verbal description above: Austria and France are the influential clusters for the five-percent significance margin, while Spain pulls the magnitude in the opposite direction

but does not threaten significance. The constructive implication is not to overstate what this panel can deliver, but to widen the sample whenever possible. With very small G, improving inference is not only about changing the correction method; it is also about changing the sample design.

F Monthly continuous-dose robustness

This appendix collects the four robustness exercises summarized in the main text of Section 7: first-difference estimates, alternative binning and reference-window choices, Rambachan–Roth sensitivity analysis, and leave-one-country-out sensitivity. All four reinforce the main conclusion that the monthly continuous-dose specification loads heavily on pre-existing peripheral convergence rather than on a clean post-2020 break.

F.1 First-difference estimates

If the level result reflects a genuine break around July 2020, it should continue to appear when the dependent variable is monthly spread changes. If instead the level result is absorbing a gradual pre-existing drift, first differencing should shrink the estimate sharply. That is what happens in the full sample.

Table 15: First-difference estimates of the continuous-dose specification

Sample	FE?	$\hat{\beta}$	SE	p_{asymp}	p_{WCR}	N	G
Full pool, 2018-01 to 2024-12	No	−11.668	10.322	0.2587	0.4121	830	10
Full pool, 2018-01 to 2024-12	Yes	−11.668	10.322	0.2587	0.4121	830	10
Drop UK	No	−5.397	8.339	0.5178	0.6758	747	9
Drop UK	Yes	−5.397	8.339	0.5178	0.6758	747	9
Drop COVID months (2020-02 to 2020-06)	No	−1.745	9.239	0.8503	0.8340	780	10
Drop COVID months (2020-02 to 2020-06)	Yes	−1.745	9.239	0.8503	0.8340	780	10
2019-07 to 2021-06 window only	No	−50.892	12.984	0.0001	0.0039	230	10
2019-07 to 2021-06 window only	Yes	−50.892	12.984	0.0001	0.0039	230	10

Notes: The outcome is the monthly change in the Bund spread (bps); the regressor is the change in dose-times-post. All specifications include month fixed effects and cluster standard errors by country. The "FE?" column indicates whether country fixed effects are added on top of the month fixed effects. For every sample the with-FE and without-FE rows are identical in coefficient and standard error. Source: Eurostat monthly sovereign yield panel (`IRT_LT_MCBY_M`)

The substantive content of Table 15 is the contrast between the full-sample and the short-window estimates. On the full sample, the first-difference coefficient collapses to a small and statistically indistinguishable number (−11.7 bps with a wild cluster bootstrap p-value of 0.412). Dropping the UK or the February–June 2020 COVID months has the same effect. Only the narrow 2019-07 to 2021-06 window delivers a large negative and significant estimate, and that short specification was chosen ex post to mute the broader pre-trend. The level estimate of −426.6 bps does not survive a generic first-difference robustness check on the same panel.

F.2 Alternative binning and reference-window sensitivity

Table 16 varies both bin width and the reference window. Eight of the nine configurations produce at least one pre-treatment bin significant at the five-percent level under exact wild cluster bootstrap. The remaining configuration is borderline rather than reassuring. The failure of the design is not an artifact of one particular choice of bin width or one particular reference period.

Table 16: Strongest pre-treatment bin under alternative binning and reference-window configurations

Config (bw × ref)	Strongest pre-bin	$\hat{\beta}$ (bps)	p_{WCR}
6×6	pre_3	596	0.004
6×12	pre_2	627	0.002
6×18	pre_1	477	0.002
12×6	pre_2	411	0.004
12×12 (baseline)	pre_1	544	0.002
12×18	pre_1	306	0.002
18×6	pre_1	322	0.055
18×12	pre_1	449	0.002
18×18	pre_1	306	0.002

Notes: Each row shows the pre-event bin with the smallest WCR p-value for that configuration. Eight of nine configurations have at least one pre-bin significant at $p_{\text{WCR}} < 0.05$; the lone exception (18×6) has $p_{\text{WCR}} = 0.055$. Source: Eurostat monthly sovereign yield panel (`IRT_LT_MCBY_M`).

F.3 Rambachan–Roth sensitivity analysis

The Rambachan and Roth (2023) exercise asks a stronger question than a conventional pre-test: does any bounded relaxation of parallel trends leave a statistically meaningful post-treatment effect standing? In this application, the answer is no. For the baseline 12-by-12 specification, the pre-treatment path is so irregular that the relevant constraint set is infeasible below a bound of roughly 800 basis points per unit intensity. At the first feasible value, all post-treatment confidence intervals include zero.

The breakdown $\bar{M}$ reported in Table 17 is the feasibility threshold: the smallest value of $\bar{M}$ at which the $\Delta^{RM}(\bar{M})$ constraint set becomes non-empty. The standard Rambachan and Roth (2023) breakdown definition is the smallest $\bar{M}$ at which the post-treatment confidence interval first includes zero under $\Delta^{RM}(\bar{M})$. In this application the two thresholds are very close: at the first feasible value of $\bar{M}$ every post-treatment confidence interval already includes zero, so the standard breakdown value lies in the same range $[550, 800]$ across the four post-treatment bins and never produces a binding signal of a non-zero post-treatment effect. The feasibility threshold is reported because it is what the `HonestDiD` implementation returns directly; the substantive conclusion is unchanged under either definition.

F.4 Leave-one-country-out sensitivity

Dropping Italy from the ten-country panel pushes the headline asymptotic p-value from 0.019 to 0.056, a significance flip at the 5% threshold. Dropping Spain increases the coefficient magnitude

Table 17: Rambachan and Roth (2023) sensitivity analysis for the baseline 12×12 binned event study

Post-treatment bin	$\hat{\beta}$ (bps)	Breakdown $\bar{M}$ (feasibility)	CI at $\bar{M} = 800$ includes zero?
0–6 months	−85.9	800	Yes
6–18 months	−334.8	800	Yes
18–30 months	42.7	800	Yes
≥30 months	−169.7	800	Yes

Notes: The reported breakdown $\bar{M}$ is the feasibility threshold (smallest value at which the $\Delta^{RM}(\bar{M})$ constraint set becomes non-empty). The standard breakdown definition (smallest $\bar{M}$ at which the CI first includes zero) lies in the same range across the four post-treatment bins, and at every value of $\bar{M}$ where the constraint is feasible, every post-treatment confidence interval already includes zero. There is no value of $\bar{M}$ at which a significant post-treatment effect survives.

by roughly two-thirds, from −427 to −708 bps, while retaining significance at the one-percent level. No country drop flips the sign. The pattern is that Spain drives the negative magnitude while Italy drives the marginal statistical significance.

Table 18: Leave-one-country-out sensitivity for the headline ten-cluster pooled monthly TWFE continuous-dose specification

Dropped country	$\hat{\beta}$	p_{asymp}	Sign flip?	5% sig. flip?	G	N
Full sample: $\hat{\beta}_{\text{full}} = -426.6$, $p_{\text{full}} = 0.0190$, $G = 10$, $N = 840$						
AT	−418.9	0.0226	No	No	9	756
BE	−427.7	0.0202	No	No	9	756
DK	−429.2	0.0227	No	No	9	756
ES	−708.2	<0.0001	No	No	9	756
FI	−419.0	0.0226	No	No	9	756
FR	−422.0	0.0208	No	No	9	756
IT	−305.8	0.0557	No	Yes	9	756
PT	−375.4	0.0438	No	No	9	756
SE	−459.5	0.0143	No	No	9	756
UK	−390.7	0.0296	No	No	9	756

Notes: The sign is stable across all drops; Italy is the only country whose removal flips significance at the five-percent threshold, while dropping Spain increases the magnitude by roughly two-thirds. Source: Eurostat monthly sovereign yield panel; OLS with cluster-robust standard errors on the indicated leave-one-out subsamples.

G Formula-based exposure: first stage, balance, and LOCO

This appendix reports the first-stage input table, the full balance regression, and the leave-one-country-out sensitivity for the formula-based exposure design of Section 8. The main text retains the first-stage figure, the reduced-form binned event-study coefficients, and the balance plot; the corresponding tables are given here.

Table 19: First-stage inputs and outputs for the formula-based RRF intensity measure

Country	2019 GDP pc (€)	2015–19 unemp. (%)	Formula intensity	Actual intensity
Spain	26,430	17.26	0.0967	0.130
Portugal	20,850	8.67	0.0497	0.077
Italy	29,600	11.23	0.0287	0.107
France	36,000	9.35	0.0111	0.016
Belgium	41,720	7.15	0.0042	0.010
Finland	43,620	8.00	0.0046	0.009
Austria	44,730	5.32	0.0019	0.009
Germany	42,000	3.55	0.0010	0.008

Notes: The formula intensity is the 70% pre-pandemic share of the RRF grant allocation reconstructed from 2019 GDP per capita, 2015–19 average unemployment, and 2019 population, divided by 2019 GDP. In the eight euro-area recipient cross-section, OLS of actual intensity on the formula intensity plus a constant gives $R^2 = 0.799$ and a homoskedastic first-stage $F = 23.8$.

G.1 First-stage inputs and outputs

G.2 Full balance regression

The balance regression cross-section contains ten cross-sectional units ($N = 10$: AT, BE, DK, ES, FI, FR, IT, PT, SE, UK; Germany is the Bund reference and drops out by construction). At this sample size, conventional asymptotic standard errors are themselves inferentially limited. The analysis therefore supplements the HC1 asymptotic p-values with an exact randomization test that enumerates all 10! permutations of the formula-intensity vector across the ten units, holding the pre-treatment outcome vectors fixed; the two-sided exact-permutation p-value is reported alongside the asymptotic p-value below.

Table 20: Balance regressions of pre-treatment spread outcomes on the formula-based exposure measure

Pre-treatment outcome	$\hat{\beta}$	SE	p_{asymp}	p_{exact}
Mean spread level (2018–19)	933	531	0.079	0.166
Mean monthly change (2018–19)	−22	11	0.046	0.100
Trend slope (bps/day, 2018–19)	−0.75	0.47	0.112	0.106

Notes: p_{asymp} is the HC1 asymptotic two-sided p-value; p_{exact} is the two-sided exact-permutation p-value enumerating all 10! reassignments of the formula intensity across the ten cross-sectional units. The asymptotically significant mean-monthly-change coefficient ($p_{\text{asymp}} = 0.046$) does not survive exact enumeration ($p_{\text{exact}} = 0.100$); no balance outcome is significant at the 5% level under the exact test. This is the substantive lesson: with $N = 10$, asymptotic standard errors materially understate sampling uncertainty, and the diagnostic correctly classifies the balance regression as inferentially limited.

The diagnostic conclusion stands at the prose level: the formula-based exposure measure correlates with pre-treatment spread dynamics, and the balance assumption that underlies a causal reading of the reduced-form coefficient is not credibly satisfied. The strength of the inferential support for this conclusion is, however, limited by the $N = 10$ cross-section: the asymptotic significance of the mean-monthly-change slope is not preserved under exact enumeration, so the balance evidence is suggestive rather than decisive on its own. The interpretation therefore leans on the panel reduced-form pre-bin coefficients in Section 8.2 (10-

country monthly panel, 840 observations) as the primary inferential basis for the diagnostic, and treats the balance test as a complementary cross-sectional check whose appropriate inference is the exact permutation column above.

G.3 Leave-one-country-out sensitivity

The full-sample reduced-form headline is not significant at the five-percent level under asymptotic inference, but dropping Spain produces a dramatic shift: the coefficient moves to −1,483 basis points and becomes highly significant. No other country drop has a comparable effect, and none changes the sign.

Table 21: Leave-one-country-out sensitivity for the formula-based reduced-form headline

Dropped country	$\hat{\beta}$	p_{asymp}	Sign flip?	5% sig. flip?	G	N
Full sample: $\hat{\beta}_{\text{full}} = -456.0$, $p_{\text{full}} = 0.1069$, $G = 10$, $N = 840$						
AT	−435.3	0.1247	No	No	9	756
BE	−449.9	0.1153	No	No	9	756
DK	−442.7	0.1216	No	No	9	756
ES	−1,483.4	0.0002	No	Yes	9	756
FI	−436.3	0.1231	No	No	9	756
FR	−445.9	0.1179	No	No	9	756
IT	−399.5	0.0624	No	No	9	756
PT	−344.1	0.0862	No	No	9	756
SE	−478.0	0.0999	No	No	9	756
UK	−389.9	0.1403	No	No	9	756

Notes: The result is sign-stable but highly sensitive to Spain: dropping Spain moves the asymptotic p-value from 0.107 to 0.0002 and triples the magnitude of the coefficient. No other country drop changes significance at the five-percent threshold. Source: Eurostat monthly sovereign yield panel; formula-based RRF intensity from Annex II of Regulation 2021/241.

H Diagnostic 1: dependence structure, HC robustness, permutation, and event-window detail

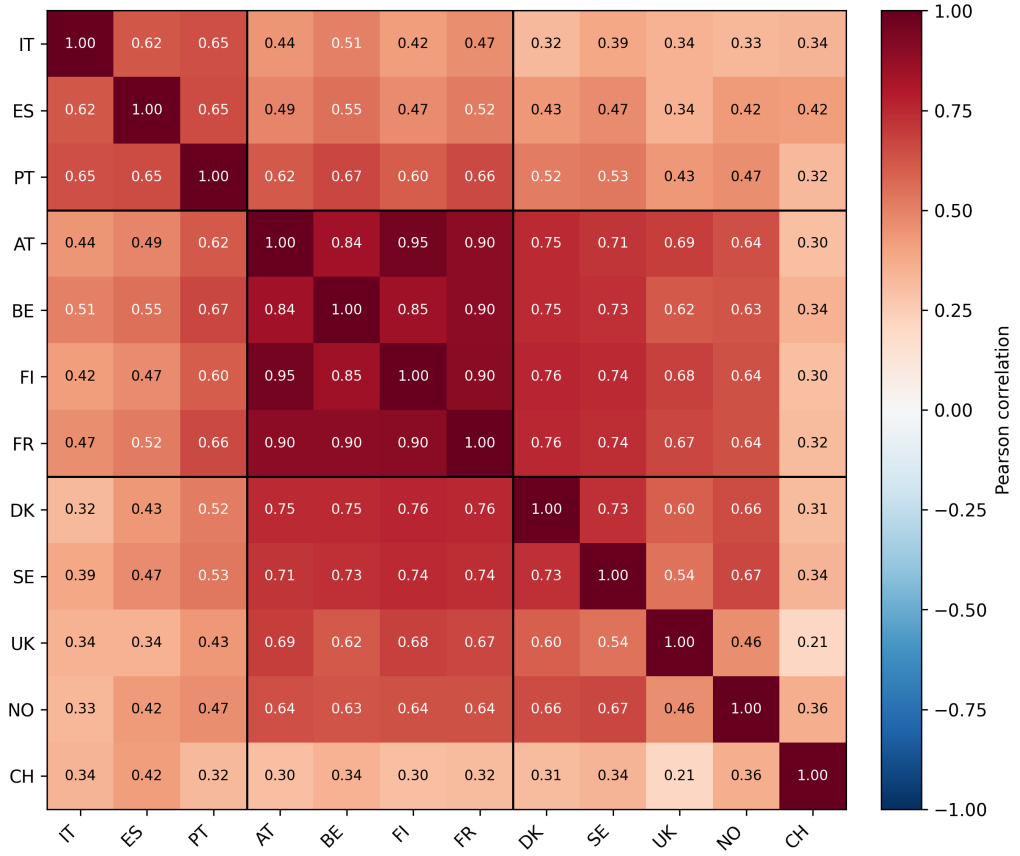
This appendix collects the extended robustness for Diagnostic 1: the full correlation heatmap and PCA loadings (Section H.1), the HC0–HC3 comparison (Section H.2), the exact permutation inference table (Section H.3), the placebo rejection-rate stability plot for the broader 627-day window (Section H.4), and event-window robustness across alternative window conventions (Section H.5).

H.1 Cross-sectional dependence: heatmap and PCA

A principal-component analysis on the standardized spread-change panel confirms the picture summarized by the group-correlation table in the main text. Table 22 reports the variance shares of the first three principal components together with the PC1 loadings.

PC1 alone explains 61.5 percent of the variance, with positive loadings on every sovereign, a

Figure 6: Pairwise correlations of daily Bund-spread changes $\Delta s_{c,t}$ across the twelve non-German sovereigns, 2018-01-01 to 2019-12-31, on the 474 trading days with no missing observations



Notes: Countries are ordered by group: peripheral euro, core euro, non-euro. Cells are annotated with the rounded correlation; thin black lines separate the three groups.

textbook "level factor" across European sovereign spreads. PC2 and PC3 together add another 17.7 percentage points. A small cross-section regression of Δs_c on a single treatment variable that is itself correlated with these factors, as the RRF allocation-intensity vector is because it loads heavily on the peripheral group, will inherit the variance of the common factors through the residual.

H.2 HC0–HC3 comparison

Table 23 repeats the placebo calibration on the broader 627-day 2018-01 to 2020-06 pre-Franco-German window using homoskedastic OLS standard errors and the four standard sandwich estimators HC0 through HC3.

The over-rejection is generic to the entire family of heteroskedasticity-robust corrections. HC3 is the most conservative but still rejects on roughly 22–27 percent of placebo days at the nominal 5% level, four to five times the correct size. Practitioners may prefer HC3 in cross-sections of this size, but should not treat HC3 p-values as decisive without first calibrating against a placebo distribution.

Table 22: Variance shares of the first three principal components of the daily Bund-spread-change panel (standardized columns, 474 trading days, 12 countries), and the PC1 loadings by country.

	PC1	PC2	PC3
Variance share	0.615	0.108	0.069
Cumulative share	0.615	0.723	0.792
PC1 loadings (peripheral | core | non-euro):			
IT	0.22		
ES	0.24		
PT	0.28		
AT	0.34		
BE	0.34		
FI	0.34		
FR	0.34		
DK	0.31		
SE	0.30		
UK	0.26		
NO	0.27		
CH	0.16		

Notes: PC1 alone explains 61.5% of the cross-sectional variance and loads positively on all twelve sovereigns.

H.3 Exact permutation inference

Table 24 reports a complementary inference exercise: a Fisher-style permutation test that holds the realized vector of cross-section spread changes fixed and permutes the RRF allocation-intensity assignment across countries. The number of distinct assignments equals $12!/\prod_j m_j!$, where m_j is the multiplicity of each tied intensity value. In the baseline twelve- country pool, five non-euro countries (DK, SE, UK, NO, CH) share zero intensity, and Austria and Finland share an additional rounded intensity of 0.009. The distinct-permutation space therefore contains $12!/(5!\,2!) = 1{,}995{,}840$ assignments. Because this is small enough, the calculation enumerates every distinct assignment exactly.

Placebo and permutation nulls answer different questions. The placebo calibration asks whether the realized event-day coefficient is unusual relative to ordinary pre-treatment trading-day coefficients generated by the same design, a question about *event-date extremeness* given the realized exposure vector. The permutation test asks instead whether, conditional on the realized vector of event-day spread changes, the observed assignment of treatment intensities is unusually aligned with that vector, a question about *cross-sectional label exchangeability* on a fixed date. The placebo test is the relevant inference object when the concern is common-factor day-to-day variation; the permutation test is informative about whether the realized exposure vector happens to project unusually onto the day's spread changes. The two tests answer different nulls.

The permutation test broadly agrees with the placebo calibration on the headline conclusion: the 18 May 2020 Franco-German proposal is the only event-window combination that produces a permutation p-value below 0.01 at every window. The two tests disagree on the 21 July 2020

Table 23: Placebo rejection rates by variance estimator

Window	SE type	N days	# rej. 5%	PRR @ 5%	# rej. 1%	PRR @ 1%
$S \pm 1$	Homoskedastic	627	270	0.431	158	0.252
$S \pm 1$	HC0	627	281	0.448	195	0.311
$S \pm 1$	HC1	627	255	0.407	168	0.268
$S \pm 1$	HC2	627	213	0.340	129	0.206
$S \pm 1$	HC3	627	136	0.217	78	0.124
$S \pm 2$	Homoskedastic	627	295	0.470	185	0.295
$S \pm 2$	HC0	627	295	0.470	205	0.327
$S \pm 2$	HC1	627	270	0.431	177	0.282
$S \pm 2$	HC2	627	220	0.351	134	0.214
$S \pm 2$	HC3	627	146	0.233	74	0.118
$S \pm 5$	Homoskedastic	627	322	0.514	204	0.325
$S \pm 5$	HC0	627	315	0.502	239	0.381
$S \pm 5$	HC1	627	290	0.463	207	0.330
$S \pm 5$	HC2	627	251	0.400	153	0.244
$S \pm 5$	HC3	627	169	0.270	89	0.142

Notes: The cross-section regression $\Delta s_c = \alpha + \beta\,\text{intensity}_c + \varepsilon_c$ is run on each of the 627 pre-NGEU trading days; the rejection rates are reported separately for homoskedastic OLS standard errors and for the HC0–HC3 family. Under correct size, the PRR should equal the nominal level; observed 5% PRRs range from 21.7% to 51.4% across the reported variance estimators and windows, while HC1 alone ranges from 40.7% to 46.3%.

Council deal at $S \pm 2$, where the exact permutation p-value is 0.0085 even though the calibrated placebo test classifies the event as "inside". The disagreement illustrates the conceptual distinction noted above: the Council-deal date has an unusual exposure alignment conditional on that day's spread changes, but it is not extreme relative to ordinary pre-NGEU common-factor movements.

H.4 Stability of the placebo rejection rate over the broader 2018–mid-2020 window

As draws accumulate from 100 up to the full 627 trading days, the cumulative PRR stabilizes by approximately 300 draws and remains in the 40–50 percent range across windows. The broader window therefore reinforces the qualitative conclusion that HC1 over-rejects in this design.

This broader-window exercise is a robustness check. The main calibration remains the clean 488-day 2018–2019 benchmark in Table 3, which excludes the early COVID episode and dates close to the NGEU policy sequence.

H.5 Event-window robustness

This subsection reports the cross-section HC1 regression of Δs_c on the RRF allocation-intensity vector under three alternative window conventions for each of the four NGEU news dates: a close-to-close one-day window ($t^* - 1 \to t^*$, denoted CC), a $t^* - 1 \to t^* + 1$ short symmetric window ($S \pm 1$), and a $t^* \to t^* + 1$ post-announcement one-day window (P1).

The headline 18 May 2020 result strengthens, rather than weakens, when the window is opened

Table 24: Exact permutation inference on the four NGEU news dates

Event	Win.	$\hat{\beta}$	p_{HC1}	Placebo verdict	p_{perm}	Mode
18 May 2020	$S \pm 1$	−180.5	0.000	outside	0.0010	exact (1,995,840)
	$S \pm 2$	−144.4	0.010	outside	0.0067	exact (1,995,840)
	$S \pm 5$	−177.2	0.006	inside	0.0054	exact (1,995,840)
21 Jul 2020	$S \pm 1$	−9.9	0.326	inside	0.4399	exact (1,995,840)
	$S \pm 2$	−81.6	0.032	inside	0.0085	exact (1,995,840)
	$S \pm 5$	−49.5	0.312	inside	0.1360	exact (1,995,840)
14 Dec 2020	$S \pm 1$	−29.4	0.008	inside	0.0624	exact (1,995,840)
	$S \pm 2$	−35.9	0.021	inside	0.0753	exact (1,995,840)
	$S \pm 5$	8.5	0.574	inside	0.6985	exact (1,995,840)
15 Jun 2021	$S \pm 1$	−4.1	0.209	inside	0.2308	exact (1,995,840)
	$S \pm 2$	15.9	0.035	inside	0.1781	exact (1,995,840)
	$S \pm 5$	4.9	0.637	inside	0.8040	exact (1,995,840)

Notes: For each event-window pair, the observed $\hat{\beta}$ and the HC1 p-value are reported alongside the two-sided permutation p-value, computed by enumerating every distinct permutation of the RRF allocation-intensity vector across the twelve sovereigns. The distinct-permutation space has $12!/(5!\,2!) = 1{,}995{,}840$ elements: five non-euro diagnostic units (DK, SE, UK, NO, CH) tied at zero exposure, and Austria and Finland tied at 0.009 exposure. All other intensities are unique. The placebo verdict column is copied from Table 5.

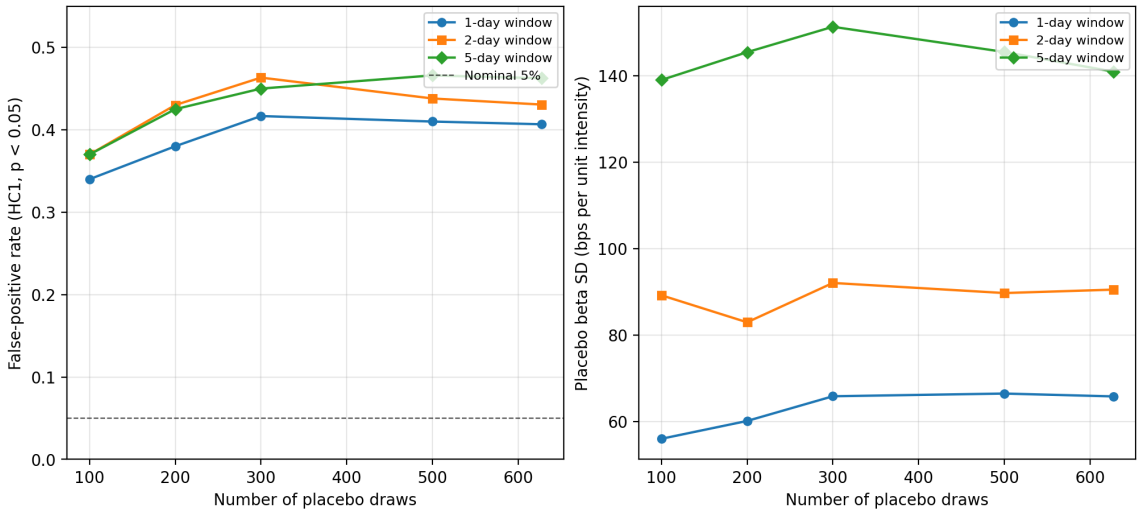
up. The close-to-close (CC) estimate is −47 bps with an HC1 p-value just above five percent (0.053); widening the window to $S \pm 1$ or P1 increases the magnitude to between −130 and −180 bps with HC1 p-values below 0.001. The other three NGEU dates remain economically small or sign-flip across windows. Only 18 May 2020 produces a treatment-correlated cross-section response that survives both influence checks and timing perturbations.

H.6 Factor-residualized and sample-composition robustness

Tables 26 and 27 add two robustness checks to the main Diagnostic 1 benchmark. The first removes common-factor loading components estimated from the 2018–2019 spread-change panel before rerunning the placebo exercise. Removing only the first principal component leaves the 18 May 2020 proposal outside the $S \pm 1$ and $S \pm 2$ placebo intervals. Removing the first two components absorbs most of the cross-sectional factor structure and moves the event inside the residualized placebo intervals. This is a diagnostic result: it confirms that common factors are central to HC1 miscalibration in this cross-section, not that the second component is a structural NGEU control.

The sample-composition check recomputes the placebo distribution rather than reusing the baseline interval. The $S \pm 1$ Franco-German result remains outside the placebo interval in all reported samples. The $S \pm 2$ result is weaker in the euro-area-only sample and remains outside in the baseline, EU-only, and drop-Denmark/Sweden samples. The $S \pm 5$ result is inside in all samples. The main conclusion is therefore stable at the shortest symmetric window and more sample-dependent at the wider symmetric window.

Figure 7: Stability of the HC1 placebo rejection rate over the broader 627-day pre-Franco-German window 2018-01 to 2020-06, as the cumulative number of placebo draws increases from 100 to 627



Notes: Lines show the cumulative PRR at each symmetric horizon. The rate stabilizes by approximately 300 draws. Source: FactSet daily 10-year benchmark sovereign yields; author's calculations. The main calibration in Table 3 uses only the clean 488-day 2018–2019 subset.

Table 25: Event-window robustness for the four NGEU news dates

Event	Window	$\hat{\beta}$ (bps)	$\mathrm{SE}_{\mathrm{HC1}}$	p_{HC1}	n
2020-05-18 Franco-German	CC ($t^*-1 \to t^*$)	−47.20	20.20	0.0533	12
	$S \pm 1$ ($t^*-1 \to t^*+1$)	−180.52	30.69	0.0002	12
	P1 ($t^* \to t^*+1$)	−133.32	19.10	<0.0001	12
2020-07-21 Council Deal	CC ($t^*-1 \to t^*$)	−0.29	5.93	0.9606	12
	$S \pm 1$ ($t^*-1 \to t^*+1$)	−9.85	10.03	0.3261	12
	P1 ($t^* \to t^*+1$)	−9.56	9.97	0.3377	12
2020-12-14 Reg. Adopted	CC ($t^*-1 \to t^*$)	−15.95	9.08	0.0789	12
	$S \pm 1$ ($t^*-1 \to t^*+1$)	−29.43	11.02	0.0076	12
	P1 ($t^* \to t^*+1$)	−13.48	4.75	0.0046	12
2021-06-15 First Issuance	CC ($t^*-1 \to t^*$)	−8.58	7.35	0.2431	12
	$S \pm 1$ ($t^*-1 \to t^*+1$)	−4.09	3.25	0.2085	12
	P1 ($t^* \to t^*+1$)	4.49	6.47	0.4875	12

Notes: Cross-section OLS of Δs_c on the RRF allocation-intensity vector with HC1 standard errors, 12-country pool. CC = close-to-close one-day window; $S \pm 1$ = short symmetric window from $t^* - 1$ to $t^* + 1$; P1 = post-announcement one-day window. HC1 p-values are uncalibrated; rejection thresholds should be read against the placebo-calibrated intervals of Section 5.3 rather than against nominal 5%.

Table 26: Factor-residualized Diagnostic 1 placebo benchmark

Adjustment	Window	PRR @ 5%	Mean $\hat{\beta}$	95% interval	18 May $\hat{\beta}$	Verdict
Remove PC1	S±1	33.6%	-2.8	[-76.9, 87.2]	-154.7	**outside**
Remove PC1	S±2	34.0%	-5.5	[-107.5, 114.0]	-128.4	**outside**
Remove PC1	S±5	38.3%	-9.8	[-165.3, 195.9]	-152.9	inside
Remove PC1-PC2	S±1	0.4%	-0.4	[-21.2, 19.2]	-14.5	inside
Remove PC1-PC2	S±2	0.0%	-0.8	[-25.3, 25.3]	-9.7	inside
Remove PC1-PC2	S±5	0.0%	-1.4	[-37.8, 38.0]	-29.6	inside

Notes: PCA loadings are estimated from standardized daily Bund-spread changes in 2018–2019 for the twelve non-German sovereigns. Each placebo and event-window spread-change vector is residualized on the first one or two loading vectors, then the residualized cross-section is regressed on RRF intensity with HC1 standard errors. Coefficients remain in basis points per unit exposure after removing the displayed common-factor loading components.

Table 27: Sample-composition robustness for the event-study placebo benchmark

Sample	Window	Units	PRR @ 5%	95% interval	18 May $\hat{\beta}$	Verdict
Baseline 12	S±1	12	35.9%	[-93.2, 112.5]	-180.5	**outside**
Baseline 12	S±2	12	39.1%	[-127.2, 129.2]	-144.4	**outside**
Baseline 12	S±5	12	43.2%	[-185.1, 243.4]	-177.2	inside
Euro-area 7	S±1	7	33.7%	[-98.2, 116.5]	-149.4	**outside**
Euro-area 7	S±2	7	35.7%	[-124.4, 148.3]	-111.2	inside
Euro-area 7	S±5	7	40.4%	[-173.4, 251.4]	-162.7	inside
Drop DK/SE	S±1	10	33.3%	[-98.8, 106.4]	-179.8	**outside**
Drop DK/SE	S±2	10	36.2%	[-131.3, 144.7]	-140.9	**outside**
Drop DK/SE	S±5	10	41.7%	[-181.9, 235.4]	-176.2	inside
EU only	S±1	9	35.1%	[-91.9, 119.4]	-161.9	**outside**
EU only	S±2	9	36.9%	[-121.5, 132.5]	-127.8	**outside**
EU only	S±5	9	42.5%	[-173.6, 253.6]	-169.4	inside

Notes: Each row recomputes the 2018–2019 placebo distribution using only the listed country sample. The euro-area sample contains AT, BE, FI, FR, IT, PT, and ES; Germany is the reference spread and is not in the cross-section. "Drop DK/SE" keeps the seven euro-area countries plus UK, NO, and CH. "EU only" keeps the seven euro-area countries plus Denmark and Sweden. Non-euro units assigned zero exposure are included only in the samples that list them.

I Software implementations of the diagnostic toolkit

Table 28 lists the main software implementations of each diagnostic across R, Stata, Python, and Julia.

Table 28: Software implementations of the diagnostic toolkit.

Diagnostic	R	Stata	Python	Julia	Notes
Wild cluster bootstrap	`fwildclusterboot`	`boottest`	`wildboottest`	`WildBootTests.jl`	Exact enumeration when G small
CR2 / CR3	`clubSandwich`	user-written	custom	custom	Report Satterthwaite df
Rambachan–Roth	`HonestDiD`	partial ports	custom	custom	Requires event-study coefs
Permutation inference	custom	`ritest` or custom	custom	custom	Account for ties
Placebo-date calibration	custom	custom	custom	custom	Straightforward loop

The main burden is therefore not computational. It is presentational. Researchers need to report these diagnostics clearly enough that the reader can see which inferential problem is relevant, which correction is being used, and whether the result survives the appropriate small-sample check.

J Country-level monthly trends

This appendix shows the per-country raw evidence underlying the binned event-study pre-trend result of 7.2. Figure 8 plots the monthly Bund spread, in basis points, for the ten non-German sovereigns in the monthly diagnostic panel, January 2018 through December 2024. Lines are colored by group: peripheral sovereigns (Italy, Spain, Portugal) in warm colors, core sovereigns (France, Belgium, Austria, Finland) in cool colors, and the three non-euro controls (Denmark, Sweden, the United Kingdom) in gray tones with dashed lines. The dotted vertical marker is the 21 July 2020 NGEU Council agreement.

The figure makes visible what the binned event-study coefficients of 7.2 report numerically. Countries with the highest RRF allocation intensity, Italy, Spain, and Portugal, are visibly converging toward Bund yields throughout 2018 and 2019, well before NGEU is on the negotiating table. Portugal's spread falls from above 130 bps to below 75 bps over the two pre-treatment years; Spain shows a similar though smaller drift; Italy's spread is more volatile but ends 2019 below where it started. The core sovereigns and the non-euro controls show essentially flat trajectories over the same window. Because the binned diagnostic and the formal HonestDiD analysis both load on these pre-treatment slopes, the figure clarifies why the monthly diagnostic verdict is what it is: the identifying conditional-parallel-trends assumption is violated for exactly the high-exposure sovereigns whose post-treatment trajectories the design relies on.

Figure 8: Monthly Bund spreads, ten non-German sovereigns, 2018-2024. Spreads are computed as $100 \cdot (y_{c,t} - y_{\mathrm{DE},t})$ in basis points. The vertical dotted line marks 21 July 2020 (NGEU Council agreement). Source: Eurostat monthly sovereign yield panel (`IRT_LT_MCBY_M`); author's calculations.