

Does routine supervisory communication move volatility?

Evidence from 285 ECB announcements

by

Andrea Monticini*

Università Cattolica del Sacro Cuore

Via Necchi 5

20123 Milan, Italy

email: andrea.monticini@unicatt.it

Abstract

Using all 285 ECB Banking Supervision releases from 2013 to 2025, we test whether routine supervisory communication moves daily bank index volatility. An EGARCH event model with randomization inference estimates a 2.4% increase ($p = 0.71$), with a pivotal 95% interval of $[-1.5\%, +6.1\%]$. Applied to 113 monetary policy decisions, the same design detects a 9.0% increase ($p = 0.023$). The evidence bounds the same-day aggregate effect of unanticipated supervisory news, not supervision's informativeness.

Keywords: Banking supervision, ECB, Single Supervisory Mechanism, EGARCH, randomization inference

JEL codes: G14, G21, G28

September 2026

* I am most grateful to Andrea Tarelli for useful comments.

1. Introduction

Prudential supervision is costly for banks and informative about them. Major institutional events have moved bank equity, including the establishment of European banking supervision and the 2014 Comprehensive Assessment: see Fiordelisi, Ricci and Stentella Lopes (2017), Loipersberger (2018) and Sahin and de Haan (2016). Stress test disclosures have also affected bank returns and volatility, as in Petrella and Resti (2013) and Durrani, Ongena and Ponte Marques (2022). These studies examine episodes selected for their importance. We ask instead whether the routine flow of guidance, consultations, sanctions and assessments is itself market-moving. If it is, the aggregate cost of supervisory communication exceeds the sum of the celebrated episodes; if it is not, the finding is consistent with effective supervisory gradualism.

This paper studies all 285 press releases published by ECB Banking Supervision between September 2013 and December 2025. The object of interest is the second moment of bank index returns. Monetary policy studies identify surprises from futures prices – Kuttner (2001), Bernanke and Kuttner (2005), and for the euro area Monticini, Peel and Vaciago (2011) – but no comparable instrument exists for qualitative supervisory news. Since its sign is unidentified, positive and negative reactions can cancel in mean returns, whereas a variance test does not require a predicted direction.

The estimand is deliberately narrow: the unanticipated, same-day, aggregate price effect of supervisory communication. It excludes anticipated content, intraday adjustment completed before the close, and bank-level effects concealed by aggregation.

The estimated change in conditional volatility is +2.4%, with a pivotal 95% interval of $[-1.5\%, +6.1\%]$. The estimate is robust across indices, event-day conventions and subsamples (Appendix B). As a positive control, the same design applied to the 113 monetary policy decisions of the period – an event class known to move bank equity, see Ricci (2015) – detects a 9.0% increase ($p = 0.023$).

2. Data

Announcements are the 285 ECB Banking Supervision press releases that remain once attachments and periodic statistical publications, identified by the ECB’s own publication type rather than by a reading of their content, are removed. Twenty-six administrative releases – appointments, relocations, nominations – are kept as a separate category. Releases before 16:00 CET are assigned to the same trading day and later ones to the next, yielding 242 substantive and 25 administrative event days.

Market data are daily from Bloomberg and comprise the EURO STOXX Banks and FTSE Italia All-Share Banks indices, their market benchmarks and the 10-year BTP–Bund spread. The euro area estimation contains 3,330 observations. The positive control uses the 113 monetary policy decisions of the Governing Council falling in that sample. Appendix A documents sample construction, publication times, overlapping events and alternative event-day conventions.

3. Empirical model and inference

Let r_t be the log return on the bank index and r_t^m the return on its market benchmark. The mean equation is

$$r_t = \mu + \rho r_{t-1} + \lambda r_t^m + \phi \Delta s_t + \varepsilon_t, \quad \varepsilon_t = h_t^{1/2} z_t, \quad (1)$$

where Δs_t is the change in the BTP–Bund spread and z_t is a standardized Student- t innovation. The contemporaneous market return isolates the bank-specific component. The conditional variance follows an EGARCH – Nelson (1991) – augmented with the event dummy:

$$\log h_t = \omega + \alpha z_{t-1} + \gamma (|z_{t-1}| - E|z_{t-1}|) + \beta_1 \log h_{t-1} + \beta_2 \log h_{t-2} + \delta D_t, \quad (2)$$

where D_t marks event days and δ is the parameter of interest. Since (2) is written in logarithms, δ is a proportional effect: volatility changes by $\exp(\delta/2) - 1$.

The lag structure, the error distribution and the model family were fixed by a deterministic diagnostic rule set in advance and reported in Appendix B, which returns EGARCH(1,2) with Student- t errors. No candidate fully removes residual ARCH, so coefficient magnitudes remain model-dependent. Because one step of the rule consults the observed event labels, the entire selection procedure is repeated within 1,000 placebo assignments: the randomization p value of the studentized event coefficient is then 0.58, confirming the fixed-specification conclusion.

Inference is by randomization. Placebo sets preserve the observed number of events within calendar year $\times$ weekday strata – monetary policy decisions are Thursdays in 96% of cases – and the true dates, overlaps included, remain in the support, while dates carrying only the other event class are excluded. The model is re-estimated on each assignment, and draws are screened on convergence alone, never on the log-likelihood. Tests are two-sided, obtained by doubling the smaller tail, because a release that resolves uncertainty rather than creating it lowers the conditional variance; pivotal intervals use the centred quantiles of the placebo distribution. Estimates carrying an interval use 10,000 draws. Appendix C gives the implementation and the quarter $\times$ weekday matching.

4. Results

The baseline estimate is $\hat{\delta} = 0.047$, a 2.4% increase in conditional volatility. It lies close to the centre of the placebo distribution ($p = 0.71$), and the upper bound of the pivotal 95% interval is 6.1%. Re-running the selection rule inside each assignment gives no evidence against the null either ($p = 0.58$). Quarter $\times$ weekday matching leaves the estimate unchanged and lowers the upper bound to 5.4% ($p = 0.80$).

The conclusion is unchanged when the contemporaneous market return is removed from the mean equation and when only system-wide releases are retained; both upper

Table 1: Effect of announcement days on the conditional variance of bank index returns

	$\hat{\delta}$	p_{RI}	$\Delta \text{ vol.}$	95% interval
<i>Panel A. Supervisory announcements</i>				
Baseline, euro area index	0.047	0.71	+2.4%	[−1.5%, +6.1%]
No market return in the mean	0.044	0.30	+2.2%	[−2.1%, +6.6%]
System-wide releases only	0.050	0.76	+2.6%	[−1.8%, +6.9%]
Quarter $\times$ weekday matched	0.047	0.80	+2.4%	[−0.7%, +5.4%]
<i>Panel B. Monetary policy decisions (positive control)</i>				
Baseline, euro area index	0.173	0.023	+9.0%	[+1.2%, +17.4%]
No market return in the mean	0.230	0.002	+12.2%	[+4.3%, +20.7%]
Quarter $\times$ weekday matched	0.173	0.13	+9.0%	[−0.9%, +20.3%]

Notes. $\hat{\delta}$ is the coefficient on the event dummy in the log-variance equation (2); $\Delta \text{ vol.}$ is $\exp(\hat{\delta}/2) - 1$. p_{RI} is the two-sided randomization p value, obtained by doubling the smaller tail of the placebo distribution and reported for the rows on which that distribution was computed, each using 10,000 draws. The 95% interval is $\hat{\delta}$ shifted by the centred quantiles of that distribution, a pivotal approximation rather than an exact inversion of the test. Placebo assignments preserve calendar year $\times$ weekday counts; the quarter-matched rows additionally preserve the calendar quarter, and are shown in both panels because they must be read together: the refinement leaves the supervisory estimate where it is but takes the positive control from 0.023 to 0.13, widening its placebo distribution and reducing its power. Sample: 3,330 observations, 2013–2025; 242 supervisory and 113 monetary policy event days. Further specifications, robust standard errors and asymptotic p values are reported in Appendix B.

bounds remain below 7%. Estimates are also small for the Italian index, under alternative event-day conventions and on the post-2018 subsample, for which publication times are known exactly (Appendix B). Mean returns on announcement days are indistinguishable from their placebo distribution ($p = 0.91$), and exactly half of the 242 days close negative (Appendix D).

The positive control gives a different answer. On monetary policy decision days conditional volatility rises by 9.0% ($p = 0.023$); without the market return in the mean equation the increase is 12.2% ($p = 0.002$). Quarter $\times$ weekday matching leaves that point estimate unchanged but widens its interval and reduces the power of the positive control ($p = 0.13$). [Figure 1](#) plots the two randomization distributions.

5. Concluding remarks

Routine ECB supervisory communication produces no detectable same-day aggregate volatility response, whereas monetary policy decisions estimated with the same design do. This may reflect anticipation generated by deliberately gradual supervisory procedures – consultations precede draft guidance and guidance precedes regulation – or

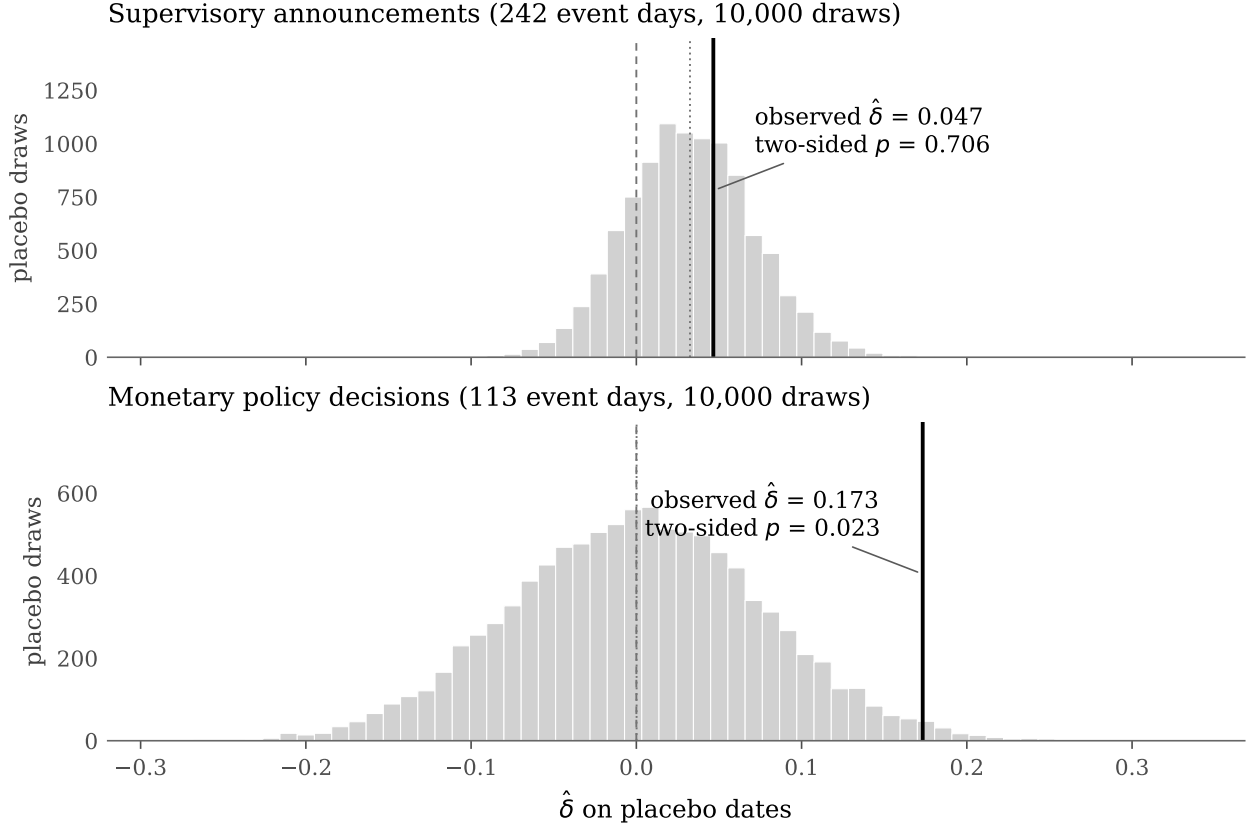


Figure 1: Randomization distributions of $\hat{\delta}$

The two panels show the distribution of $\hat{\delta}$ over sets of placebo dates matched on year and weekday, with the observed estimate marked by the solid vertical line and the two-sided p value of Table 1 alongside it. The dashed grey line is zero and the dotted grey line the mean of the placebo distribution; in the lower panel the two coincide. Both panels share the same horizontal scale.

the limited economic content of many releases; daily aggregate data cannot distinguish the two.

The result delimits rather than contradicts the evidence from exceptional institutional and stress test events. Three things bound it: aggregation across banks, the daily frequency, and a variance equation that no candidate fits. Because the same estimation procedure is applied to every placebo assignment, inference is less reliant on the misspecified model’s asymptotic approximation, although the magnitude remains model-dependent. Appendix E reports what intraday monetary policy surprises add.

References

Bernanke, B. S. and K. N. Kuttner (2005). “What Explains the Stock Market’s Reaction to Federal Reserve Policy?”, *Journal of Finance*, **60**, 1221–1257.

- Durrani, A., S. Ongena, and A. Ponte Marques (2022). “The Certification Role of the EU-wide Stress Testing Exercises in the Stock Market. What Can We Learn from the Stress Tests (2014–2021)?”, *ECB Working Paper* 2711.
- Fiordelisi, F., O. Ricci, and F. S. Stentella Lopes (2017). “The Unintended Consequences of the Launch of the Single Supervisory Mechanism in Europe”, *Journal of Financial and Quantitative Analysis*, **52**, 2809–2836.
- Kuttner, K. N. (2001). “Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market”, *Journal of Monetary Economics*, **47**, 523–544.
- Loipersberger, F. (2018). “The Effect of Supranational Banking Supervision on the Financial Sector: Event Study Evidence from Europe”, *Journal of Banking and Finance*, **91**, 34–48.
- Monticini, A., D. Peel, and G. Vaciago (2011). “The Impact of ECB and FED Announcements on the Euro Interest Rates”, *Economics Letters*, **113**, 139–142.
- Nelson, D. B. (1991). “Conditional Heteroskedasticity in Asset Returns: A New Approach”, *Econometrica*, **59**, 347–370.
- Petrella, G. and A. Resti (2013). “Supervisors as Information Producers: Do Stress Tests Reduce Bank Opacity?”, *Journal of Banking and Finance*, **37**, 5406–5420.
- Ricci, O. (2015). “The Impact of Monetary Policy Announcements on the Stock Price of Large European Banks during the Financial Crisis”, *Journal of Banking and Finance*, **52**, 245–255.
- Sahin, C. and J. de Haan (2016). “Market Reactions to the ECB’s Comprehensive Assessment”, *Economics Letters*, **140**, 1–5.

Online Appendix to
Does routine supervisory communication
move volatility?

Evidence from 285 ECB announcements

by

Andrea Monticini

Università Cattolica del Sacro Cuore

Via Necchi 5

20123 Milan, Italy

email: **andrea.monticini@unicatt.it**

September 2026

A. Construction of the announcement sample

Announcements were assembled twice and independently. The first extraction scraped the press release archive of ECB Banking Supervision page by page, taking each date from the annual index and from the file name. The second queried the publication database that the website loads to render its index, whose records carry a `pub_timestamp` field; both extractions are reproduced by the script accompanying this paper.

The two agree completely. Restricted to press releases published up to 19 December 2025 they return 297 and 295 documents, *agree on every one of the 294 dates they share*, and differ only over four attachments: three FAQ and annex documents that the scrape collects and the database does not list as publications, and one FAQ the database carries separately. Every one of the 285 releases in the sample appears in both. For the 203 releases whose publication time is recorded the two sources give *the same time to the minute, in all 203 cases*, and the remaining 82 are flagged as untimed by both. A third route, reading the date declared on each release page itself, agrees with the scrape on every shared date as well.

The database also settles two questions that would otherwise rest on judgement. The nine periodic statistical publications excluded from the sample are exactly the nine press releases that the ECB’s own classification assigns to the publication type reserved for supervisory banking statistics – the correspondence is exact in both directions, no release of that type is kept and no release of another type is dropped. The three attachments excluded are not listed in the database as publications at all. Neither exclusion required a judgement about content.

Publication times come from this database alone: the release pages carry a date but no time. Of the 82 releases treated as having no recorded time, 70 are stored at midnight and the remaining twelve between 00:10 and 02:00, which is why any stamp before 04:00 is read as a placeholder rather than as a publication time.

From 297 documents three attachments are dropped, since they duplicate the event date of a parent release, together with nine periodic statistical publications (“Supervisory banking statistics” releases), leaving 285.

Event days. The website records publication times for 203 of the 285 releases, essentially all of them from mid-2017 onwards. A release published before 16:00 CET is assigned to the same trading day, a later release to the next trading day, and a weekend release to the following Monday. Of the 203 timed releases two fall before 08:00, 113 between 08:00 and 12:00, 26 between 12:00 and 16:00, and 62 at or after 16:00, thirty-two of them at 18:00. Those 62 are all moved forward, but one falls on a Saturday and is counted under the weekend rule, so 61 carry the late-release rule. The 82 releases whose time is not recorded are assigned to the same trading day, which is an assumption and not a measurement: had they followed the timed distribution, roughly a quarter of them would belong to $t + 1$ instead. Two estimates in [Table B.3](#) bound the consequence: the specification that moves every event to $t + 1$, and the post-2018 subsample, in which publication times are known and no assumption is needed.

Twelve days carry more than one release, so the dummy is constructed per day rather than per release. The 285 releases map to 242 substantive and 25 administrative event days: the twenty-six administrative releases fall on twenty-six days, one of which (15 December 2016) also carries a substantive release and is therefore counted with the substantive days. The 16:00 cut-off is the one used throughout the main text; the robustness row labelled “cut-off at 17:30” re-assigns to the same trading day the ten releases published between 16:00 and 17:30, and changes the estimate from 0.0466 to 0.0467.

Administrative releases. Twenty-six releases concern appointments, office relocations and board nominations and carry no prudential content. They are kept as a separate category rather than pooled.

Scope. Releases are classified as institution-specific or system-wide from the titles alone, never from returns, by two independent criteria: verbs identifying action on a single intermediary (*sanctions, failing or likely to fail, asset quality review of, withdraws banking licence, temporary administrators*), and the presence of a bank’s proper name in the title. Fifty of the 259 substantive releases (19%) are institution-specific. They fall on 46 distinct dates, 45 of them trading days; four of those also carry a system-wide release and are assigned to the system-wide group, leaving the 41 institution-specific and 201 system-wide event days of the main text.

B. Selection of the variance specification

The lag structure, the error distribution and the model family were selected by a rule fixed in advance and applied in this order: among specifications that converge, that respond to the event regressor and that are numerically stable, take the one with the lowest Nyblom (1989) joint stability statistic among those not rejected by an ARCH LM test at 5%; if no candidate survives that test, take the lowest Nyblom among the stable ones. The rule never looks at the sign of $\hat{\delta}$, at its magnitude or at its p value. It is not, however, blind to the event dates: the screen described below establishes whether a family responds to a variance regressor at all, and to do so it estimates each family once with the true dummy. The main text is worded accordingly.

The second branch is the one that fires. No candidate passes the ARCH LM test, a fact recorded in the selection output rather than discovered afterwards, and the rule therefore selects on stability alone. This is stated here because it is the weakest link in the design: the variance equation is misspecified for every family examined, which is why the main text treats $\hat{\delta}$ as provisional in magnitude and relies on randomization rather than asymptotics for inference.

A prior screen. Each family is estimated twice, once with the true dummy and once with the dummy shifted forward by one day. The gap $\Delta \log L$ between the two fits separates the two ways a candidate fails. A gap in the hundreds or thousands means the optimiser lands in different regions depending on which dummy it is given, which matters because the randomization test refits the model a thousand times; the screen

Table B.1: Candidate specifications

	$\hat{\delta}$	AIC	ARCH LM	Nyblom	$\Delta \log L$	usable
eGARCH(2,1)- t	0.033	2.6494	0.029	4.20	8605.07	no
NAGARCH(1,1)- t	0.000	2.6520	0.005	5.10	0.00	no
eGARCH(1,2)-t	0.047	2.6522	0.003	3.66	0.29	yes
gjrGARCH(1,1)- t	0.000	2.6530	0.003	4.60	0.00	no
apARCH(1,1)- t	0.000	2.6531	0.002	4.43	0.00	no
sGARCH(1,1)- t	0.000	2.6535	0.002	3.96	0.00	no
TGARCH(1,1)- t	0.000	2.6543	0.001	4.09	0.00	no
AVGARCH(1,1)- t	0.012	2.6544	0.000	3.67	0.27	yes
eGARCH(1,1)-skew t	0.018	2.6853	0.001	17.84	14017.35	no
eGARCH(1,1)- t	0.164	2.6935	0.000	15.93	528.16	no

Notes. All ten candidates converge. ARCH LM is the p value of the test at five lags; none exceeds 0.05. Nyblom is the joint parameter-stability statistic, the criterion on which the rule selects: lower is better, and for the selected eleven-parameter model the asymptotic critical values are 2.49, 2.75 and 3.27 at the 10%, 5% and 1% levels. $\Delta \log L$ is the gap in log-likelihood between the fit with the true dummy and the fit with the dummy shifted by one day; a gap of zero means the regressor does not enter the fit, a large gap means the optimum is not stable. “Usable” is true when $\Delta \log L < 50$ and $\hat{\delta}$ is not zero to four decimals; the second condition is applied here rather than by the estimation code, as explained above. Reported $\hat{\delta}$ is rounded to three decimals; the five entries shown as 0.000 are zero to four decimals and pair with $\Delta \log L$ of zero. The selected specification is in bold.

discards those, and they are the three specifications marked unstable in Table B.1. A gap of exactly zero means the opposite: the fit does not respond to the dummy at all. This happens in five of the six non-logarithmic families, where the optimum for the coefficient on a variance regressor lies on the boundary that the positivity of the conditional variance imposes. A boundary optimum is not a failure of identification; it means that under those constraints the maximum is at the edge of the parameter space, where symmetric inference does not apply, and that the family cannot deliver an unconstrained estimate of δ .

One point of implementation should be stated plainly, since it is visible in the replication files. The screen tests for a non-zero coefficient as $|\hat{\delta}| > 10^{-8}$, and the five degenerate families return values that are zero to four decimals but not below that threshold, so they passed the test and entered the pool from which the Nyblom criterion selected. They are excluded here on the stronger evidence of $\Delta \log L = 0$, which is decisive and which the threshold should have encoded. The selection is unaffected: eGARCH(1,2)- t has the lowest Nyblom statistic among all stable candidates, degenerate ones included.

The screen repeated on a placebo. Because the screen uses the true event dates, it was repeated with a placebo dummy of the same size drawn from event-free days. Restricted to the branch that fires on the true data – lowest Nyblom among the usable

candidates – the placebo screen selects eGARCH(1,2)- t again, with a Nyblom statistic of 3.77 against 15.34 for the only other usable family. Applied in full, however, the rule returns eGARCH(2,1)- t : under the placebo dummy that family passes the ARCH LM test ($p = 0.14$), the first branch of the rule therefore fires, and it selects on ARCH LM rather than on stability. Two things follow. The criterion actually used here is invariant to the event dates. The other branch is not robust, since whether it fires at all turns on a single test crossing 0.05, and a rule for a future paper should not leave that hinge in place.

What the screen cannot do is manufacture the null. It keeps the families whose fit *responds* to the true dummy and discards those whose optimum ignores it, so the step that looks at the event dates selects towards a non-zero $\hat{\delta}$, not towards zero. The main estimate is indistinguishable from its placebo distribution all the same. Contamination from selection therefore works against the reported result rather than for it, and the same holds for the alternative usable candidate: AVGARCH(1,1)- t would have returned 0.012 in place of 0.047.

The rule re-run inside every draw. The argument above bounds the problem; it does not remove it. A fully valid randomization test re-runs the selection under every label assignment, so that the observed statistic enjoys no conditioning the placebo statistics lack. That was done. For each of 1,000 matched placebo assignments the entire rule is applied afresh – all ten families, each fitted with the placebo dummy and with it shifted by one day, Nyblom and ARCH LM recomputed – and the coefficient is taken from whichever specification the rule selects *for those labels*. Twenty-one fits per draw, five and a half hours on seven cores.

The statistic is the robust t of the event coefficient, $\hat{\delta}$ over its standard error, and not $\hat{\delta}$ itself. The candidates are not all logarithmic: in eGARCH δ is a proportional effect, while in AVGARCH the dummy enters the conditional standard deviation, so a null distribution of $\hat{\delta}$ would mix parameters that are not the same quantity. Two draws illustrate the point, returning $\hat{\delta}$ of 1.66 and -32.2 with standard errors of 10.6 and 67.1; as t they are 0.16 and -0.48 , which is what they are worth. This is therefore a check on the validity of the design and not a restatement of [Table 1](#) of the main text, whose statistic is different.

The rule reselects eGARCH(1,2)- t in 67.8% of draws. It picks eGARCH(1,1)- t in 16.1%, AVGARCH(1,1)- t in 10.8%, eGARCH(2,1)- t in 4.7% and eGARCH(1,1)-skewt in 0.6%: the specification does depend on the labels, but in two draws out of three it does not change. The observed t is 1.093 against a null distribution with mean 0.701 and standard deviation 0.773, so the estimate lies half a standard deviation above the centre and 290 draws out of 1,000 exceed it. The two-sided p value is 0.58, with a Monte Carlo error of 0.03. As in the fixed-specification test the null is not centred on zero: the procedure returns a positive t on any matched set of 242 days, which is the same property recorded for $\hat{\delta}$ above.

What this measures is the size of the contamination, and it is worth stating plainly rather than as a vindication. Holding the specification fixed the same 1,000 draws give $p = 0.69$; re-selecting inside each draw gives 0.58. The conditioning on the true dates is

therefore worth about a tenth of a p value: the objection is real and now quantified, and the conclusion does not turn on it, since 0.58 is as far from any conventional threshold as 0.69. The exercise covers the selection of the variance specification. The set of ten candidates and the mean equation were not chosen by a rule, and no re-randomization can repair that.

Three features of [Table B.1](#) bear on the main text. The information criteria are reported as `rugarch` computes them, divided by the number of observations; in ordinary units they span 147 points across the ten candidates and 17 across the best eight, so they do discriminate. They do not, however, select: AIC ranks eGARCH(2,1)- t first and the Schwarz criterion sGARCH(1,1)- t , and both are excluded before any criterion is consulted, the first as numerically unstable and the second because it does not respond to the variance regressor. Between the two usable candidates the two criteria agree, and by the same margin since both have eleven parameters, favouring eGARCH(1,2)- t over AVGARCH(1,1)- t by 7.3 points. Of the two usable candidates, eGARCH(1,2)- t has the lower Nyblom statistic, 3.66 against 3.67 – a margin so thin that the alternative deserves stating: AVGARCH(1,1)- t would have given $\hat{\delta} = 0.012$ instead of 0.047, that is a smaller estimated effect, so the selection rule did not pick the specification most favourable to a null. And no candidate passes the ARCH LM test, which establishes only that none of the families examined describes the residual dynamics adequately – not that a family outside this set would fail as well.

The mean equation. The selection rule above concerns the variance equation. The mean equation of the main text was not selected by a rule, and one of its terms deserves scrutiny of its own: the contemporaneous market return r_t^m . If a supervisory release moved the entire market, including r_t^m would remove that movement from ε_t before the conditional variance is estimated, and $\hat{\delta}$ would be biased towards zero by construction. [Table B.2](#) estimates the same variance equation under five mean equations, with the supervisory dummy and with the monetary policy dummy.

Three things follow. The market return is worth about 1,437 log-likelihood points, so the rows that drop it are deliberately misspecified checks and not competing specifications; they are reported because the objection they answer is about bias, not about fit. Dropping it does not raise the supervisory estimate, which is what a market-wide supervisory effect concealed by the control would require: the estimate moves from 0.047 to 0.044, and the randomization interval does not widen, so the check costs no precision either. And the same removal raises the monetary policy estimate from 0.173 to 0.230, and to 0.268 once Δs_t goes as well. Dropping the spread alone – which the market return leaves in place, and which is itself a possible outcome of the news – changes neither series much, 0.035 and 0.174. A common component on event days is therefore present and measurable in these data; it is in the monetary policy decisions. The autoregressive term changes nothing anywhere, 0.047 against 0.048.

Robustness of the event coefficient. [Table B.3](#) collects the specifications summarised in the main text. Every row re-estimates the variance equation of the main text on the stated sample or dummy, holding the selection rule fixed.

Table B.2: The event coefficient under alternative mean equations

mean equation	supervisory		monetary policy		log L
	$\hat{\delta}$	s.e.	$\hat{\delta}$	s.e.	
$\mu + \rho r_{t-1} + \lambda r_t^m + \phi \Delta s_t$ (baseline)	0.047	0.043	0.173	0.122	−4405.0
without λr_t^m	0.044	0.040	0.230	0.101	−5842.1
without $\phi \Delta s_t$	0.035	0.043	0.174	0.122	−4408.0
without λr_t^m and $\phi \Delta s_t$	0.015	0.043	0.268	0.107	−6119.6
without ρr_{t-1}	0.048	0.044	0.179	0.122	−4407.9
constant only	0.015	0.044	0.273	0.107	−6120.1

Notes. The variance equation is the one of the main text throughout, estimated on the euro area index over the same 3,330 observations; only the mean equation changes. Standard errors are Bollerslev–Wooldridge robust. log L is that of the supervisory fit; the monetary policy fits differ from it by less than three points in every row. The two rows without r_t^m were also subjected to the full randomization test of Appendix C: the supervisory estimate gives, on 10,000 draws, a two-sided p value of 0.30 with a 95% interval of $[-2.1\%, +6.6\%]$ in volatility, and the monetary policy estimate a p value of 0.002 with $[+4.3\%, +20.7\%]$, as reported in Table 1 of the main text.

C. The randomization distribution and the estimator

Matching. Placebo dates are matched on calendar year *and* weekday: each placebo set contains, for every year and every day of the week, as many dates as the true event set does. Year alone is not enough. Monetary policy decisions are announced on a published calendar and 96.5% of them fall on a Thursday, while supervisory releases lean towards Monday (26.9% against 20.0% of event-free days) and away from Thursday (14.0%). Whether this matters is an empirical question and the answer is that it does not: on days free of known events the variance of market-adjusted returns is 1.040 on Thursdays against 1.012 overall, a log ratio of 0.033 with a randomization p value of 0.71. A calendar regularity of that size cannot generate a $\hat{\delta}$ of 0.17. The matching is retained because it costs nothing, not because it rescues anything.

The support. The true event dates are eligible as placebo dates. The alternative, which the first implementation used, is to draw only from days free of every known event; the true dates are then never drawn and sit in the control group of all 1,000 re-estimations. If event days carry higher variance this depresses the placebo estimates and inflates the contrast with the observed one, and the p value is too small – which is what happened: the monetary policy control gave 0.008 under that design and gives 0.023 under this one. For the supervisory test the same bias works against the null, which survives it. Days carrying *only* the other event class are excluded from the support. Overlapping dates remain eligible, because they belong to the target event set: two of the 242 supervisory days are also monetary policy days and are drawn from, while the 111 monetary policy days that carry no release are not.

Table B.3: The event coefficient across specifications

	$\hat{\delta}$	s.e.	p_{as}	Δ vol.
<i>Panel A. Supervisory announcements</i>				
Baseline, euro area index	0.047	0.043	0.27	+2.4%
Italian bank index	0.016	0.054	0.76	+0.8%
Cut-off at 17:30	0.047	0.043	0.27	+2.4%
Event day $t + 1$	0.034	0.044	0.44	+1.7%
Day before the release	0.036	0.046	0.43	+1.8%
Normal errors	0.040	0.050	0.42	+2.0%
Post-2018 subsample	0.017	0.054	0.75	+0.8%
From November 2014	0.039	0.047	0.40	+2.0%
No market return in the mean	0.044	0.040	0.27	+2.2%
System-wide releases only	0.050	0.049	0.30	+2.6%
Institution-specific only	0.058	0.097	0.55	+2.9%
Administrative releases (placebo)	-0.126	0.204	0.54	-6.1%
<i>Panel B. Monetary policy decisions</i>				
Baseline, euro area index	0.173	0.122	0.16	+9.0%
No market return in the mean	0.230	0.101	0.02	+12.2%
Both dummies: monetary policy	0.193	0.122	0.11	+10.1%
Both dummies: supervision	0.057	0.044	0.19	+2.9%

Notes. Standard errors are Bollerslev–Wooldridge robust, see Bollerslev and Wooldridge (1992), and p_{as} is the associated asymptotic p value, reported here rather than in the main text because it disagrees systematically with the randomization p value: the baseline and the 17:30 cut-off have the same coefficient and the same standard error and give 0.71 against 0.27, which is why the paper relies on p_{RI} . The rows carrying a randomization p value and an interval are in Table 1 of the main text. About a fifth of the sample concerns individual institutions, which cannot move an aggregate index except through contagion, so pooling might dilute any effect mechanically; it does not detectably, 201 system-wide event days giving 0.050 against 0.058 on 41 institution-specific days, though 41 days is imprecise enough that only a large dilution would show.

Why not match more finely. Adding the calendar quarter to the strata is the obvious refinement, and it was run on both event classes. It does not move the supervisory estimate: $\hat{\delta}$ is unchanged at 0.047 and its p value goes from 0.71 to 0.80. But it substantially reduces the power of the positive control. On the monetary policy decisions the same refinement takes the p value from 0.023 to 0.13 and opens the interval to $[-0.9\%, +20.3\%]$, so an effect of 9% that the coarser design detects is no longer distinguishable from zero. Both are run on 10,000 draws, as are the central specifications.

The mechanism is capacity. Monetary policy decisions are almost all Thursdays; adding the quarter spreads 113 of them over 54 strata of nine to twelve eligible days each, from which two must be drawn. Each draw therefore takes a large share of a small pool and lands on the true dates often – an expected overlap of 18% – and those

draws return a large $\hat{\delta}$. The null distribution is contaminated towards the alternative: its standard deviation rises from 0.076 to 0.098 and its 95th percentile from 0.117 to 0.177, which is almost the observed 0.173. The supervisory design suffers the same effect more mildly, 242 events spread over 154 strata and an overlap rising from 10% to 16%.

The conclusion is not that the supervisory null survives a more demanding test. It is that this particular refinement buys stringency by reducing the number of admissible placebo assignments, and with it the power that the positive control measures. The matching of the main text was fixed before these results, on the calendar regularities set out above; what the quarter-matched rows add is a measure of what a finer design would cost.

Implementation. Every re-estimation starts from the coefficients of the baseline fit, without which a substantial share converges to degenerate optima: in a first implementation without starting values, 39% of draws returned $|\hat{\delta}| > 0.5$ against a standard error of 0.095, and 9% piled up on a single spurious optimum near $\hat{\delta} = -32$.

A draw is retained if the optimizer converges and its log-likelihood is not below that of the nested fit with $\delta = 0$, which adding a parameter cannot lower. No draw fails either condition in any test reported. Nothing else is screened, and in particular the log-likelihood is not used as a measure of the quality of a draw. It cannot be: across ten thousand placebo draws it correlates with $|\hat{\delta}|$ at 0.94, a placebo that falls on turbulent days raising the likelihood and producing a large coefficient by the same mechanism, so a likelihood screen is a screen on the statistic and would truncate the null distribution exactly in the tails where the p value is read. Mean $\hat{\delta}$ rises monotonically across deciles of the log-likelihood, from 0.000 to 0.098. A tolerance set relative to the largest log-likelihood in the batch is worse again, since the maximum of B draws grows with B : on the baseline the same rule discards nothing at a thousand draws and three quarters of them at ten thousand, and what it keeps is the upper tail of $|\hat{\delta}|$ – enough to move the p value from 0.71 to 0.06. None of the estimates reported here uses such a rule. Every row of Table 1 that reports an interval uses 10,000 draws rather than 1,000. Two considerations motivate the larger number. For the positive control the p value rested on twelve draws in the tail at a thousand; at ten thousand it rests on 116 and moves from 0.026 to 0.023. The same control without the market return rested on a single draw, which is a resolution floor and not a measurement: at ten thousand it rests on eight and the p value is 0.002. For the supervisory rows the p value was never in doubt – 0.71 is far from any threshold, and its Monte Carlo error at a thousand draws is about 0.03 – but the interval endpoints are quantiles estimated from some twenty-five draws per tail, with a Monte Carlo standard error of 0.16 percentage points on the upper endpoint. At ten thousand that falls to 0.05, which is what the reported decimal requires. Moving from one thousand draws to ten thousand changes the baseline interval from $[-1.45\%, +6.01\%]$ to $[-1.50\%, +6.09\%]$.

The p value. Reported p values are two-sided, obtained by doubling the smaller tail of the null distribution rather than by referring $|\hat{\delta}|$ to it: the null distribution is not centred on zero, so the absolute-value convention would compare the observed statistic

Table C.1: Monte Carlo distribution of $\hat{\delta}$ under $\delta = 0$

event days	mean	s.d.	t
52	-0.008	0.115	-1.65
117	0.003	0.077	0.92
246	-0.001	0.050	-0.45

with the wrong reference. Two-sidedness is substantive and not a formality, since a release that resolves uncertainty rather than creating it lowers the conditional variance; the administrative placebo of [Table B.3](#) is the one estimate with that sign.

The confidence interval. Writing d_b for the placebo estimates and $\bar{d}$ for their mean, the 95% interval reported for $\hat{\delta}$ is

$$\left[\hat{\delta} - q_{0.975}(d_b - \bar{d}), \hat{\delta} - q_{0.025}(d_b - \bar{d}) \right],$$

and is then mapped to volatility through $\exp(\cdot/2) - 1$. It is centred on the observed estimate by construction and inherits the width of the null distribution, which is the quantity the randomization design measures reliably. It should be called what it is: a pivotal approximation, which shifts one null distribution by $\hat{\delta}$ rather than recomputing the placebo distribution under each δ_0 on a grid. An exact inversion would require a thousand re-estimations per grid point and is not attempted here; the approximation is the standard one, and it is exact only if the null distribution of $\hat{\delta} - \delta_0$ does not depend on δ_0 .

No detectable bias in the estimator. The randomization distribution is not centred on zero: its mean is 0.032 in the supervisory design and 0.000 in the monetary policy design. Because a finite-sample bias could depend on event frequency and on the structure of the assignments, this was checked directly. Five hundred series were simulated from the fitted model with $\delta = 0$, holding the mean-equation regressors at their observed values, and a dummy on randomly chosen dates was estimated on each. The estimator is unbiased at every event frequency. The simulated dispersion matches the randomization distribution at the monetary policy frequency (0.077 against 0.072 at 117 and 113 event days) but exceeds it at the supervisory one (0.050 against 0.039 at 246 and 242 event days), the simulated series being drawn from a model that is itself misspecified for the data; the comparison bears on scale, not on the location that the table establishes. Where the null distribution sits on one realization of the data is therefore a property of that sample and carries no interpretation; the test does not require it to, comparing the observed statistic with the distribution of the same statistic on the same data under randomly assigned dates.

This is also why the asymptotic p values of [Table B.3](#) diverge from the randomization p values of [Table 1](#) of the main text. Under a variance equation that no candidate specification fits, asymptotic inference is unreliable in both location and scale, whereas the null distribution used here is generated under the same specification as the estimate compared with it.

A residual-based test without repeated variance-model estimation. Every estimate above depends on an optimizer, on starting values, and on a variance equation that Appendix B shows no candidate fits well. One test removes the first two. The model is estimated once with $\delta = 0$ – no dummy in the variance, log-likelihood -4405.576 on ten parameters – and its standardized residuals z_t are kept fixed ($\text{var}(z) = 0.9996$ on all 3,330 observations). Only the event labels are then permuted, over the same matched draws used elsewhere, and each statistic is the difference between its mean on labelled days and its mean on the rest. Nothing is re-optimized, so the p values below cannot be attributed to repeated optimization, although the standardized residuals remain dependent on the fitted variance model: they are divided by a conditional standard deviation that model produced, so the test is lighter than the others, not free of them.

Three statistics are reported. The mean of z_t^2 is the quantity the EGARCH dummy is meant to capture and is dominated by the largest residuals; the mean of $|z_t|$ is less sensitive to them; the mean rank of z_t^2 responds to a shift of the whole distribution. On 20,000 permutations the supervisory p values are 0.263, 0.054 and 0.012; the monetary policy ones are 0.012, 0.005 and 0.024. A fourth statistic, the mean of $\log z_t^2$, is reported by the accompanying script only to show why it must be discarded: six observations have $|z_t| < 0.001$ and the smallest $\log z_t^2$ is -21.4 , so a handful of days on which the model happened to fit almost exactly dominate the average. With three usable statistics on two event classes the multiplicity is real, and a single p of 0.01 in isolation would not be evidence.

What separates the two event classes is which part of the distribution moves. For supervisory days the ratio of the event to the non-event quantile of $|z_t|$ is 1.44 at the tenth percentile, 1.35 at the quarter, 1.15 at the median, 1.08 at the third quarter and 1.02 at the ninetieth: the distribution is compressed towards zero at the bottom and almost unchanged at the top. The share of days with $|z_t| < 0.25$ falls from 23.5% to 16.5% ($p = 0.006$) while the share with $|z_t| > 2$ is 5.0% against 5.1% ($p = 0.79$). On monetary policy days the two reverse: the lower-tail share moves by -4.6 points with $p = 0.41$, the upper-tail share by $+9.4$ points with $p = 0.026$. This is why the mean rank separates the supervisory days and the mean of z_t^2 does not, and it is consistent with the main estimates: a conditional variance whose likelihood is driven by its tails registers the monetary policy events and not the supervisory ones.

This evidence is exploratory. The pattern was found by inspecting the residuals and was not predicted beforehand, and it does not survive intact under finer matching: adding the calendar quarter to the year and weekday strata moves the mean-rank p from 0.012 to 0.072 and the lower-tail p from 0.006 to 0.035. It is reported because it is the one place where the two event classes differ in a way the variance equation cannot express, not as a finding the design was built to test.

Table D.1: Returns by event class, percentage points

	<i>n</i>	mean	median	% negative	<i>p</i>
<i>Panel A. Market-adjusted returns, bank index</i>					
Supervisory announcements	242	−0.003	0.018	50.0	0.91
system-wide	201	−0.015	−0.002	50.7	0.99
institution-specific	41	0.055	0.179	46.3	0.76
Administrative releases	25	0.205	0.191	44.0	0.20
Monetary policy decisions	113	0.059	0.172	44.2	0.60
Days without known events	2,962	−0.005	−0.031	51.6	–
<i>Panel B. Raw returns, bank index</i>					
Supervisory announcements	242	−0.030	0.061	48.3	0.77
Administrative releases	25	0.128	0.128	44.0	0.62
Monetary policy decisions	113	0.157	0.446	38.1	0.31
Days without known events	2,962	0.019	0.048	47.9	–
<i>Panel C. Raw returns, market benchmark</i>					
Supervisory announcements	242	−0.024	0.029	48.3	0.59
Administrative releases	25	−0.070	−0.137	56.0	0.66
Monetary policy decisions	113	0.089	0.291	39.8	0.34
Days without known events	2,962	0.022	0.053	47.0	–

Notes. Two-sided randomization *p* values on the difference in means against all other days, 20,000 draws stratified by calendar year and weekday, the same matching used for the variance tests. Panel A is the bank index net of 1.103 times the market, 1.103 being the estimate of λ in the mean equation; Panel B is the same index unadjusted; Panel C is the EURO STOXX 50 on the same dates. The 2,962 days without known events are the 2,963 event-free trading days of Appendix C less the first day of the sample, for which no return is defined.

D. First-moment evidence

Table D.1 reports returns by event class, first market-adjusted – the bank index net of 1.10 times the market – and then raw, both for the bank index and for the market benchmark alone.

Exactly half of the announcement days close negative, 121 of 242. Monetary policy days are not distinguishable in the mean either, which is the empirical counterpart of the identification argument in the main text: the sign of the surprise is not identified *ex ante* for either event class, so pooled mean tests are uninformative for both.

Panels B and C qualify that statement for monetary policy, without overturning it. Raw bank index returns average +0.157 on decision days against +0.019 on event-free days, with a median of +0.446 and only 38 per cent of days closing negative; the market benchmark alone averages +0.089 with a median of +0.291. This is the pattern documented for FOMC days on United States data, where the announcement-day return is a market-wide phenomenon rather than a sectoral one – Lucca and Moench

Table E.1: Dispersion of the surprise and estimated effects

<i>Panel A. Standard deviation of the surprise, basis points</i>			
	n	3-month	10-year
2013–2015	32	1.87	4.06
2016–2021 (lower bound)	48	1.81	2.09
2022–October 2023	15	6.16	6.28
<i>Panel B. Estimates</i>			
	$\hat{\beta}$	s.e.	p
3-month surprise, mean equation	0.015	0.047	0.75
3-month surprise, with long end	0.012	0.040	0.76
10-year surprise, mean equation	0.089	0.043	0.04
10-year surprise , variance equation	0.013	0.033	0.70

(2015), Savor and Wilson (2013) – and it is the component that the market adjustment of Panel A removes. It is also, with 113 decisions and a daily standard deviation near 1.7 percentage points, not distinguishable from zero: $p = 0.40$ raw against $p = 0.52$ adjusted. The gap between mean and median says where the difficulty lies – what signal there is sits in the centre of the distribution and is swamped by the tails – and it is a further reason to read the second moment rather than the first. For supervisory announcements nothing changes between panels: -0.003 adjusted, -0.030 raw, neither distinguishable from zero.

Separating supervisory releases by the direction of the signal, which the collected texts would permit, is left to further work. So is a second test those texts would support. Many ECB actions have a two-stage structure – a consultation is launched, the final act follows months later – and the stage is explicit in the release titles: anticipation implies a reaction to the first announcement and none to the second, immateriality implies neither. That comparison would separate the two readings the main text leaves observationally equivalent.

E. Monetary policy surprises

The intraday OIS surprises of Altavilla, Brugnolini, Gürkaynak, Motto and Ragusa (2019) are matched to 95 of the 113 monetary policy decisions; 13 of these carry a surprise of exactly zero. Of the eighteen unmatched decisions, seventeen fall beyond the database’s coverage, which ends on 26 October 2023, and the eighteenth is the emergency decision of 18 March 2020, announced at 23:45 and therefore outside the intraday window the database is built on. Surprises are dated to the announcement day and mapped to the trading day on which the news reaches the market.

The short-end surprise has a standard deviation below two basis points for eighty of the ninety-five matched decisions, the policy rate being at its lower bound over most of the sample, so its insignificance is a statement about the instrument rather than about the design. The long-end surprise, through which forward guidance and asset

purchases operate, enters the mean equation with a coefficient of 0.089: the design recovers a signed effect where a measure of surprise exists. The estimate provides no evidence that the conditional variance varies in proportion to the magnitude of the measured rate surprise. That is consistent with the reading adopted in the main text, and weaker than it: an absence of evidence over ninety-five decisions, most of them at the lower bound, does not establish that the variance responds to the arrival of information rather than to its size.

References

- Altavilla, C., L. Brugnolini, R. S. Gürkaynak, R. Motto, and G. Ragusa (2019). “Measuring Euro Area Monetary Policy”, *Journal of Monetary Economics*, **108**, 162–179.
- Bollerslev, T. and J. M. Wooldridge (1992). “Quasi-maximum Likelihood Estimation and Inference in Dynamic Models with Time-varying Covariances”, *Econometric Reviews*, **11**, 143–172.
- Lucca, D. O., and E. Moench (2015). “The Pre-FOMC Announcement Drift”, *Journal of Finance*, **70**, 329–371.
- Nyblom, J. (1989). “Testing for the Constancy of Parameters over Time”, *Journal of the American Statistical Association*, **84**, 223–230.
- Savor, P., and M. Wilson (2013). “How Much Do Investors Care About Macroeconomic Risk? Evidence from Scheduled Economic Announcements”, *Journal of Financial and Quantitative Analysis*, **48**, 343–375.