EPC Mirage: Grid decarbonisation can overstate retrofit progress in carbon-weighted building energy ratings

Alex Melendez-Ramos* Carles Vergara-Alert†

Abstract

Energy labels shape decarbonisation policy and investment, yet carbon-weighted ratings can improve when electricity systems change but buildings do not. We exploit the June 2022 revision to the methodology for non-domestic Energy Performance Certificates (EPCs) in England and Wales as a register-scale accounting discontinuity. Of 106,228 premises with repeat EPC assessments spanning the revision, 80.5% received better ratings, with gains concentrated among electricity-exposed premises and no comparable break in aggregate, unlinked metered-energy indicators. A cross-fitted, doubly robust estimator yields 94,911 expected EPC-B reclassifications across 57.2 million m^2, a notional retrofit-expenditure equivalent of £5.7 billion. France's 2026 electricity primary-energy coefficient revision similarly moves 103,085 certificates out of F/G, 89.1% among electrically heated certificates. Policy should separate carbon accounting from physical-performance standards, enabling a least-cost mix of cleaner energy supply and targeted retrofit rather than counting grid decarbonisation as building improvement.

As countries pursue net-zero targets, buildings have become a central focus of decarbonisation policy because they combine substantial energy demand with long-lived physical assets. Energy Performance Certificates (EPCs) were introduced as disclosure instruments that translate assessed building performance into comparable rating bands [1, 2]. Over time, labels have become central to regulation, investment decisions and market signals across the building stock. Government policy documents frame EPC thresholds as targets achieved through cost-effective improvement measures, while evidence from property markets shows that energy labels are capitalised into values and rents [3–10]. Yet, where headline ratings are carbon-weighted, certificate ratings can improve without corresponding improvements in building performance. As electricity grids decarbonise, a building can receive a better certificate because the emissions factor assigned to its electricity use has changed, rather than because its fabric, demand or systems have improved. This creates a measurement problem for policy, since the same band movement can represent physical retrofit, cleaner electricity accounting or a mixture of both.

The England and Wales non-domestic EPC regime provides a useful empirical setting because the same rating that carries regulatory and appraisal weight is mechanically exposed to revisions in fuel-emission factors. Under the National Calculation Methodology, a certificate compares a Unique Property Reference Numbers' (UPRNs) modelled carbon performance with that of a standard reference building, producing an asset rating that is mapped onto the A–G EPC scale [1, 2]. The resulting scale underpins the Minimum Energy Efficiency Standard, requiring privately rented non-domestic buildings to achieve EPC-E unless a valid exemption applies [11]. We treat EPC-B

*Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom. Email: alex.ramos24@imperial.ac.uk.

†IESE Business School, Av. Pearson 21, 08034 Barcelona, Spain. Banco Sabadell Chair in Business and Sustainability. Email: cvergara@iese.edu.

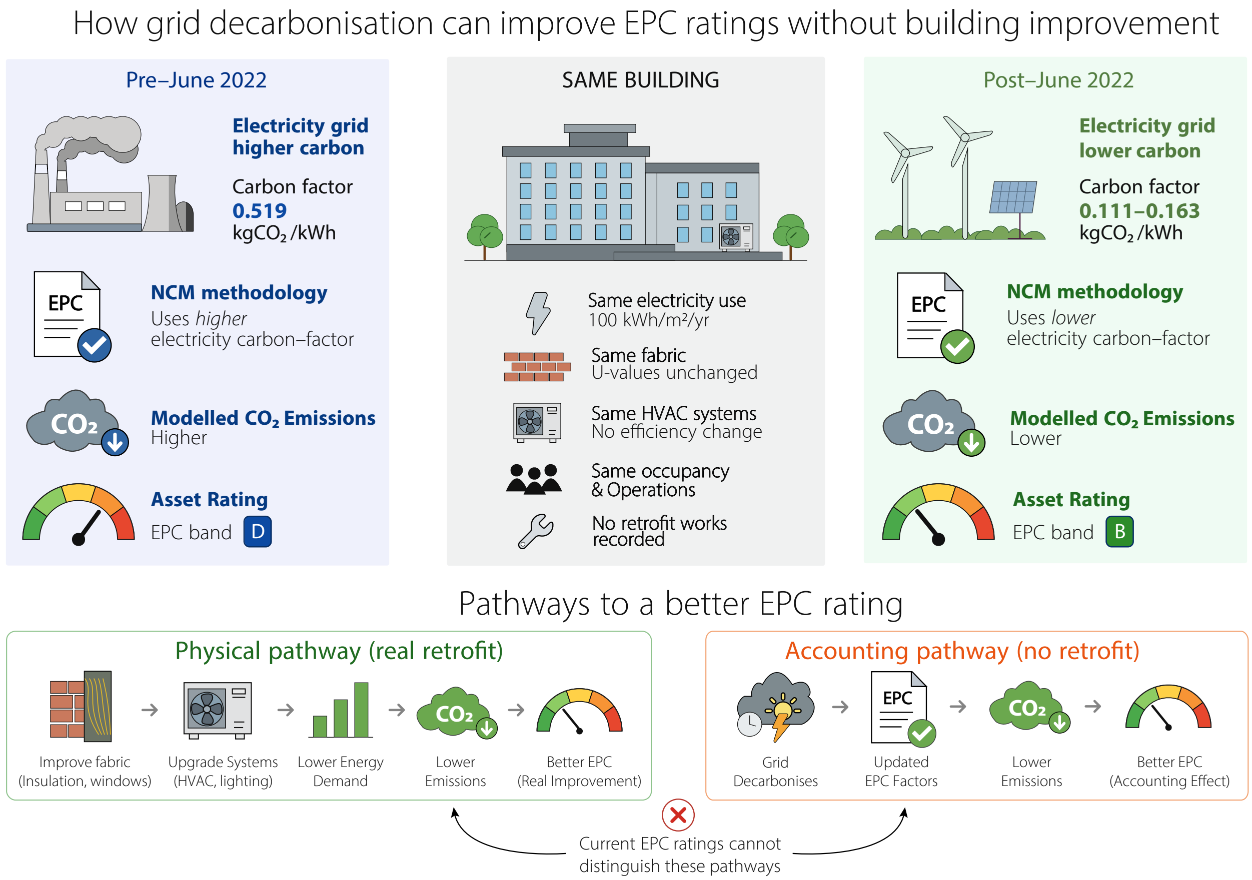

Figure 1: Two pathways to an improved carbon-weighted EPC rating. Conceptual schematic showing how an improved EPC band can reflect either physical building improvement or a revision to the accounting factors embedded in the rating. National Calculation Methodology (NCM) 2013 assigned grid electricity a fixed carbon factor of 0.519 $kgCO_2/kWh$, whereas NCM 2021 applies monthly factors ranging from 0.111 to 0.163 $kgCO_2/kWh$ (Table S1). For an otherwise unchanged building, the revised factors lower modelled emissions and can improve the Asset Rating. The schematic isolates the electricity-factor channel. The NCM revision also changed the reference-building calculation and other methodological inputs.

as a policy and appraisal benchmark because the government has stated its intention—subject to secondary legislation—to require privately rented non-domestic buildings over 1,000 m^2 to reach EPC-B from 2031, while retaining EPC-E for smaller buildings and dropping the earlier EPC-C milestone [3, 12]. Implemented on 15 June 2022 to reflect electricity-system decarbonisation, the revision updated the fuel-emission and primary-energy factors used in the EPC calculation, including a substantial reduction in the carbon assigned to grid electricity. This created a discrete methodology discontinuity in which electricity-exposed UPRNs could receive numerically lower, and therefore better, Asset Ratings without physical improvement. The revision therefore provides an empirical setting in which methodology-associated changes in certified performance can be separated, subject to stated counterfactual assumptions, from changes measured on a constant accounting basis.

Performance-gap studies show that modelled asset ratings can diverge from metered energy use, while UK non-domestic renewal studies question whether repeated certificates should be interpreted as stable measures of performance [13–18]. The UK government's own metrics review acknowledges

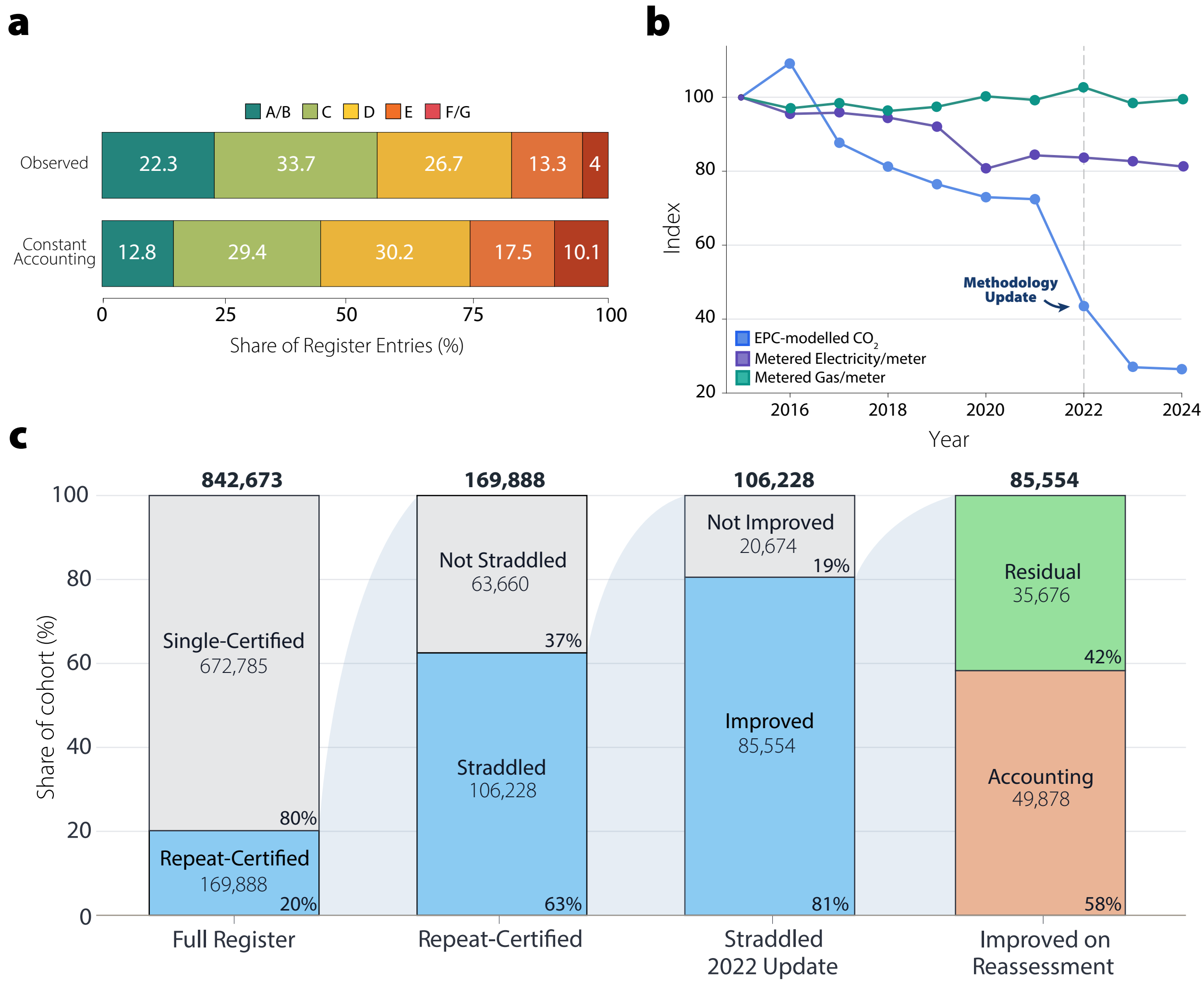

Figure 2: Register-scale decomposition of apparent EPC improvement. (a) EPC-band distribution in the England and Wales latest-certificate register frame. The upper bar shows the reported distribution, whereas the lower bar shows the transparent fixed-bridge baseline expressed on the estimated pre-2022 accounting basis. (b) Indexed comparison of mean EPC-modelled CO_2 with aggregate non-domestic electricity and gas consumption per meter, with all series normalised to 2015 = 100. The dashed line marks the June 2022 National Calculation Methodology update. The metered series are contextual aggregate indicators and are not building-level measurements linked to the EPC-register Unique Property Reference Numbers (UPRNs). (c) Sequential construction of the within-UPRN analysis sample. Improvers are divided according to whether the fixed bridge produces a larger methodology-associated component (accounting) or a larger residual component (residual).

that revisions to carbon or primary-energy factors can move UPRNs between bands without altering underlying energy performance [4]. Previous work has documented performance gaps and the sensitivity of building ratings to methodological choices, but has not quantified, at register scale, how revisions to system-level accounting factors alter apparent building-performance progress and policy-threshold attainment.

We address this gap using the June 2022 National Calculation Methodology (NCM) revision as an accounting discontinuity, linking repeat certificates using Unique Property Reference Numbers and constructing constant-accounting counterfactuals to estimate the scale and policy relevance of the resulting reclassification. We test whether apparent improvement follows the predicted fuel-exposure pattern, whether it persists among UPRNs with no administrative evidence of works, and whether the certified-emissions break is absent from independently metered energy-use indicators. We then estimate the aggregate number of register entries whose threshold status changes because of the accounting revision and assess whether the same metric-design problem extends beyond England and Wales using a European exposure typology and a same-certificate re-computation of France's 2026 primary-energy coefficient reform [19–21].

We show that methodology-associated reclassification materially alters the interpretation of national EPC progress. Our primary register-conditional estimator yields 94,911 expected EPC-B reclassifications across 57.2 million m^2 of affected floor area. A transparent fixed-bridge baseline independently raises the below-B share from 77.7% to 87.2% and produces 80,565 hard reclassifications. These complementary estimates are concentrated among electricity-exposed UPRNs, remain visible in a restricted sample with no administrative evidence of works and are not mirrored by a comparable break in aggregate, unlinked metered-energy indicators. Our contribution is not to establish that carbon-responsive ratings mechanically respond to changing energy-system factors, but to quantify the register-scale consequences for reported building-performance progress and policy-threshold attainment.

1 Results

1.1 Certified performance changes abruptly at the 2022 methodology revision

Figure 2B locates the apparent improvement in certified carbon performance at the June 2022 methodology break. Across 842,673 cleaned EPC-register UPRNs in the analysis register, mean asset rating falls from 95.9 in 2012 to 59.2 in 2025, while mean modelled Building Emission Rate falls by roughly 72%, from 106.8 to 30.0 $kgCO_2\,m^{-2}\,yr^{-1}$. Indexed to 2015, EPC-modelled CO_2 is broadly stable through 2021 and then falls discontinuously after the methodology update, with mean Building Emission Rate dropping from 84.4 in 2021 to 50.7 in 2022 and 31.4 in 2023. The break aligns with the introduction of the 2021 National Calculation Methodology, which replaced the fixed grid-electricity carbon factor of 0.519 $kgCO_2\,kWh^{-1}$ with monthly factors of 0.111–0.163 $kgCO_2\,kWh^{-1}$; 0.136 $kgCO_2\,kWh^{-1}$ is the October value used in the mechanism diagnostic (Table S1). Because the reference-building Standard Emission Rate fell by approximately 55–65% across fuel groups at the same revision, the asset rating declines less steeply than modelled emissions. The same panel shows no comparable discontinuity in the aggregate metered-energy indicators, which are not linked to the assessed UPRNs. Metered non-domestic gas consumption per meter changes by only 1% between 2021 and 2023 and remains within a $\pm 4\%$ band across 2020–2024, while modelled emissions recorded against gas-heated UPRNs more than halve over a comparable window. The carbon intensity of the electricity grid declines gradually over the period, whereas the factor embedded in the EPC calculation moves in a single step.

1.2 Rating gains follow electricity exposure and the accounting change

Within-building reassessments show that the post-2022 improvement was not limited to changes in the mix of certified UPRNs. Of the 106,228 UPRNs assessed on both sides of the June 2022 revision, 85,554 (80.5%) recorded a lower, and therefore better, asset rating (Fig. 2C). Under the transparent fuel-multiplier bridge, 49,878 improvers (58%) have a larger methodology-associated

Table 1: No-recorded-works mechanism diagnostic and fuel-specific counterfactual inputs across the 2022 methodology revision. The sample comprises repeat-certified UPRNs with unchanged recommendation counts and less than 2% change in reported floor area. The implied Building Emission Rate (BER) ratio is the new-to-old main-fuel carbon-factor ratio; electricity uses the representative October NCM 2021 factor. Observed BER and Standard Emission Rate (SER) ratios are post/pre medians, while m_f is the median pre/post asset-rating ratio used to construct the constant-accounting rating in Eq. (5). The observed gas BER ratio lies below the main-fuel implied ratio, reflecting grid-electricity use for lighting, ventilation, cooling and auxiliary loads in nominally gas-heated certificates. Rating windfall is the mean reduction in asset-rating points.

Heating Fuel	Counts	Carbon factor ($kgCO_2$/kWh)	Implied BER ratio	Observed BER ratio	SER ratio	Multiplier m_f	Rating windfall (pts)
Electric	1,578	0.519 → 0.136	0.262	0.231	0.360	1.633	35.2
Gas	746	0.216 → 0.210	0.972	0.377	0.452	1.197	15.3
Other	118	mixed	–	0.292	0.347	1.384	25.3

bridge component than residual component, whereas 35,676 (42%) have a larger residual component. On the same bridge-based decomposition, methodology-associated repricing represents approximately 81% of the aggregate improvement magnitude, rising to about 85% among electrically heated UPRNs and 67% among gas-heated UPRNs. These quantities do not provide individually validated classifications of physical retrofit.

If the rating changes follow the proposed accounting mechanism, the additional improvement associated with electricity exposure should be concentrated among reassessments spanning the 2022 update. Figure 3B shows asset ratings declining by 41.3 points on average among electrically heated UPRNs reassessed across the revision, compared with 18.3 points for gas-heated UPRNs and 26.1 points for other fuels. The electric–gas improvement gap widened from 18 points among pairs assessed entirely before the revision to 23 points among straddling pairs, then reversed among pairs assessed entirely afterwards, when gas-heated UPRNs improved by 24.3 points compared with 9 points for electrically heated UPRNs. This timing is consistent with a one-off re-pricing effect that is no longer present once both certificates share the post-revision accounting basis (Table S5). Formalising this change in the electric–gas differential, a pooled difference-in-differences model estimates an additional 8.6-point improvement for electrically heated UPRNs spanning the revision (95% CI 7.8–9.4; Table S4). The result is similar across narrow cut-off windows, building characteristics and transaction types (Tables S5 and S7). However, electrically heated UPRNs were already improving faster before the revision, so the fuel contrast supports the proposed mechanism but does not by itself identify the accounting effect. We therefore test a sharper implication of the revision by asking whether changes among UPRNs in the restricted sample with no administrative evidence of works match the magnitude implied by the published factor change.

Table 1 shows that rating changes in the restricted sample with no administrative evidence of works are consistent in direction and magnitude with the published factor revision. Among electrically heated UPRNs in this sample, the median post-to-pre ratio of modelled Building Emission Rate was 0.231, compared with 0.262 implied by the representative electricity carbon-factor change alone.

This correspondence provides the empirical basis for the fuel-specific re-pricing multipliers used in the constant-accounting counterfactual. Because nominally gas-heated certificates also include grid electricity for lighting, ventilation, cooling and auxiliary loads, the electric–gas contrast understates the full accounting effect (Supplementary Section 1.10). The revision also creates a rating advantage associated with recorded electrification. Among UPRNs initially assessed as gas-heated, 46.3% of

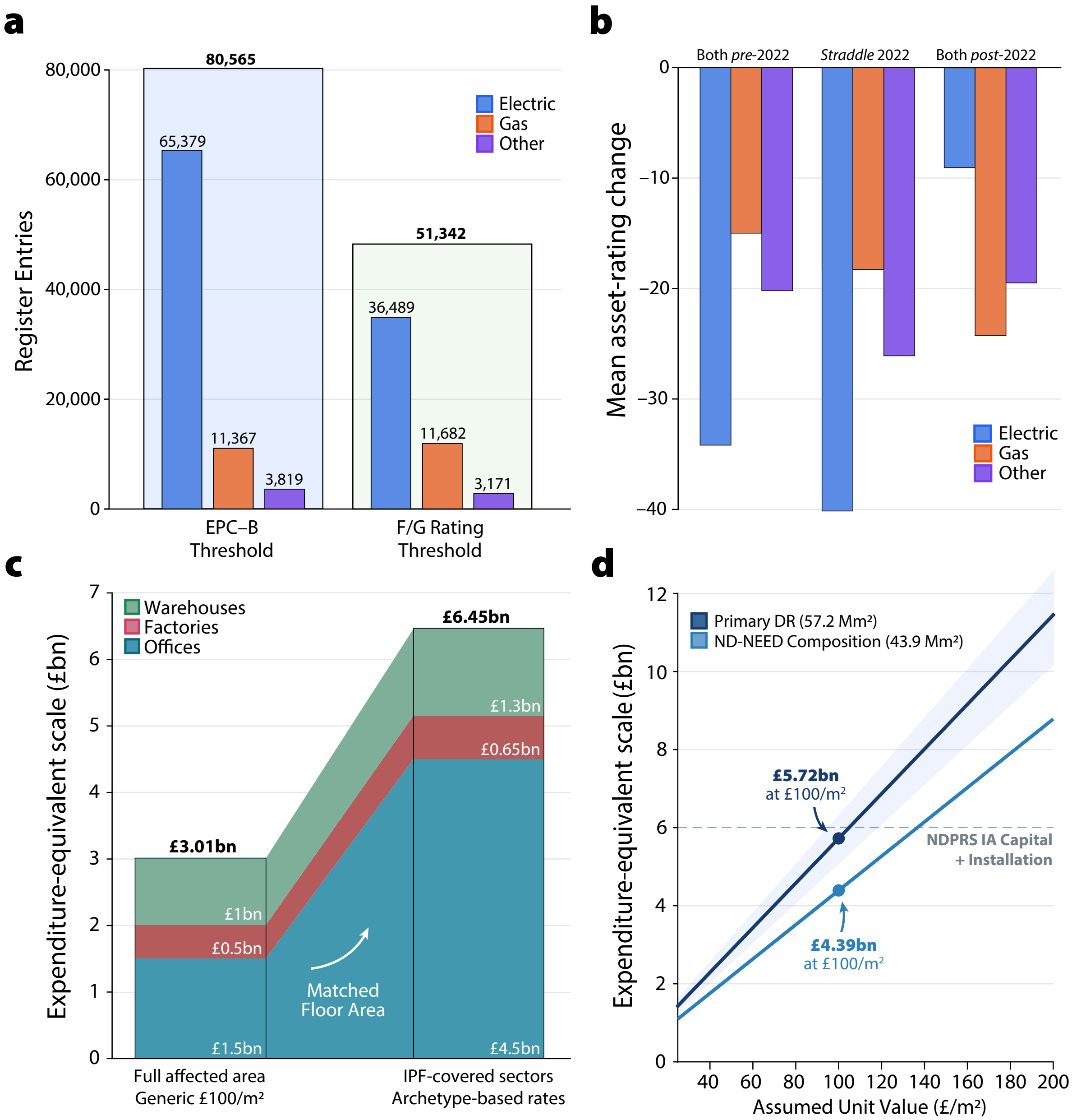

Figure 3: Threshold reclassification and expenditure-equivalent scale. (a) EPC-B and F/G reclassifications under the transparent fixed-accounting bridge, by heating fuel. (b) Mean within-UPRN asset-rating change by heating fuel and certificate period. (c) Generic and sector-specific expenditure-equivalent range for the IPF-covered share of expected EPC-B affected area. (d) EPC-B expenditure-equivalent scale over £25–200 m^{-2} using the primary doubly robust estimate (57.2 million m^2) and the ND-NEED composition sensitivity (43.9 million m^2). Shading shows the 95% bootstrap interval; the dashed line marks the impact-assessment capital-and-installation estimate.

those subsequently recorded as electrically heated reached EPC-B, compared with 16% of those remaining gas-heated, with asset ratings improving by 32.4 points on average, compared with 9.4 points among UPRNs remaining on gas. As registers fail to decouple heat pumps from resistance

Table 2: Policy-threshold reclassification of the latest-register frame under constant pre-revision accounting. Panel A reports the fixed-accounting bridge, which assigns post-2022 entries to old-basis bands using fuel-specific multipliers estimated from the no-recorded-works calibration sample. Parentheses in Panel A are 95% non-parametric intervals from 1,000 multiplier-bootstrap resamples, capturing uncertainty in the fuel-specific multipliers. Panel B reports the primary cross-fitted doubly robust Model C specification. Parentheses in Panel B are 95% full-pipeline bootstrap intervals from 1,000 resamples of the labelled calibration source with the register application frame held fixed. Both panels are conditional on the latest-register frame. Estimator validation and structural sensitivities are reported in Supplementary Sections 1.21–1.25.

Quantity	**Current basis**	**Constant-accounting estimate**
Panel A — Fixed-accounting bridge		
Below-EPC-B threshold rate (%)	77.7	87.2 (86.8–87.4)
F/G threshold rate (%)	3.96	10.1 (9.5–10.5)
Entries at EPC-B or better reclassified below B (n)	–	80,565 (76,493–81,865)
Entries at EPC-E or better reclassified into F/G (n)	–	51,342 (46,714–54,982)
Panel B — Primary cross-fitted doubly robust specification		
Expected entries reclassified below EPC-B (n)	–	94,911 (94,432–95,358)
Expected affected floor area (Mm^2)	–	57.2 (50.7–62.5)

heating, the pattern suggests a rating incentive associated with recorded electrification rather than improvements in delivered-energy efficiency.

1.3 Constant accounting reverses policy-threshold attainment

Figure 2A and Table 2 first report the transparent fixed-accounting bridge (Supplementary Section 1.18), which provides a readily interpretable baseline but does not recover precise old-basis ratings for individual entries. Under this bridge, the share of the latest-certificate register frame below EPC-B rises from 77.7% to 87.2%, and 80,565 entries currently rated EPC-B or better move below the threshold. The F/G share rises from 3.96% to 10.1%, with 51,342 entries currently rated EPC-E or moving into F/G under the fixed bridge (Supplementary Section 1.21). Our primary aggregate specification is the cross-fitted doubly robust threshold model, selected through held-out validation and transport diagnostics. It estimates 94,911 expected EPC-B threshold crossings (95% conditional bootstrap interval 94,432–95,358) and 36,847 expected F/G crossings.

Reclassification is concentrated among entries most exposed to the accounting revision. Figure 3A shows how electrically heated entries account for 65,379 of the 80,565 EPC-B-threshold crossings (81.2%), despite representing approximately 56% of the register, while being similarly over-represented at the F/G threshold. The same pattern appears within the repeated-UPRN sample, where electrically heated units account for 62.8% of EPC-B crossers compared with 52.9% of all straddlers. The observed-to-constant-accounting gap remains stable across alternative sample definitions, including certificates within their ten-year validity window, post-2022 certificates only, UPRNs above 1,000 m^2, separate analysis for England and Wales, stricter no-recorded-works calibrations and alternative multiplier specifications (S6, S9, S10, and S11).

Table 3: F/G exits by main heating energy in the France DPE factor-revision counterfactual. Rows group QC-cleaned certificate records by reported main heating carrier. Share of certificate flow is calculated relative to the 2,734,987 certificates retained after quality control. Pre-reform F/G certificates are classified under the 2.3 electricity primary-energy coefficient, and exits are certificates reclassified to E or better under the 1.9 coefficient. Counts describe certificate-flow quantities from the analysed register window.

Heating Energy	Certificates	Flow Share(%)	Pre-reform *passoires*	Exits	Exit-rate *passoires*(%)	Share of all Exits(%)
Electricity	1,122,161	41.0	189,920	91,836	48.36	89.09
Gas	1,065,224	39.0	88,107	6,818	7.74	6.61
Wood or biomass	89,460	3.3	16,268	2,006	12.33	1.95
Heat network	334,035	12.2	8,633	1,298	15.04	1.26
Oil/other	124,107	4.5	59,056	1,127	1.91	1.09

1.4 Accounting-conferred status has a multi-billion-pound expenditure scale

Accounting-induced threshold movement has a multi-billion-pound capital-expenditure-equivalent scale. We convert the expected affected floor area at the EPC-B threshold into investment units using a unit-cost range of £25–£200/m^2, with £100/m^2 as the central reference. The preferred cross-fitted estimator yields 94,911 expected EPC-B reclassifications across 57.2 million m^2 of affected floor area, corresponding to £5.7 billion at the central reference and £1.4–£11.4 billion across the full range. Alternative estimator and external-composition specifications yield central values of approximately £4.4–£6.5 billion at £100/m^2 (Supplementary Sections 1.18–1.25). Office and industrial/storage premises account for 52.6% of the primary affected area, with sector-specific IPF conversions reported separately (Supplementary Sections 1.23–1.24).

Government appraisal treats EPC-B attainment as evidence of cost-effective physical upgrades that deliver energy and carbon savings, yet the accounting revision moves certificates across the same threshold without those interventions. We therefore keep the expenditure-equivalent scale separate from the pathway's appraised social net present value, since the former translates affected floor area into investment units whereas the latter measures discounted net social welfare. Holding accounting constant leaves 735,100 register entries (87.2% of the latest-certificate frame) covering 380 million m^2 below EPC-B, with the estimated below-B register population concentrated among gas-heated premises. The scale of this divergence makes the design of the rating metric itself a broader policy concern. This register-level quantity is not a count of legally regulated premises, non-compliant properties or required retrofit projects.

1.5 Structural exposure across Europe and a same-certificate test in France

Metric design determines where similar reclassification is structurally possible beyond England and Wales, which provide register-scale evidence of methodology-associated reclassification. Figure 4 identifies where comparable factor sensitivity is structurally possible; it does not estimate realized reclassification outside the empirical cases. Delivered-energy metrics are insulated from factor revisions, whereas primary-energy and carbon-weighted ratings embed conversion coefficients or fuel-specific emissions factors that can change independently of building performance. Outside the UK, Spain and Portugal have the highest values on our ordinal structural-exposure index, while the Netherlands is notable for coupling a sensitive primary-energy metric to a current regulatory threshold (Table S3). Most other highly exposed systems remain linked to planned rather than binding thresholds, whereas France provides a separate same-certificate test of the same mechanism through an explicit coefficient revision in a comparatively low-carbon electricity system.

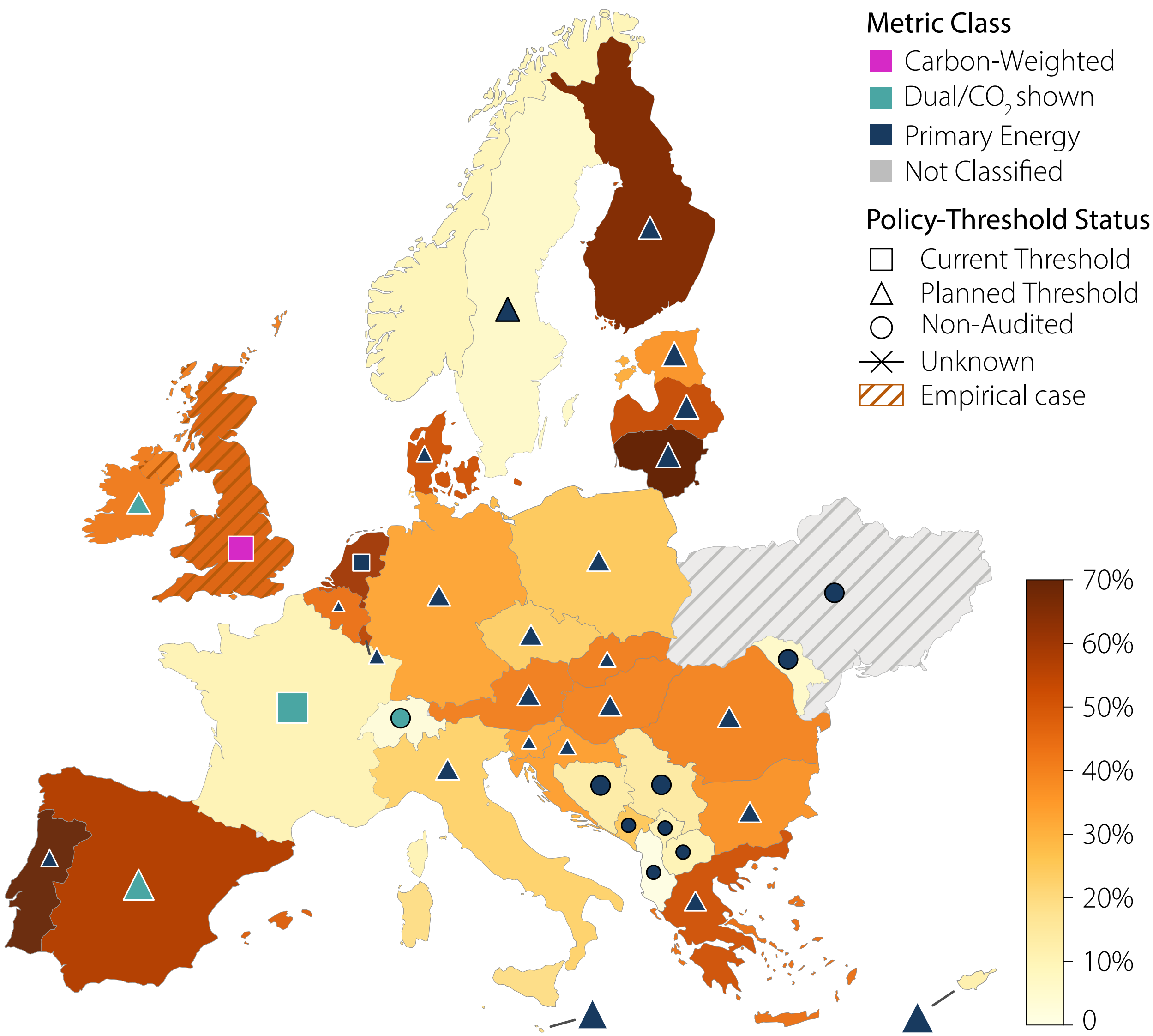

Figure 4: Structural exposure of European building-rating schemes to revisable energy factors. Country shading shows the 2015–2023 decline in electricity-generation carbon intensity. Marker colour denotes the headline rating metric, and marker shape denotes policy-threshold status. The hatched overlay identifies the England and Wales non-domestic EPC register analysed in this study; the grid-intensity series is UK-wide. Country coding, sources and confidence levels are reported in Table S3.

France occupies a distinctive position in the typology. Its relatively low structural-exposure score reflects limited recent grid-carbon change, yet the 2026 DPE reform directly revises the electricity primary-energy coefficient embedded in the rating. France therefore provides a focused test of whether factor revision itself can generate policy-relevant reclassification independently of rapid grid decarbonisation. Figure 5 examines this case through a same-certificate counterfactual of France's 2026 DPE reform, which lowers the electricity final-to-primary-energy coefficient from 2.3 to 1.9

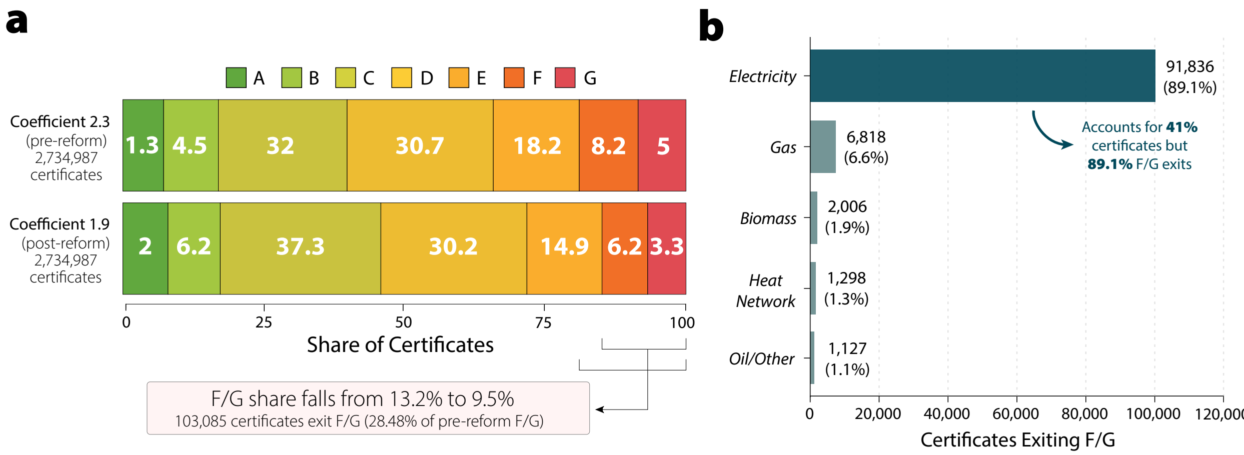

Figure 5: France's 2026 primary-energy coefficient reform reclassifies DPE labels through an electricity-exposure channel. (a) Same-certificate DPE label distributions under the pre-reform electricity primary-energy coefficient of 2.3 and the post-reform coefficient of 1.9, holding final energy, greenhouse-gas emissions, dwelling surface area and all other diagnostic inputs fixed. Labels are assigned using the official surface-dependent thresholds for dwellings below 40 m^2. The share of QC-cleaned certificates in classes F/G falls from 13.24% to 9.47%, with no certificate entering F/G. (b) Certificates moving from F/G under the 2.3 coefficient to E or better under the 1.9 coefficient, grouped by reported main heating energy. The coefficient revision produces 103,085 exits, of which 91,836 (89.1%) occur among electrically heated certificates. Counts describe certificate flows in the analysed register window.

[19]. Because the final label is determined by the worse of the primary-energy and greenhouse-gas classes, the reform can change a label only where the primary-energy criterion binds. Holding final energy, greenhouse-gas emissions, dwelling surface area and all other diagnostic inputs fixed, the revision moves 103,085 certificates in the analysed one-year flow from F/G to E or better, equal to 28.48% of pre-reform *passoires*, with none moving into F/G. Reclassification follows the predicted electricity-exposure channel. Electrically heated certificates account for 89.1% of exits and 48.4% of electrically heated *passoires* leave F/G, compared with 7.7% of gas-heated *passoires*, while no exit occurs where the greenhouse-gas class remains F/G. The recomputation reproduces 97.93% of published post-reform labels, and restricting the analysis to exact matches retains 102,004 exits (98.95%) of the headline estimate. Proxy deduplication and threshold-margin restrictions preserve the result (Supplementary Section 2.14). France therefore shows that accounting-driven threshold movement does not depend on carbon-weighted ratings or rapid grid decarbonisation, but arises whenever policy classifications embed revisable system-level coefficients.

2 Discussion

Building-performance policy requires knowing where improvement has occurred. Certificates can improve through lower energy demand, more efficient building systems or a reduction in the emissions factor assigned to electricity, while recording each pathway as the same band movement [4]. Exploiting the June 2022 NCM revision as a register-scale accounting discontinuity, we show that a single methodology update produced policy-relevant reclassification across the England and Wales non-domestic EPC register. Our primary cross-fitted doubly robust estimator yields 94,911 expected EPC-B threshold crossings and 57.2 million m^2 of expected affected floor area in the latest-certificate register frame. The transparent fixed-bridge baseline independently shows that the F/G share roughly triples from 3.96% to 10.1% when post-revision ratings are expressed on the estimated pre-revision basis. In doing so, the expenditure-equivalent measure translates the affected floor area into investment units, whereas the government's appraised social net present value measures discounted net social welfare [22]. The pattern is concentrated among electricity-exposed UPRNs, remains visible in a restricted sample with no administrative evidence of works and is not mirrored by a comparable break in aggregate metered-energy indicators that are not linked to the assessed UPRNs.

The measurement problem arises from bundling two quantities that policy analysis needs to keep distinct. One is the carbon intensity assigned to the energy a building uses, which falls as the electricity system decarbonises. The other is building-level performance, spanning demand, fabric quality and system efficiency, which changes through design, operation and retrofit. A carbon-weighted band collapses these channels and reattributes power-sector decarbonisation to the building stock. EPC bands function as regulatory thresholds, influence property values and rents, and underpin policy appraisal frameworks [5–7, 22]. Factor revisions therefore alter more than certificate labels: they change recorded threshold status and may affect regulatory classification within the legally covered subset, market signals, and the baseline against which future policy is assessed. The same mechanism can also create rating incentives. Once the electricity carbon factor falls, a carbon-weighted threshold rewards recorded movement towards electricity regardless of whether the installed system is efficient, crediting resistance heating alongside heat pumps while leaving delivered-energy demand, running costs and peak load unexamined. Where a band is reached through re-pricing, the register records progress without the physical interventions from which projected energy savings, bill reductions and comfort benefits would follow.

The mechanism follows from metric design rather than from the jurisdiction. Any rating system that converts modelled energy performance into bands using periodically revised factors is exposed to accounting-driven change. The exposure is lowest for delivered-energy metrics, where the reported quantity is the building's energy use itself, greater for primary-energy metrics through their dependence on conversion coefficients, and highest for operational-carbon metrics, which explicitly incorporate the current carbon intensity of fuels. Exposure therefore depends jointly on metric sensitivity to factor revisions and the pace of grid decarbonisation, with Figure 4 identifying where similar accounting-driven reclassification is structurally possible. France demonstrates that the mechanism extends beyond carbon weighting and rapidly decarbonising grids, as revising a single primary-energy coefficient moves 28.5% of F/G certificates across a threshold with rental consequences [19–21, 23, 24]. Under the recast EPBD, Member States must establish maximum performance thresholds for non-residential premises and residential renovation trajectories expressed in average primary-energy use. Because the underlying primary-energy factors may be periodically updated, analogous classification effects remain possible [25, 26].

The implication is not to choose grid decarbonisation over retrofit, but to prevent one from being counted as the other. For society, the relevant objective is the least-cost portfolio of cleaner energy supply, efficient electrification and targeted building improvement. Cleaner electricity can reduce operational emissions across many electrified premises without building-level capital expenditure, whereas well-targeted retrofits can reduce energy demand, bills and peak loads and improve comfort. Carbon-responsive labels should therefore disclose current operational emissions, while regulation intended to induce physical improvement should use stable building-level indicators and explicit cost-effectiveness tests. Major methodology revisions should also provide old- and new-basis ratings so that regulators, investors and markets can distinguish cleaner energy supply from changes to the building itself.

Several limitations qualify these conclusions. First, asset ratings are modelled rather than metered, so our counterfactual estimates methodology-associated reclassification rather than changes in actual energy use. Second, physical interpretation relies on aggregate, unlinked metered series, and unobserved assessor, input or operational changes cannot be excluded. Methodology changes not captured by the fixed bridge remain in the residual, which should not be interpreted as physical retrofit. Third, the primary threshold estimator is validated within the labelled calibration population and transported using observed characteristics; its register-wide interpretation requires conditional exchangeability between the calibration and application frames and may remain sensitive to unobserved service configurations, electricity end uses and assessment inputs. Fourth, the European analysis is a structural typology rather than an estimate of realized reclassification, and the France analysis describes certificate flows rather than the standing stock.

Carbon-weighted EPCs can remain an informative indicator of modelled operational-carbon performance under current factors while becoming an imperfect measure of physical improvement as grids decarbonise. The appropriate response is to pair carbon-responsive metrics with stable building-performance indicators, attributing gains from cleaner energy systems to the energy system and gains from renovation to the building stock itself.

3 Methods

3.1 Study Design

We use the 15 June 2022 introduction of NCM 2021 as an accounting discontinuity in the carbon and primary-energy factors embedded in non-domestic EPCs. Reassessments of the same UPRN on both sides of the cut-off provide the main within-UPRN comparison, while a restricted sample with stable recommendations and reported floor area calibrates methodology-associated repricing in the absence of administrative evidence of works.

We construct certificate-level, latest-register and within-UPRN datasets; estimate the discontinuity among straddling reassessments; and build constant-accounting counterfactuals for threshold reclassification. We assess physical interpretation separately using aggregate, unlinked metered-energy series and translate expected affected floor area into clearly labelled expenditure-equivalent quantities.

3.2 EPC data and rating framework

We used the official England and Wales non-domestic EPC register as the certificate-level record of assessed building performance [27]. The government bulk-data service provides data for domestic EPCs, non-domestic EPCs and Display Energy Certificates, and the non-domestic register contains certificate identifiers, inspection dates, asset ratings, floor area, property characteristics, building services information and recommendations generated at assessment. Importantly, the non-domestic EPC asset rating, produced using the National Calculation Methodology and approved software such as Simplified Building Energy Model (SBEM), is an asset measure, not a metered-energy measure. Under the NCM, the EPC asset rating is a modelled carbon-emissions index defined as

$$\mathrm{AR}_i = 50 \times \frac{\mathrm{BER}_i}{\mathrm{SER}_i}, \tag{1}$$

where BER_i is the building emission rate for the assessed building and SER_i is the standard emission rate for the reference building. Note that the NCM states that the standard emission rate corresponds to an asset rating of 50, which lies at the boundary between EPC bands B and C [2]. We therefore used $\mathrm{AR} \leq 50$ as the EPC-B threshold, $50 < \mathrm{AR} \leq 75$ as band C, $75 < \mathrm{AR} \leq 100$ as band D, $100 < \text{AR} \leq 125$ as band E, and $\text{AR} > 125$ as bands F/G. Full verification of the register identity is reported in the Supplementary Material.

The NCM 2021 guide came into force on 15 June 2022, creating an accounting discontinuity in the carbon and primary-energy factors used in non-domestic EPC calculations [2]. Under the 2013 NCM guide, grid-supplied electricity had a carbon factor of 0.519 $kgCO_2$/kWh and a primary-energy factor of 3.07 kWhPE/kWh, while natural gas had a carbon factor of 0.216 $kgCO_2$/kWh. Under NCM 2021, grid electricity is represented with monthly carbon factors, including 0.136 $kgCO_2$/kWh in October, and monthly primary-energy factors of approximately 1.4 to 1.6 kWhPE/kWh. The natural-gas carbon factor was essentially unchanged across the revision, moving only from 0.216 to 0.210 $kgCO_2$/kWh, so gas-heated UPRNs were far less exposed to the accounting change than electrically heated UPRNs.

3.3 Certificate cleaning and cohort construction

The initial dataset comprised 1,245,956 non-domestic EPC certificates. We first restricted the dataset to certificates that contained the core information required for either stock-level analysis or within-building comparison. Specifically, we excluded certificates with invalid, extreme or missing asset ratings, missing inspection dates, non-positive reported floor area, missing Unique Property Reference Number (UPRN), or inspection years outside the period from 2012 to 2025. The UPRN is a persistent identifier assigned to each addressable location in the United Kingdom and designed to remain stable over the life cycle of a property [28]. UPRNs are address-level identifiers rather than whole-building identifiers, so a single physical building may contain several UPRNs and some certificates may describe units within a building; we therefore refer throughout to cleaned EPC-register UPRNs, and the "stock" quantities are register-level counts rather than a validated census of the national non-domestic building stock. After these exclusions, certificates were sorted by UPRN and inspection date.

We linked certificates sharing the same UPRN. Following the 2022 methodology revision, which altered the electricity carbon factor substantially more than the gas carbon factor, certificates were also grouped by recorded main heating fuel [1, Appendix C, Table 27][2, Appendix C, Tables 29-30]. This classification allowed us to distinguish UPRNs whose assessed energy uses were more directly exposed to the accounting change from those less affected by it.

The cleaned certificate frame contained 842,673 unique UPRNs. For stock-level analyses, we retained the latest certificate for each UPRN, producing a latest-certificate frame that represented the most recent observed EPC status of each cleaned register UPRN in the register. For within-building analyses, we identified UPRNs with at least two valid certificates separated by more than three months, excluding near-duplicate lodgements and administrative reissues that were unlikely to represent a meaningful reassessment interval. This yielded 169,888 reassessed UPRNs. Of these, 57,892 had both first and last certificates before the 15 June 2022 cut-off, 106,228 straddled the cut-off, and 5,768 had both first and last certificates after the cut-off. Among the straddling UPRNs, 85,554 had lower asset ratings after reassessment, 19,711 had higher asset ratings, and 963 showed no change (Table S2, Fig. 2).

3.4 Within-building comparisons

The within-building analysis compared certificates for the same UPRN over time. For the main straddling analysis, we used the first and last valid certificates for each reassessed UPRN to maximize the contrast between the pre-2022 and post-2022 accounting regimes while holding the building identifier fixed. For UPRN i, we defined the change in asset rating as $\Delta\text{AR}_i =$

$\mathrm{AR}_{i,\mathrm{last}} - \mathrm{AR}_{i,\mathrm{first}}$ where lower asset ratings represent better certified performance and negative values of $\Delta\mathrm{AR}_i$ represent apparent improvement. Additionally, we computed a modelled-emissions ratio, BER Ratio$_i = \frac{\mathrm{BER}_{i,\mathrm{last}}}{\mathrm{BER}_{i,\mathrm{first}}}$, the time between assessments in years, the change in the number of recommendations, and the fractional change in reported floor area defined as,

$$\Delta\mathrm{FA}_i = \frac{|\ \mathrm{FA}_{i,\mathrm{last}} - \mathrm{FA}_{i,\mathrm{first}}\ |}{\mathrm{FA}_{i,\mathrm{first}}}. \tag{2}$$

Each pair was classified as both-pre, straddling or both-post according to the dates of the first and last certificates. First-last pairs were used when the target quantity was the total change across the methodology cut-off. Consecutive pairs were used for transition-rate estimates, where the estimand is the adjacent reassessment step rather than the endpoints of a building's full certificate history.

3.5 No-recorded-works calibration sample

To estimate the rating movement caused by the accounting revision, we constructed a conservative calibration sample of straddling UPRNs with no recorded change in recommendations and less than 2% change in reported floor area. The calibration sample helped identify UPRNs that were observed on both sides of the methodology change but had no clear administrative evidence of physical alteration. The calibration set was

$$\mathrm{NW} = \{i : dt_{i,\mathrm{first}} < \mathrm{CUT} \leq dt_{i,\mathrm{last}},\ \Delta\mathrm{rec}_i = 0,\ \Delta\mathrm{FA}_i < 0.02\}\,. \tag{3}$$

The 2% floor-area threshold was used as a conservative tolerance that allowed small assessor remeasurement differences while excluding UPRNs with material changes in reported size. In addition, since recommendations are regenerated at assessment and can change even when no major retrofit is recorded, we further included unchanged recommendation counts as a secondary guardrail. The no-recorded-works definition was therefore not interpreted as proof that a building had no physical change and was used as a calibration device for estimating the accounting discontinuity. Although small ($n = 2{,}443$), the calibration sample is broadly representative of the register in size and sector composition (median floor area 194 m^2 versus 181 m^2 register-wide, with roughly three-quarters of UPRNs below 500 m^2 in both, and the same office and retail sectors predominating). Fuel-specific estimation addresses differences in fuel shares but does not by itself establish transportability; we therefore evaluate and correct transport using the procedures described in Supplementary Sections 1.19–1.22.

3.6 Constant-accounting counterfactual

The main counterfactual re-expressed post-2022 certificates on the pre-2022 accounting basis. We estimated, for each fuel group f, a multiplier from no-recorded-works straddlers,

$$m_f = \mathrm{median}_{\ i\in\mathrm{NW}_f} \left(\frac{\mathrm{AR}_{i,\mathrm{pre}}}{\mathrm{AR}_{i,\mathrm{post}}}\right) = \mathrm{median}_{\ i\in\mathrm{NW}_f} \left[\frac{\mathrm{SER}_{\ i,\mathrm{post}}/\mathrm{SER}_{i,\mathrm{pre}}}{\mathrm{BER}_{i,\mathrm{post}}/\mathrm{BER}_{\ i,\mathrm{pre}}}\right]. \tag{4}$$

Equation (4) follows directly from AR = 50 × BER/SER. For no-recorded-works straddlers, the pre-to-post ratio of ratings estimates the factor by which a post-2022 rating must be scaled to recover the rating that would have been observed under the old accounting basis. In addition, we then defined the constant-accounting rating for each cleaned register UPRN in the latest-certificate frame as

$$\mathrm{AR}_i^{\mathrm{const}} = \begin{cases} \mathrm{AR}_i, & dt_i < CUT, \\ \mathrm{AR}_i \times m_{\mathrm{fuel}_i}, & dt_i \geq CUT. \end{cases} \tag{5}$$

In doing so, the procedure maintained pre-2022 certificates unchanged and rescaled post-2022 certificates to a common pre-2022 accounting framework. Because the multiplicative method is transparent and preserves relative differences within each fuel group, we retain it as a descriptive baseline and use it for the bridge-based decomposition of reassessment changes. Held-out validation shows that it does not recover precise individual old-basis ratings; policy-threshold counts and affected floor area are therefore based primarily on the cross-fitted doubly robust threshold model described in Supplementary Sections 1.18–1.25. The 2021 NCM update bundled the electricity carbon-factor revision with other methodology changes, including revised primary-energy factors, the 2021 notional building and updated activity-database assumptions. The empirically estimated fuel-specific multiplier therefore captures the full fuel-differential methodology-associated re-pricing rather than the carbon-factor change in isolation. The multiplier captures the bundled, fuel-differential methodology revision rather than the electricity carbon-factor change in isolation. The size of the electricity-factor reduction and the correspondence between the predicted and observed BER ratios indicate that this reduction is likely the dominant fuel-specific channel, but the counterfactual does not isolate it exactly. We therefore interpret the electric–gas contrast as mechanism-consistent evidence rather than as a standalone causal estimate.

By contrast, because the additive windfall estimator applies an average rating-point correction rather than a proportional adjustment, it was used as a robustness check. Uncertainty in the constant-accounting estimates was quantified using 1,000 bootstrap resamples of the no-recorded-works calibration sample. Resampling was performed separately within fuel groups under an identical random seed (Supplementary). In particular, for each bootstrap replicate, we re-estimated the fuel-specific multipliers, applied them to the latest-certificate frame, and recomputed the national quantities. We report the 2.5th, 50th and 97.5th percentiles of the bootstrap distribution (Table 2). This bootstrap captures sampling uncertainty in the baseline no-recorded-works multiplier only.

To assess whether the counterfactual depends on transporting a single fuel-group multiplier to heterogeneous UPRNs, we separately repeated the reclassification under progressively more granular multiplier definitions, estimating multipliers by fuel crossed with sector, floor-area bin, air-conditioning status and pre-revision rating band, and collapsing cells with fewer than 50 calibration observations to the fuel-level multiplier. We also varied the no-recorded-works floor-area tolerance and estimated the multiplier by cross-fitting. We report the resulting range as the measure of structural sensitivity to the counterfactual specification, separately from the bootstrap interval (Supplementary Section 1.9). Stratification by pre-revision rating band is reported for completeness but is not a valid sensitivity check, because the multiplier $\mathrm{AR}_{\mathrm{pre}}/\mathrm{AR}_{\mathrm{post}}$ is mechanically correlated with a building's pre-revision rating, so conditioning on it induces regression to the mean.

3.7 Decomposition of observed improvement

To quantify the proportion of the observed improvement attributable to the accounting revision after applying the fuel-specific counterfactual correction for straddling UPRNs, we decomposed the observed asset-rating change into a formula component and a residual component. Specifically, for each UPRN, we defined

$$\Delta\mathrm{AR}_i = (\mathrm{AR}_{i,\mathrm{post}} - m_{\mathrm{fuel}_i}\mathrm{AR}_{i,\mathrm{post}}) + (m_{\mathrm{fuel}_i}\mathrm{AR}_{i,\mathrm{post}} - \mathrm{AR}_{i,\mathrm{pre}}) \tag{6}$$

where the first term is the formula component, capturing the difference between the observed post-2022 rating and the same post-2022 certificate re-expressed on the old accounting basis, and the second term is the residual component, capturing the remaining change after removing the fuel-specific formula effect. Specifically, the residual includes physical upgrades, modelling changes not captured by the fuel-specific adjustment, and measurement error. We further reported the aggregate improvement-point shares as

$$\text{Share}_k = \frac{\sum_i \text{Component}_{ik}}{\sum_i \Delta\text{AR}_i}, \tag{7}$$

where k denotes the formula or residual component. Improvers were defined as UPRNs with $\Delta\text{AR}_i < 0$; because lower asset ratings indicate better performance, an improvement is a negative ΔAR_i, and component shares are reported on the signed improvement scale so that the formula and building components sum to the total observed change. Importantly, to prevent classifying negligible rounding changes as substantive rating movements, the decomposition was applied to UPRNs with $\mid \Delta\text{AR}_i \mid \geq 1$, assuming a common formula effect within each fuel group that can be transported from no-recorded-works UPRNs to the broader straddling sample. Any remaining methodology changes not captured by the fuel-specific adjustment were assigned to the residual component, so the formula share captures only the component represented by the fixed bridge.

The two categories shown in Figure 2c are obtained from this continuous decomposition by a per-improver majority rule. Writing $f_i = \text{AR}_{i,\text{post}}(1 - m_{\text{fuel}_i})$ for the (non-positive) formula component and $r_i = m_{\text{fuel}_i}\text{AR}_{i,\text{post}} - \text{AR}_{i,\text{pre}}$ for the residual component, an improver is classified as accounting-based when the formula component accounts for at least half of its observed improvement, $f_i/\Delta\text{AR}_i \geq 0.5$ (equivalently $|f_i| \geq |r_i|$), and as residual/building-attributed otherwise. This is a comparison of magnitudes and does not constrain the sign of the residual. It assigns 49,878 improvers (58%) to the accounting category and 35,676 (42%) to the residual category. Reporting the sign structure separately, and because every fuel multiplier exceeds one so the formula component always improves the rating: about 68% of improvers (approximately 58,200) have accounting and residual components acting in the same direction, both improving the rating; about 32% (approximately 27,300) have a residual component that offsets the accounting component, so the rating is no better – indeed worse – on the old accounting basis and the accounting component alone therefore exceeds the total observed improvement; and no improver has an accounting component that worsens the rating. The accounting category is thus conservative, since an improver is counted as building-attributed whenever a genuine old-basis improvement is the larger of the two components.

3.8 Treatment of modelled primary energy

The non-domestic register includes a modelled primary-energy field, but the main empirical design uses asset ratings and modelled emission rates for the accounting-discontinuity analysis. Primary energy is computed by applying primary-energy conversion factors to modelled delivered-energy components, and NCM 2021 also revised the electricity primary-energy factors. It is therefore a modelled, factor-weighted quantity rather than a delivered-energy measurement. We use primary energy to describe the structure of the NCM and the exposure of primary-energy metrics to coefficient revision. Physical energy-use validation is carried by the external metered non-domestic energy-consumption series described below.

3.9 Aggregate metered-energy comparison

To test whether changes in EPC-modelled emissions tracked changes in observed energy consumption, we compared certificate-based trends with external metered-energy data. In particular, we aggregated the EPC register by inspection year and indexed mean modelled building emissions to 2015 = 100. Leveraging the non-domestic electricity consumption from the Department for Energy Security and Net Zero (DESNZ), and including both mean consumption per meter, we further constructed comparable 2015 = 100 indices for DESNZ non-domestic mean electricity and gas consumption per meter for England and Wales [29, 30]. Since UPRNs assessed in one year are not necessarily the same UPRNs assessed in another year the EPC series constitutes a repeated certificate cross-section. Because of that, we therefore interpreted divergence between both sources as evidence that certificate-based modelled emissions and observed energy consumption were measuring different quantities, since a certificate index can exceed 100 after the base year when the UPRNs certified in that year have higher average modelled emissions than UPRNs certified in the base-year cohort.

For the grid comparison, we use annual UK electricity-carbon-intensity data from Ember via Our World in Data [31]. We compare the gradual annual change in the grid series with the discrete change in the factors embedded in NCM 2021. In doing so, the comparison tested whether the EPC calculation factor followed the gradual decarbonisation of the electricity grid or introduced a discrete re-pricing shock when the NCM methodology changed.

3.10 EPC-B crossings and accounting-conferred threshold status

We first identified observed EPC-B crossers, defined as UPRNs that moved from below EPC-B before the cut-off to EPC-B or better after the cut-off. Since lower asset ratings indicate better EPC performance, this corresponds to UPRNs with a pre-cut-off asset rating above the EPC-B threshold and a post-cut-off asset rating at or below the threshold. Formally, building i was classified as an observed EPC-B crosser when $AR_{i,\text{pre}} > 50$ and $AR_{i,\text{post}} \leq 50$. For this subset, we then tested whether the EPC-B crossing remained after holding the accounting method constant. UPRNs were classified as reaching EPC-B on paper when the observed post-2022 certificate was EPC-B but the same certificate would not have been EPC-B under constant accounting:

$$PaperB_i = 1\{AR_{i,\text{pre}} > 50,\ AR_{i,\text{post}} \leq 50,\ AR^{const}_{i,\text{post}} > 50\}. \tag{8}$$

This test separates threshold crossings that remain under a constant accounting basis from threshold crossings generated by the accounting change. Physical energy-use interpretation is based on the external aggregate metered-energy comparison described above.

3.11 Policy benchmarks

We treat EPC-B as a policy and appraisal benchmark rather than as the current legal standard for all non-domestic rented premises. Earlier consultation and appraisal documents proposed a tightening trajectory to EPC-B by 2030, subject to cost-effectiveness and implementation details [3]. The June 2026 interim response states that the proposed EPC-C milestone will not be taken forward and that the intended higher standard is EPC-B from 2031 for privately rented non-domestic buildings over 1,000 m^2 in England and Wales, where cost-effective and subject to secondary legislation; UPRNs below 1,000 m^2 remain subject to the current EPC-E minimum unless policy changes further [12]. We therefore use EPC-B by 2030 only as the historical consultation and impact-assessment benchmark; the government's current stated intention is EPC-B from 2031 for the specified subset. On the latest-certificate frame, we reported the observed and constant-accounting shares of register entries below EPC-B and in F/G. Below-B was defined as AR > 50 and F/G was defined as AR > 125. We also reported the number of UPRNs observed at EPC-B that would be below B under constant accounting, and the number of UPRNs observed at E or better that would remain F/G under constant accounting. We report the multiplicative fixed-factor recomputation as a transparent baseline. The primary aggregate threshold estimates use the cross-fitted doubly robust model, with alternative model architectures, transport specifications, external-composition adjustments, and the additive windfall estimator reported as distinct robustness checks (Supplementary Section 1.8).

3.12 Reassessment-throughput sensitivity (Markov)

As an illustrative sensitivity to reassessment throughput, we asked one narrow question: does the accounting-sensitive below-EPC-B gap disappear as the register is reassessed, or does it persist when the starting distribution and the transition matrix are defined on commensurate rating bases? This exercise is illustrative and is not a forecast of legal compliance, national stock, or physical retrofit progress. Its only requirement for internal validity is that the starting distribution and the transition matrix be defined on the same accounting basis within a scenario; the earlier version, which combined a constant-accounting starting distribution with a transition matrix estimated from observed post-revision ratings, mixed bases and is not used. Because the fixed-accounting bridge does not support exact individual old-basis ratings, the preferred specification works at the threshold level with expected states rather than hard full-band classifications. Each certificate has a probability of being below a threshold; for a reassessment pair with previous below-threshold probability p_0 and current below-threshold probability p_1, the expected transition mass is $(1-p_0)(1-p_1)$, $(1-p_0)p_1$, $p_0(1-p_1)$ and p_0p_1 across the above/below states. We construct two internally commensurate scenarios (Table S13). The observed / current-accounting scenario uses the observed latest-certificate below-threshold share and a transition matrix from observed post-2022 consecutive reassessments, both on the reported post-revision basis. The constant-accounting scenario expresses every starting and transition rating on the pre-2022 basis (pre-2022 certificates unchanged; post-2022 ratings mapped to their expected old-basis threshold state through the empirical no-works re-pricing ratio distribution). For each scenario with starting below-threshold share s_0 and threshold transition matrix T, we projected the below-threshold share after reassessing a cumulative share p of the frame as $(1-p)s_0 + p\,(s_0T)$, and reported it against p. The hard full-band matrices are retained only as a Supplementary sensitivity (Table S18). Implementation, back-testing, structural-uncertainty and transition-sample robustness details are in Supplementary Section 1.16 and Tables S12–S17.

3.13 Capital-expenditure equivalent

To express the scale of accounting-induced threshold movement in investment units, we defined a capital-expenditure equivalent. This estimate is a scale conversion, not an estimate of observed retrofit expenditure, avoided compliance cost or social welfare. For a set of UPRNs C that crossed a policy threshold on paper, and a unit cost c in pounds per square metre, we computed:

$$\text{Capex}(C, c) = \sum_{i \in C} \text{FloorArea}_\text{i} \times c \tag{9}$$

and evaluated c with £100/m^2 as the central sensitivity value. For full implementation details see Supplementary Section 1.17. The threshold-crossing sets were defined using the fixed-factor counterfactual. The EPC-B set contained UPRNs observed at EPC-B that would be below B under constant accounting, and the legal-floor set contained UPRNs observed at E or better that would remain in F/G under constant accounting. These two sets are disjoint by construction: a building in the EPC-B set has an observed rating $\text{AR} \leq 50$, so even under the largest fuel-specific multiplier (≈ 1.63) its constant-accounting rating remains below 82 and cannot exceed the $\text{AR} > 125$ threshold required for the legal-floor set. We verified that no building belongs to both hard sets, so a hard-set union would equal the sum; we nonetheless report the two thresholds separately and do not combine them, because under the preferred probability model the expected F/G-crossing area is nested within the expected EPC-B-crossing area rather than disjoint from it.

The unit-cost grid is not a costed measure schedule and is deliberately reported as a range. Its span is chosen to bracket published non-domestic energy-efficiency improvement costs, which vary widely by measure, building type and the depth of upgrade required: light-touch measures on better-rated stock sit near the lower bound, while deep fabric-and-services retrofits of the worst-rated premises sit near the upper bound [22, 32]. A single common rate is therefore a simplification; because the affected floor area is concentrated in particular sectors (Table S24), a sector- or band-specific unit-cost vector would refine the scale, and the affected-area-by-sector composition needed to apply one is reported in the area-weighted validation above. We retain the common-rate grid for transparency and because the dominant uncertainty in the scale is the affected area itself, not the unit cost.

We do not benchmark the expenditure-equivalent scale against the government's appraised social net present value for the EPC-B pathway. The two are different estimands – an area-to-cost scale conversion using an assumed common unit-cost grid, versus a discounted net social welfare estimate net of costs, energy and carbon benefits – and a direct numerical or visual comparison would be misleading. We therefore report the expenditure-equivalent scale on its own terms, as a bounded floor-area-to-cost conversion whose affected-area input is itself only partially validated (Supplementary Section 1.19).

3.14 International structural-exposure typology

The international analysis is a structural exposure typology rather than a causal estimate. The England and Wales non-domestic register provides register-scale evidence of methodology-associated label reclassification, while the international analysis identifies where similar factor sensitivity is structurally possible. For each European country or territory in the typology, we combined electricity-grid decarbonisation between 2015 and 2023 with a metric-sensitivity score based on the headline building-rating metric. Country coding, sources and confidence levels are reported in Table S3.

Metric sensitivity was coded on a four-level scale. Delivered-energy metrics were assigned 0 because their headline quantity does not depend on carbon or primary-energy conversion factors; the underlying delivered energy may itself be modelled or measured. Primary-energy metrics were assigned 1 because primary-energy conversion coefficients can be revised and can move labels even when delivered energy is unchanged. Dual primary-energy/CO_2 or CO_2-reported systems were assigned 1.5 because carbon is visible but is not necessarily the sole headline threshold metric. Carbon-weighted headline metrics were assigned 2 because the label directly incorporates fuel-specific carbon factors and is therefore most exposed to grid decarbonisation and factor revision. Countries that could not be classified were coded as missing, not zero.

Policy-threshold status was coded as a separate variable: current threshold, planned threshold, no audited threshold, or unknown/not audited. It is not combined into the metric-sensitivity score. A country can therefore have a sensitive metric without a binding threshold, or a less sensitive metric with a binding threshold. Grid decarbonisation was measured as the percentage reduction in electricity-generation carbon intensity between 2015 and 2023 using Ember/Our World in Data. For the exposure index, we used a non-negative decarbonisation term, $\max(0, \mathrm{GridReduction}_j)$, because the mapped risk is apparent improvement through falling grid intensity rather than apparent worsening through rising grid intensity. For country j, the exposure index is:

$$\mathrm{Exposure}_j = \mathrm{MetricSensitivity}_j \times \max(0, \mathrm{GridReduction}_j). \tag{10}$$

The index is used only as an ordinal ranking of structural exposure. It is not a predicted reclassification rate, and Figure 4 should not be interpreted as an estimate of realised reclassification outside England and Wales. Bloc reference rows, if included in the supplementary table, are excluded from country ranks and plotted country markers. France is treated separately as a coefficient-revision case:

its grid-speed exposure score is low, but the 2026 DPE reform demonstrates the primary-energy factor-revision channel.

3.15 External validity: France case study

To test whether the mechanism generalises from a carbon-factor revision to a primary-energy coefficient revision, we recomputed residential DPE labels using the French Agency for Ecological Transition (ADEME) register of existing-dwelling certificates under both the pre-reform and post-reform electricity primary-energy coefficients [24]. The post-reform coefficient is 1.9, while the pre-reform coefficient is 2.3. Because the published register already reports primary energy using the 1.9 coefficient, the counterfactual 2.3 value can be recovered at certificate level by adding back the difference in the electricity conversion factor. Let A_i denote dwelling surface area, E_i^{elec} total final electricity consumption, and $P_i^{1.9}$ total five-use primary energy reported under the 1.9 electricity primary-energy coefficient. The counterfactual total primary energy under the previous 2.3 coefficient is

$$P_i^{2.3} = P_i^{1.9} + 0.4\,E_i^{\text{elec}}, \tag{11}$$

and primary-energy intensity is $p_i^r = P_i^r/A_i$ for $r \in \{1.9, 2.3\}$. DPE classes are assigned from p_i^r and the fixed greenhouse-gas intensity using the official residential thresholds, including the surface-dependent class cut-points that apply to dwellings below 40 m^2 [19]. The expression follows directly from the DPE accounting identity, since the reform changes only the primary-energy coefficient applied to electricity and leaves non-electric carriers unchanged. Conditional on the carrier-reconciliation quality-control screen described in the Supplementary Information, the recomputation is therefore a same-certificate accounting transformation rather than a model of behavioural change, renovation, or altered energy demand. Once the official small-surface thresholds are implemented, it reproduces the published post-reform label for 97.9% of QC-cleaned certificates overall and for at least 97.8% within every surface group, so the headline is the full certificate flow rather than a surface subset. Residential DPE labels are assigned using both a primary-energy class and a greenhouse-gas class, with the final label given by the worse of the two. The coefficient reform can therefore affect only the primary-energy axis. A dwelling can leave classes F/G under the 1.9 coefficient only if its greenhouse-gas class is already E or better. If the greenhouse-gas axis is itself F or G, lowering the electricity primary-energy coefficient cannot move the final DPE label out of F/G. We report certificate-level reclassification flows in the main analysis. Separate official national dwelling-stock estimates are reported in Supplementary Section 2; they are not derived by scaling the ADEME certificate flow.

3.16 Use of generative AI

During manuscript preparation, the authors used OpenAI's ChatGPT to assist with editorial review, language refinement and consistency checking. The authors independently verified all suggested changes and take full responsibility for the final manuscript. The tool was not used to generate or analyse the research data.

4 Data and code availability

The England and Wales non-domestic EPC register and the ADEME DPE register for France are public open data; DESNZ metered-consumption series and Ember/Our World in Data grid-intensity series are public. All processing and analysis code, and the derived tables underlying every figure and supplementary table are available at https://github.com/Mele0/EPC; the repository will be made openly accessible to reviewers and readers.

The study used publicly available administrative data and involved no intervention with human participants.

References

[1] Department for Communities and Local Government. National Calculation Methodology (NCM) modelling guide for buildings other than dwellings in England: 2013 edition. Technical report, Department for Communities and Local Government, November 2017. URL https://www.uk-ncm.org.uk/filelibrary/NCM_Modelling_Guide_2013_Edition_20November2017.pdf.

[2] Department for Levelling Up, Housing and Communities. National Calculation Methodology (NCM) modelling guide for buildings other than dwellings in England: 2021 edition. Technical report, Department for Levelling Up, Housing and Communities, April 2022. URL https://www.uk-ncm.org.uk/filelibrary/NCM_Modelling_Guide_2021_Edition_England_14April2022.pdf.

[3] Department for Business, Energy and Industrial Strategy. Non-Domestic Private Rented Sector Minimum Energy Efficiency Standards: EPC B Future Trajectory Implementation. Technical report, Department for Business, Energy and Industrial Strategy, March 2021. URL https://assets.publishing.service.gov.uk/media/6051337ed3bf7f0453f7b992/non-domestic-prs-mees-epc-b-future-trajectory-implementation.pdf. Includes Technical Annex. Published 17 March 2021.

[4] Ministry of Housing, Communities and Local Government and Department for Energy Security and Net Zero. Technical annex for chapter 2: What EPCs measure. GOV.UK consultation outcome, March 2026. URL https://www.gov.uk/government/consultations/reforms-to-the-energy-performance-of-buildings-regime/technical-annex-for-chapter-2-what-epcs-measure. Updated 9 March 2026.

[5] Piet Eichholtz, Nils Kok, and John M. Quigley. Doing Well by Doing Good? Green Office Buildings. *American Economic Review*, 100(5):2492–2509, 2010. doi: 10.1257/aer.100.5.2492.

[6] Dirk Brounen and Nils Kok. On the Economics of Energy Labels in the Housing Market. *Journal of Environmental Economics and Management*, 62(2):166–179, 2011. doi: 10.1016/j.jeem.2010.11.006. URL https://doi.org/10.1016/j.jeem.2010.11.006.

[7] Franz Fuerst, Pat McAllister, Anupam Nanda, and Peter Wyatt. Does Energy Efficiency Matter to Home-Buyers? An Investigation of EPC Ratings and Transaction Prices in England. *Energy Economics*, 48:145–156, 2015. doi: 10.1016/j.eneco.2014.12.012. URL https://doi.org/10.1016/j.eneco.2014.12.012.

[8] Franz Fuerst, Pat McAllister, Anupam Nanda, and Peter Wyatt. Energy performance ratings and house prices in Wales: An empirical study. *Energy Policy*, 92:20–33, 2016. doi: 10.1016/j.enpol.2016.01.024.

[9] Nils Kok and Maarten G. J. Jennen. The Impact of Energy Labels and Accessibility on Office Rents. *Energy Policy*, 46:489–497, 2012. doi: 10.1016/j.enpol.2012.04.015. URL https://doi.org/10.1016/j.enpol.2012.04.015.

[10] Piet Eichholtz, Nils Kok, and Xudong Sun. The impact of minimum energy performance standards on the commercial real estate market. *Nature Communications*, 17:4038, 2026. doi: 10.1038/s41467-026-70684-w.

[11] HM Government. The Energy Efficiency (Private Rented Property) (England and Wales) Regulations 2015. Statutory Instrument 2015 No. 962, February 2015. URL https://www.le

gislation.gov.uk/uksi/2015/962/contents/made. Minimum Energy Efficiency Standards regulations for privately rented property.

[12] Department for Energy Security and Net Zero. Minimum Energy Efficiency Standards (MEES) in the non-domestic Private Rented Sector: interim response. GOV.UK consultation outcome, June 2026. URL https://www.gov.uk/government/consultations/non-domestic-private-rented-sector-minimum-energy-efficiency-standards-epc-b-implementation/outcome/minimum-energy-efficiency-standards-mees-in-the-non-domestic-private-rented-sector-interim-response. Updated 18 June 2026.

[13] Minna Sunikka-Blank and Ray Galvin. Introducing the prebound effect: the gap between performance and actual energy consumption. *Building Research & Information*, 40(3):260–273, 2012. doi: 10.1080/09613218.2012.690952.

[14] Daša Majcen, Laure C. M. Itard, and Henk Visscher. Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and policy implications. *Energy Policy*, 54:125–136, 2013. doi: 10.1016/j.enpol.2012.11.008.

[15] Pieter de Wilde. The Gap Between Predicted and Measured Energy Performance of Buildings: A Framework for Investigation. *Automation in Construction*, 41:40–49, 2014. doi: 10.1016/j.autcon.2014.02.009. URL https://doi.org/10.1016/j.autcon.2014.02.009.

[16] Chris van Dronkelaar, Mark Dowson, Emma Burman, Catalina Spataru, and Dejan Mumovic. A Review of the Energy Performance Gap and Its Underlying Causes in Non-Domestic Buildings. *Frontiers in Mechanical Engineering*, 1:17, 2016. doi: 10.3389/fmech.2015.00017. URL https://doi.org/10.3389/fmech.2015.00017.

[17] Stefano Cozza, Jonathan Chambers, Arianna Brambilla, and Martin K. Patel. Do Energy Performance Certificates Allow Reliable Predictions of Actual Energy Consumption and Savings? Learning from the Swiss National Database. *Energy and Buildings*, 224:110235, 2020. doi: 10.1016/j.enbuild.2020.110235. URL https://doi.org/10.1016/j.enbuild.2020.110235.

[18] Mingda Yuan and Ruchi Choudhary. Energy Performance Certificate Renewal: An Analysis of Reliability of Simple Non-Domestic Buildings' EPC Ratings and Pragmatic Improving Strategies in the UK. *Energy Policy*, 178:113581, 2023. doi: 10.1016/j.enpol.2023.113581. URL https://doi.org/10.1016/j.enpol.2023.113581.

[19] République française. Arrêté du 13 août 2025 modifiant le facteur de conversion de l'énergie finale en énergie primaire de l'électricité relatif au diagnostic de performance énergétique. Journal officiel de la République française, 2025. URL https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000052134589. JORFTEXT000052134589. Published 26 August 2025. Accessed 29 June 2026.

[20] Ministère de la Transition écologique. Évolution du calcul du DPE au 1er janvier 2026 : favoriser l'électrification du chauffage des logements. Press release, August 2025. URL https://www.ecologie.gouv.fr/presse/evolution-du-calcul-du-dpe-1er-janvier-2026-favoriser-lelectrification-du-chauffage. Press release announcing the reduction of the electricity primary-energy coefficient from 2.3 to 1.9, effective 1 January 2026. Accessed 29 June 2026.

[21] Ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique. Un nouveau DPE au 1er janvier 2026 pour favoriser le chauffage électrique. economie.gouv.fr, August 2025. URL https://www.economie.gouv.fr/actualites/un-nouveau-dpe-au-1er-janvier-2026-pour-favoriser-le-chauffage-electrique.

[22] Department for Business, Energy and Industrial Strategy. Non-domestic Private Rented Sector consultation Impact Assessment. Technical report, Department for Business, Energy and Industrial Strategy, October 2019. URL https://assets.publishing.service.gov.uk/media/5da5c614ed915d17b7f321a1/non-domestic-prs-consultation-ia.pdf.

[23] Direction de l'information légale et administrative. Les passoires thermiques les moins bien isolées ne peuvent plus être mises en location. Service-Public.fr, January 2025. URL https://www.service-public.gouv.fr/particuliers/actualites/A17975. Accessed 8 July 2026.

[24] ADEME. DPE Logements existants (depuis juillet 2021). Open data dataset, 2026. URL https://data.ademe.fr/datasets/dpe03existant. Dataset identifier: dpe03existant. Accessed 30 June 2026.

[25] European Parliament and Council of the European Union. Directive (EU) 2024/1275 of the European Parliament and of the Council of 24 April 2024 on the energy performance of buildings (recast). Official Journal of the European Union, L, 8 May 2024, April 2024. URL https://eur-lex.europa.eu/eli/dir/2024/1275/oj. Energy Performance of Buildings Directive Recast.

[26] European Commission. Energy Performance of Buildings Directive, 2026. URL https://energy.ec.europa.eu/topics/energy-efficiency/energy-performance-buildings/energy-performance-buildings-directive_en. Accessed 23 June 2026.

[27] Ministry of Housing, Communities and Local Government. Get energy performance of buildings data. GOV.UK service, n.d. URL https://get-energy-performance-data.communities.gov.uk/. Accessed 22 June 2026.

[28] GeoPlace. The Unique Property Reference Number (UPRN), n.d. URL https://www.geoplace.co.uk/addresses-streets/location-data/the-uprn. Accessed 23 June 2026.

[29] Department for Energy Security and Net Zero. Non-domestic National Energy Efficiency Data Framework (ND-NEED), 2024. GOV.UK official statistics, August 2024. URL https://www.gov.uk/government/statistics/non-domestic-national-energy-efficiency-data-framework-nd-need-2024.

[30] Department for Energy Security and Net Zero. Subnational electricity and gas consumption summary report 2024. GOV.UK official statistics, December 2025. URL https://www.gov.uk/government/statistics/subnational-electricity-and-gas-consumption-summary-report-2024/subnational-electricity-and-gas-consumption-summary-report-2024--2.

[31] Hannah Ritchie, Pablo Rosado, and Max Roser. Data Page: Lifecycle carbon intensity of electricity generation. Our World in Data. Data adapted from Ember, 2026. URL https://archive.ourworldindata.org/20260601-105254/grapher/carbon-intensity-electricity.html. Archived 1 June 2026.

[32] Adam Mactavish and Azita Dezfouli. Costing Energy Efficiency Improvements in Existing Commercial Buildings. Full report, Investment Property Forum, October 2017. URL https:

//www.ipf.org.uk/asset/C22367C2-84CD-4CAB-BBBABEC10B25E23F/. Research undertaken by Currie & Brown through the IPF Research Programme 2015–2018.

[33] HM Treasury. GDP Deflators at Market Prices, and Money GDP June 2026 (Quarterly National Accounts). Accredited official statistics, June 2026. URL https://www.gov.uk/government/statistics/gdp-deflators-at-market-prices-and-money-gdp-june-2026-quarterly-national-accounts. Published 30 June 2026.

[34] SDES. Le parc de logements par classe de performance énergétique au 1er janvier 2025. Données et études statistiques, Ministère de la Transition écologique, 2025. URL https://www.statistiques.developpement-durable.gouv.fr/le-parc-de-logements-par-classe-de-performance-energetique-au-1er-janvier-2025. Published 12 November 2025. Accessed 29 June 2026.

[35] Ministère de la Transition écologique. Évolutions du calcul du DPE : les réponses à vos questions. Government FAQ, 2025. URL https://www.ecologie.gouv.fr/actualites/evolutions-du-calcul-du-dpe-reponses-vos-questions. Published 15 July 2025. Accessed 29 June 2026.

1 Supplementary Material

1.1 NCM electricity factors

NCM 2013 applied fixed carbon-emission and primary-energy factors to grid electricity, whereas NCM 2021 introduced month-specific values. Table S1 reports the full schedule used after the June 2022 revision. Figure S1 illustrates the table. The arithmetic mean is shown only as a descriptive summary, since EPC calculations apply the relevant monthly factors to modelled electricity use in each month.

Table S1: Electricity carbon-emission and primary-energy factors before and after the 2022 NCM revision. NCM 2013 applied fixed factors to grid-supplied electricity, whereas NCM 2021 applies month-specific factors to modelled electricity use. The arithmetic mean is included only as a descriptive summary and is not used directly in the EPC calculation, which applies the corresponding factors to each month of electricity use.

Accounting basis	**Month**	**CO_2 factor ($kgCO_2/kWh$)**	**Primary-energy factor (kWh_{PE}/kWh)**
NCM 2013	Fixed annual factor	0.519	3.070
NCM 2021	January	0.163	1.602
	February	0.160	1.593
	March	0.153	1.568
	April	0.143	1.530
	May	0.132	1.487
	June	0.120	1.441
	July	0.111	1.410
	August	0.112	1.413
	September	0.122	1.449
	October	0.136	1.504
	November	0.151	1.558
	December	0.163	1.604
NCM 2021	Arithmetic mean	0.139	1.513

1.2 Cohort description

Table S2 reports the composition of the England and Wales non-domestic EPC register used throughout the study. It distinguishes certificates, unique UPRNs, repeat-certified UPRNs and the straddling cohort used for within-building analyses, and it summarises the fuel, sector, transaction-type and latest-band composition of the latest-certificate frame.

1.3 International structural-exposure typology

Table S3 reports the country-level coding underlying the structural-exposure map in Figure 4 combining rating-metric sensitivity, grid-decarbonisation speed and policy-threshold status to identify where comparable factor-driven reclassification is structurally possible.

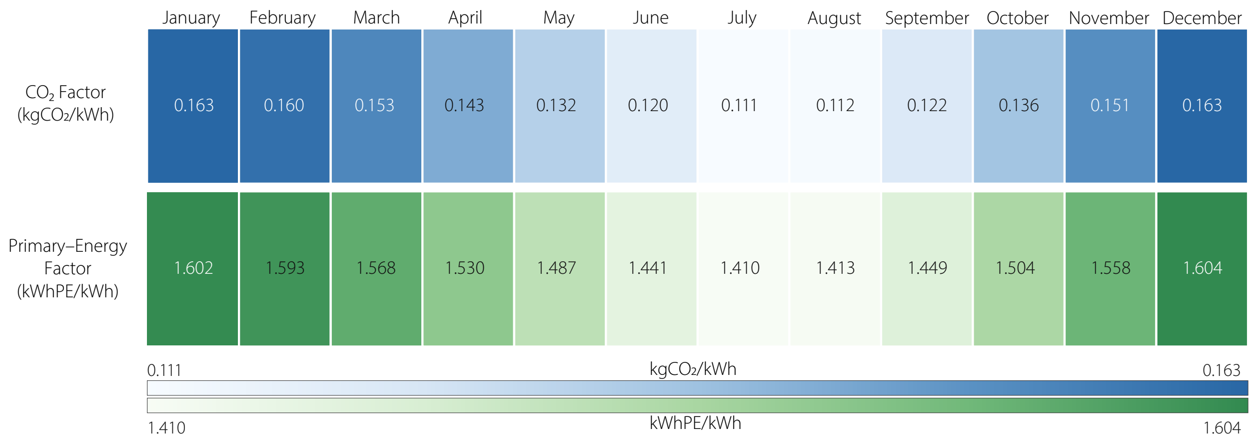

Figure S1: Monthly electricity carbon-emission and primary-energy factors under NCM 2021. Month-specific factors applied to modelled grid-electricity use are shown for carbon emissions and primary energy, with darker shading indicating higher values within each series.

1.4 Event-study specification

To test whether electrically heated UPRNs diverged from gas-heated UPRNs at the 2022 methodology change, we estimated an event-study model. We use the asset rating as the outcome because EPC policy thresholds are defined on this scale and because it is the register quantity directly affected by the accounting factors. The estimation sample contained electric- and gas-heated UPRNs with at least two valid certificates. The event study includes all repeat-certified electric- or gas-heated UPRNs with at least two valid certificates, without the three-month separation filter applied to the first-last within-building panel. As a result it includes more electric/gas UPRNs (180,264) than the 169,888 all-fuel first-last panel. The pooled difference-in-differences row is estimated on the corresponding first-last pairs (N=160,754). We therefore estimated:

$$\mathrm{AR}_{it} = \alpha_i + \sum_{y \neq 2021} \lambda_y \{\mathrm{year}_t = y\} + \sum_{y \neq 2021} \beta_y \left[\mathrm{Electric}_i \times \{\mathrm{year}_t = y\}\right] + \varepsilon_{it}. \tag{S1}$$

where α_i is a UPRN fixed effect and 2021 is the omitted reference year, β_y measures the electric minus gas asset-rating differential in year y relative to 2021, and ε_{it} is the idiosyncratic error term. We estimated the model by demeaning the outcome and regressors within UPRNs, equivalent to absorbing building fixed effects. To avoid making the year-specific divergence dependent on additional control choices, the specification intentionally used only building identity, inspection year and heating-fuel group.

1.5 Fuel-by-straddle mechanism diagnostic

We estimated a first-difference difference-in-differences model on repeat-certified pairs:

$$\Delta\mathrm{AR}_i = \beta_0 + \beta_1 \cdot \mathrm{Electric}_i + \beta_2 \cdot \mathrm{Straddle}_i + \beta_3 \cdot (\mathrm{Electric}_i \times \mathrm{Straddle}_i) + \beta_4 \cdot \mathrm{Gap}_i + \epsilon_i \tag{S2}$$

Here, $\Delta\mathrm{AR}_i$ is the first–last asset-rating change for building i, where negative values indicate improvement because lower asset ratings denote better certified performance. $\mathrm{Electric}_i$ is an indicator for UPRNs recorded as electrically heated at the first certificate, and $\mathrm{Straddle}_i$ indicates that the certificate pair crosses the 15 June 2022 methodology change. Gap_i is the inter-certificate gap, measured in years, and ϵ_i is the error term. The coefficient of interest is β_3, the interaction

Table S2: Characteristics of the England & Wales non-domestic EPC register, 2012-2025. Repeat-certified cohort comprises UPRNs with two or more valid certificates lodged more than three months apart, the basis for all within-building analyses. Counts are at the certificate level unless stated. Band distribution is computed on the latest certificate per building. Percentages are of the relevant denominator. Transaction type identifies new-build versus existing stock.

Panel	Category	Counts	%
Scope	Certificates (2012–2025)	1,245,956	100
	Unique UPRNs	842,673	67.63
	Single-certified UPRNs	672,785	79.8
	Repeat-certified UPRNs	169,888	20.2
	Of which straddle the June-2022 cut	106,228	62.5
	Of which both certs pre-2022	57,892	34.1
	Of which both certs post-2022	5,768	3.4
Heating fuel	Electric	696,006	55.9
	Gas	462,297	37.1
	Other	87,653	7.0
Air conditioning	Present	252,374	20.3
Transaction type	New build	60,497	4.9
	Existing building	1,185,459	95.1
Sector	Retail/Financial and Professional Services	428,958	34.4
	Offices and Workshop Businesses	332,102	26.7
	Restaurants and Cafes/Drinking Establishments	146,429	11.8
	Industrial / storage / distribution	181,533	14.6
	Hotels	20,815	1.7
	Education	26,771	2.1
	Hospitals or Care homes	44,419	3.6
	Assembly, leisure, night clubs, theatres	22,599	1.8
	Community / day centres	15,890	1.3
	Universities / colleges / residential schools	11,633	0.9
	Residential spaces / other residential institutions	7,440	0.6
	Libraries, museums, galleries	2,711	0.2
	Courts, emergency services, terminals, other	4,656	0.4
Asset-rating band	A/B	188,138	22.3
	C	283,805	33.7
	D	224,884	26.7
	E	112,457	13.3
	F/G	33,389	4.0

between Electric_i and Straddle_i. It estimates the additional asset-rating change for electrically heated UPRNs, relative to gas-heated UPRNs, when the certificate pair crosses the accounting cut, after adjusting for the main electric-heating difference, the average straddling-period difference and the time between certificates. Standard errors were clustered by UPRN. Table S4 reports the year-specific electric–gas interactions and the pooled fuel-by-straddle estimate used as mechanism evidence in the main text.

Table S3: International structural-exposure typology: per-country coding, grid decarbonisation, exposure index, sources and confidence. Metric-sensitivity score: delivered/final energy = 0, primary energy = 1, dual or CO_2-reported = 1.5, carbon-weighted = 2, not classified = missing. Grid decline is the percentage reduction in electricity-generation carbon intensity 2015–2023 (Ember via Our World in Data). EU-27 members default to the EPBD primary-energy headline with planned minimum standards (Directive (EU) 2024/1275); non-EU/EEA and Energy-Community entries are coded by framework obligation at lower confidence, and verified national exceptions are coded individually.

Country	Metric class	Sens.	Grid decl. (%)	Threshold	Exposure	Conf.	Source
United Kingdom	Carbon-weighted	2	40.8	Current	81.6	High	This study; NCM/MEES
Spain	Dual / CO_2	1.5	48.3	Planned	72.4	Medium	RD 390/2021; BPIE
Portugal	Primary energy	1	61.7	Planned	61.7	Medium	EPBD 2024/1275; BPIE
Lithuania	Primary energy	1	59.8	Planned	59.8	Medium	EPBD 2024/1275; BPIE
Finland	Primary energy	1	55.4	Planned	55.4	Medium	EPBD 2024/1275; BPIE
Luxembourg	Primary energy	1	55.1	Planned	55.1	Medium	EPBD 2024/1275; BPIE
Ireland	Dual / CO_2	1.5	35.7	Planned	53.6	Medium	SEAI DEAP
Denmark	Primary energy	1	51.6	Planned	51.6	Medium	EPBD 2024/1275; BPIE
Netherlands	Primary energy	1	51.4	Current	51.4	Medium	NTA 8800
Latvia	Primary energy	1	45.9	Planned	45.9	Medium	EPBD 2024/1275; BPIE
Greece	Primary energy	1	44.3	Planned	44.3	Medium	KENAK; EPBD
Belgium	Primary energy	1	37.5	Planned	37.5	Medium	EPBD 2024/1275; BPIE
Slovakia	Primary energy	1	35.1	Planned	35.1	Medium	EPBD 2024/1275; BPIE
Austria	Primary energy	1	35.1	Planned	35.1	Medium	EPBD 2024/1275; BPIE
Hungary	Primary energy	1	34.4	Planned	34.4	Medium	EPBD 2024/1275; BPIE
Romania	Primary energy	1	34.1	Planned	34.1	Medium	EPBD 2024/1275; BPIE
EU-27 (ref.)	Primary energy	1	32.3	Planned	32.3	Medium	EPBD 2024/1275
Bulgaria	Primary energy	1	31.6	Planned	31.6	Medium	EPBD 2024/1275; BPIE
Estonia	Primary energy	1	31.0	Planned	31.0	Medium	EPBD 2024/1275; BPIE
Croatia	Primary energy	1	29.1	Planned	29.1	Medium	EPBD 2024/1275; BPIE
Slovenia	Primary energy	1	29.0	Planned	29.0	Medium	EPBD 2024/1275; BPIE
Germany	Primary energy	1	27.9	Planned	27.9	Medium	EPBD 2024/1275; BPIE
Montenegro	Primary energy	1	21.6	None	21.6	Low	Energy Comm. (EPBD)
Poland	Primary energy	1	21.2	Planned	21.2	Medium	EPBD 2024/1275; BPIE
Czechia	Primary energy	1	20.0	Planned	20.0	Medium	EPBD 2024/1275; BPIE
Italy	Primary energy	1	18.9	Planned	18.9	Medium	EPBD 2024/1275; BPIE
Malta	Primary energy	1	18.8	Planned	18.8	Medium	EPBD 2024/1275; BPIE
France	Dual / CO_2	1.5	10.0	Current	15.0	High	This study; arrêtés DPE
Serbia	Primary energy	1	14.2	None	14.2	Low	Energy Comm. (EPBD)
Bosnia and Herzegovina	Primary energy	1	13.2	None	13.2	Low	Energy Comm. (EPBD)
Cyprus	Primary energy	1	11.7	Planned	11.7	Medium	EPBD 2024/1275; BPIE
Kosovo	Primary energy	1	10.5	None	10.5	Low	Energy Comm. (EPBD)
North Macedonia	Primary energy	1	10.0	None	10.0	Low	Energy Comm. (EPBD)
Sweden	Primary energy	1	6.1	Planned	6.1	Medium	EPBD 2024/1275; BPIE
Moldova	Primary energy	1	5.8	None	5.8	Low	Energy Comm. (EPBD)
Switzerland	Dual / CO_2	1.5	2.2	None	3.3	Medium	GEAK
Iceland	Primary energy	1	-1.1	Planned	0.0	Low	EEA; EPBD
Norway	Delivered energy	0	9.2	None	0.0	High	NVE
Albania	Primary energy	1	-2.5	None	0.0	Low	Energy Comm. (EPBD)
Liechtenstein	Primary energy	1	–	Planned	–	Low	EEA; EPBD
Andorra	Not classified	–	–	Unknown	–	Low	–
Ukraine	Primary energy	1	–	None	–	Low	Energy Comm. (EPBD)

1.6 Mechanism diagnostics

We used three mechanism diagnostics to test whether the post-2022 rating change followed the structure expected from carbon-factor re-pricing. First, we compared the expected electricity re-pricing ratio, $\frac{EF_{\text{elec,new}}}{EF_{\text{elec,old}}} = \frac{0.136}{0.519}$, with the observed within-building modelled-emissions ratio $\text{BER}_{\text{post}}/\text{BER}_{\text{pre}}$ among no-recorded-works straddlers, to test whether UPRNs most exposed to electricity accounting changed by approximately the magnitude implied by the electricity carbon-

Table S4: Within-building event-study estimates of the electric–gas asset-rating differential. Coefficients are year-by-electric interactions from a building-fixed-effect event-study model with 2021 omitted. Year rows therefore measure the change in the electric-gas asset-rating differential relative to the 2021 baseline, not the absolute electric-gas gap in each year. The pooled row is a separate first-difference fuel-by-straddle diagnostic. Negative post-2022 coefficients indicate that electrically heated UPRNs improved relative to their 2021 electric-gas baseline. Because pre-2021 coefficients evolve, the event study is interpreted as supporting mechanism evidence rather than as a standalone parallel-trends causal design.

Year	Electric-Gas differential (pts) [95% CI]	p-value
Pooled DiD	−8.58 [−9.40, −7.76]	< 0.001
2012	17.56 [16.66, 18.46]	< 0.001
2013	15.25 [14.27, 16.24]	< 0.001
2014	12.95 [11.82, 14.07]	< 0.001
2015	13.97 [12.85, 15.09]	< 0.001
2016	14.55 [13.25, 15.86]	< 0.001
2017	7.84 [6.68, 9.01]	< 0.001
2018	4.39 [3.24, 5.54]	< 0.001
2019	0.10 [−0.96, 1.16]	0.848
2020	−1.19 [−2.34, −0.05]	0.0416
2021	*ref*	*ref*
2022	−14.66 [−15.59, −13.72]	< 0.001
2023	−21.65 [−22.39, −20.91]	< 0.001
2024	−21.85 [−22.67, −21.04]	< 0.001
2025	−22.71 [−23.53, −21.89]	< 0.001

factor change. Here 0.136 is the representative October monthly NCM 2021 factor used as an illustrative benchmark for this diagnostic; the headline constant-accounting counterfactual does not depend on any single monthly value, because it rescales post-2022 ratings by the per-fuel multiplier m_{fuel} estimated empirically from the no-recorded-works calibration sample. To exclude invalid denominators and extreme ratio values that would otherwise dominate ratio-based summaries, ratios were restricted to the $0 <$ BER Ratio < 5 interval. Second, we examined electricity exposure among straddling gas and electric UPRNs by assigning UPRNs to four categories using main heating fuel and air-conditioning status, measured at the first certificate: gas without air conditioning, gas with air conditioning, electric without air conditioning and electric with air conditioning. The primary contrast is between heating fuels, tested within each air-conditioning stratum; air conditioning is treated as a secondary modifier and is not assumed to produce a strictly monotonic ordering. We then examined redistribution around the EPC-B threshold. For straddling UPRNs, we defined a building as moving into EPC-B if its asset rating exceeded 50 at the first certificate and was 50 or below at the last certificate. We then compared the electric share among UPRNs crossing into EPC-B with the electric share among all straddling UPRNs to test whether apparent compliance gains around the policy threshold were concentrated in UPRNs with greater exposure to the electricity carbon-factor reduction.

1.7 Robustness and heterogeneity of within-UPRN jumps.

Table S5 summarises the main robustness and heterogeneity checks for the within-UPRN asset-rating jump. The purpose is to show that the post-2022 discontinuity is not an artefact of transaction type, major sector composition, the no-recorded-works calibration sample, or ordinary temporal drift in repeated certification. The placebo comparisons are especially important because both-certificate pre-cut and both-certificate post-cut pairs remain under a common accounting regime, whereas straddling pairs cross the methodology revision.

Table S5: Robustness and heterogeneity of the 2022 within-UPRN asset-rating jump. Mean change in EPC asset rating for repeat-certified UPRNs, estimated from first–last certificate pairs. Composition block restricts straddling cohort by transaction type, sector and a no-recorded-works calibration sample. Temporal-placebo block compares the straddling cohort with pairs whose two certificates were both before or both after the June-2022 methodology cut. Electric-gas differential remains large and negative across existing UPRNs, major sectors and UPRNs with no recorded works. Straddle denotes cut between certificates. No recorded works represent UPRNs with Δrec $= 0$ and Δarea $< 2\%$. Δ Elec-Gas Diff. vs. 2021 (ARP) denotes the change in the electric-gas differential relative to its 2021 value, measured in asset-rating points (ARP).

Block	**Stratum**	*All*		*Electric*		*Gas*		
		Count	**Mean ΔAR**	**Count**	**Mean ΔAR**	**Count**	**Mean ΔAR**	**Δ Elec-Gas Diff. vs. 2021 (ARP)**
Composition	All straddlers	106,228	−30.9	56,247	−41.3	44,372	−18.3	−23.0
	Existing UPRNs	104,128	−30.7	55,307	−41.3	43,373	−17.8	−23.5
	New UPRNs	2,100	−39.9	940	−39.8	999	−39.6	−0.2
	Property to let	62,852	−32.8	35,531	−42.3	23,989	−19.2	−23.1
	Marketed/non-marketed sale	33,356	−26.7	15,799	−39.1	15,899	−15.1	−23.9
	Retail/financial services	38,197	−32.8	28,717	−38.3	8,689	−16.0	−22.4
	Offices and workshops	27,770	−34.8	12,595	−47.8	13,252	−22.6	−25.2
	Restaurants/cafes	13,699	−31.2	5,323	−48.6	7,746	−20.1	−28.5
	Industrial	17,025	−26.1	7,199	−35.7	8,133	−17.5	−18.2
	No recorded works	2,443	−28.7	1,578	−35.2	746	−15.3	−20.0
Temporal placebo	Straddle 2022	106,228	−30.9	56,247	−41.3	44,372	−18.3	−23.0
	Both certificates pre-2022	57,892	−26.3	31,792	−34.2	22,955	−16.2	−18.0
	Both certificates post-2022	5,768	−15.1	3,339	−9.0	2,049	−24.3	15.2

1.8 Additive no-works windfall estimator

As an additive robustness specification, we estimated a fuel-specific rating windfall from the no-recorded-works calibration sample. For each fuel group, the windfall was defined as the negative of the mean change in asset rating among UPRNs with no recorded works, so that reductions in asset rating corresponded to positive rating-point windfalls. This estimator therefore measures the average rating points gained by each fuel group among straddling UPRNs with no recorded works. We used this additive estimator only as a robustness check, reported alongside the headline multiplicative fixed-factor results for the national policy estimates; the headline Markov stress test uses the fixed-factor-corrected starting stock.

1.9 Robustness to multiplier specification

The headline counterfactual transports a single per-fuel multiplier to heterogeneous UPRNs. To assess whether this drives the result, we re-estimated the reclassification under progressively more granular multiplier definitions: fuel crossed with sector, floor-area bin and air-conditioning status, with sparse cells (fewer than 50 no-recorded-works calibration observations) collapsed to the fuel-level multiplier. We also varied the no-recorded-works floor-area tolerance (1%, 2%, 5%), estimated the multiplier by repeated out-of-sample cross-fitting (multipliers estimated on a random half of the no-works straddlers and applied to the other half, over 200 random splits, reporting the median and 2.5–97.5 percentile range), and report the additive-windfall estimator. Table S6 shows the result is stable across every specification that stratifies on UPRN attributes uncorrelated with the crossing: the share of transition crossings that dissolve under constant accounting stays between 65% and 69% and the latest-stock counts within a few per cent of the headline. We therefore treat the range across these attribute-stratified specifications, not the bootstrap interval, as the measure of structural sensitivity.

Stratification by pre-revision rating band is reported separately as a bad-control diagnostic rather than as a sensitivity check. Conditioning on the pre-revision band collapses the estimate to 26% (33,587 reaching EPC-B), but this is an artefact rather than a genuine sensitivity: the multiplier AR_{pre}/AR_{post} is mechanically correlated with the pre-revision rating, so band-specific multipliers regress toward the mean and absorb precisely the accounting variation the counterfactual is designed to isolate. It is therefore a control that lies on the causal path and should not be used to bound the estimate.

1.10 Interpretation of the fuel-specific multiplier components

The fuel-specific Building Emission Rate ratios, Standard Emission Rate ratios and asset-rating multipliers are reported in main-text Table 1. As $AR = 50\,BER/SER$, the building-level asset-rating multiplier equals the reference-building emission-rate change divided by the assessed-building emission-rate change. The reported table entries are separate fuel-specific medians.

The components clarify why gas-heated UPRNs show a substantial modelled-emissions decline but a more modest asset-rating multiplier. Nominally gas-heated UPRNs retain grid-electricity loads for lighting, ventilation, cooling and auxiliary services, while the reference-UPRN SER also falls across the revision. Electrically heated UPRNs receive the larger multiplier as their modelled emissions are more directly exposed to the electricity carbon-factor revision.

1.11 Local discontinuity around the June 2022 cut

The main within-UPRN comparison uses first and last certificates, which can span several years and could in principle blend the accounting shock with unrelated retrofit or renewal. As a supportive local check, Table S7 restricts the straddling cohort to UPRNs whose two certificates both fall within a symmetric window of ±3, ±6, ±12 and ±24 months of 15 June 2022 (by inspection date). The electric–gas differential remains large and negative in every window, including the tightest ±3-month window (−22.3 points). Because very few UPRNs are recertified within a few months on both sides of the cut, the narrowest windows contain small samples (41 UPRNs at ±3 months) and should be read as a supportive local-window check on the direction and magnitude of the step rather than as a powered regression-discontinuity or month-by-month event study; the differential is nonetheless consistent across all windows and stable in the larger ±12- and ±24-month samples.

Table S6: Robustness of constant-accounting threshold reclassification to multiplier specification. Each row re-estimates the fuel-specific re-pricing multiplier under a different definition and recomputes the fraction of the 37,357 within-UPRN EPC-B crossings that would not be EPC-B under constant accounting, together with the latest-stock counts of UPRNs reaching EPC-B and clearing the legal floor on paper. Cells with fewer than 50 no-recorded-works calibration observations are collapsed to the fuel-level multiplier. The cross-fitted row reports the median and 2.5–97.5 percentile range over 200 random half-sample splits. The estimate is stable across attribute-stratified specifications. The additive-windfall estimator over-corrects deep-A UPRNs and is reported as an alternative. Stratification by pre-revision rating band is a bad control (the multiplier is mechanically correlated with the pre-revision rating) and is discussed in the text rather than tabulated here.

Multiplier specification	**Crossings dissolving (%)**	**Reached (n)**	**Cleared legal floor (n)**
Baseline (fuel only)	67.9	80,565	51,342
Fuel × sector	67.1	81,170	52,399
Fuel × floor-area bin	67.9	80,832	51,886
Fuel × air-conditioning status	67.4	80,623	49,319
No-works tolerance $\lvert\Delta\text{FA}\rvert < 1\%$	69.4	81,756	52,920
No-works tolerance $\lvert\Delta\text{FA}\rvert < 5\%$	67.9	80,464	52,034
Cross-fitted (200 splits, median)	66.8	79,896	50,925
Cross-fitted (200 splits, [95%])	[64.4, 69.6]	[76,883, 82,131]	[46,660, 54,830]
Additive windfall (alternative)	88.5	108,460	29,865

Table S7: Local discontinuity: within-UPRN asset-rating change for straddlers with both certificates near the cut. Straddling UPRNs are restricted to pairs whose first and last certificates both fall within the stated symmetric window of 15 June 2022 (inspection date). Mean ΔAR (negative = improvement) by heating fuel, with counts in parentheses.

Window	**Counts**	**Electric ΔAR**	**Gas ΔAR**	**Elec−Gas**
±3 months	41	−27.4 (21)	−5.1 (17)	−22.3
±6 months	316	−23.5 (193)	−7.1 (108)	−16.5
±12 months	1,495	−26.8 (909)	−13.6 (474)	−13.3
±24 months	5,639	−29.7 (3,308)	−16.0 (1,924)	−13.7

1.12 Observed-to-constant band transition and rounding sensitivity

Table S8 reports the full seven-band (A–G) transition from observed to constant-accounting ratings on the latest-certificate frame, using the SBEM band cut-points (A $\leq$ 25, B $\leq$ 50, C $\leq$ 75, D $\leq$ 100, E $\leq$ 125, F $\leq$ 150, G $>$ 150). The matrix is monotone: holding carbon accounting constant can only move a UPRN to a worse band, and the reclassification is spread across the distribution rather than concentrated at a single boundary, with substantial A→B, B→C, B→D and E→F/G movement.

Because the register stores asset ratings rounded to the nearest integer, we checked threshold sensitivity at the EPC-B (AR = 50) and legal-floor (AR = 125) cut-points. Only 4.1% of the 80,565 EPC-B reclassifications and 4.5% of the 51,342 legal-floor reclassifications have a constant-accounting rating within one point of the threshold, and re-rounding the constant-accounting rating to the nearest integer changes the EPC-B count by less than 1.2% (to 79,573) and leaves the legal-floor count unchanged. The headline counts are therefore not driven by UPRNs sitting on the rounding boundary.

Table S8: Observed and constant-accounting EPC band transition matrix for the latest-stock sample. Rows denote the observed EPC band under the current carbon-accounting methodology, and columns denote the corresponding band under the constant-accounting counterfactual. Entries are UPRN counts. The transition matrix is lower triangular because holding carbon accounting constant cannot improve a building's EPC band relative to the observed rating.

Obs.\Const.	**A**	**B**	**C**	**D**	**E**	**F**	**G**
A	18,733	12,693	0	0	0	0	0
B	0	76,147	61,623	18,942	0	0	0
C	0	0	186,561	59,529	37,715	0	0
D	0	0	0	176,473	19,922	20,672	7,817
E	0	0	0	0	89,604	12,354	10,499
F	0	0	0	0	0	14,641	1,399
G	0	0	0	0	0	0	17,349

1.13 Robustness to certificate validity

The latest-certificate frame retains the most recent certificate per UPRN regardless of its age, and EPCs are valid for ten years, so a reviewer could object that some "latest" certificates are expired and the current-stock framing is unsafe. Table S9 repeats the headline national-policy quantities under alternative validity cut-offs. The below-EPC-B and F/G shares shift with the denominator, but the observed-to-constant gap persists in every subset, including certificates lodged within the last ten years and post-2022 certificates only. The on-paper crossing counts are unchanged across the first three rows because a UPRN can only cross under constant accounting if its latest certificate is post-2022. As a result, each crossing UPRN is therefore recent and valid by construction, so the validity concern cannot reduce the artefact. Restricting further to UPRNs above 1,000 m^2 with a valid certificate leaves 10,273 EPC-B crossings (£3.91 billion).

1.14 England-only and England–Wales split

The analysis applies a single 15 June 2022 methodology cut-off to the England and Wales register, but the National Calculation Methodology is administered separately in the two nations and Welsh notices carry their own version and timing details, so a reviewer could question whether the discontinuity applies identically in Wales. Table S10 recomputes the headline quantities for England only and Wales only (classified from the local-authority ONS code), and re-estimates the within-UPRN electric–gas straddler contrast and the no-recorded-works multiplier separately by nation. England, which is 95% of the register, reproduces the full-sample result almost exactly. Wales shows the same direction and a nearly identical electric–gas contrast (-22.5 versus -23.0 points) and no-works multipliers (electric ≈ 1.63–1.65, gas ≈ 1.16–1.20 in both), so the single-cut-off assumption is not driving the result.

Table S9: Robustness of the national-policy quantities to certificate validity. Latest-certificate frame under alternative validity cut-offs. Below-B and F/G are observed → constant-accounting shares; reached-B and cleared-floor are on-paper crossing counts with capex-equivalent at £100/m^2. Validity uses the lodgement date against a ten-year window to a 15 June 2026 as-of date.

Subset	**Counts**	**Below-B (%)**	**F/G (%)**	**Reached EPC-B (n; £bn)**	**Cleared floor (n; £bn)**
All latest certificates	842,673	77.7 → 87.2	3.96 → 10.1	80,565 (£5.50)	51,342 (£1.77)
Lodged within 10 years	683,451	74.6 → 86.4	1.80 → 9.3	80,565 (£5.50)	51,342 (£1.77)
Post-2022 latest only	319,013	62.7 → 88.0	0.85 → 16.9	80,565 (£5.50)	51,342 (£1.77)
>1,000 m^2 and within 10 years	77,887	57.3 → 70.5	1.05 → 4.7	10,273 (£3.91)	2,863 (£0.80)

Table S10: England-only and Wales-only robustness. National-policy quantities (as Table S9) and the within-UPRN electric–gas straddler contrast and no-recorded-works multiplier, estimated separately by nation. Nation is classified from the local-authority ONS code. Note that country split excludes 168 latest-stock UPRNs with missing or unclassified local-authority country codes.

Nation	**Counts (stock)**	**Below-B (%)**	**Reached-B (Counts)**	**Cleared floor (Counts)**	**Straddler Counts**	**Elec−Gas ΔAR**	**No-works mult. (E / G)**
England	801,118	77.5 → 87.2	77,525	48,260	101,441	−23.0	1.631 / 1.197
Wales	41,387	81.2 → 88.5	3,014	3,078	4,782	−22.5	1.646 / 1.159

1.15 Stricter no-recorded-works calibration

Headline no-recorded works sample requires an unchanged recommendation count together with a floor area change below 2%. Since recommendation count alone provides only a weak proxy for the absence of physical changes, Table S11 progressively strengthens this definition by additionally requiring unchanged main heating fuel and air conditioning status, an unchanged short, medium, and long payback mix, and finally an identical set of recommendation codes across both certificates. The electric multiplier remains stable throughout and declines only from 1.63 to 1.51 under the strictest definition indicating the electric repricing result is not driven by the relatively weak recommendation count proxy. Despite gas multiplier remaining close to one, calibration cell becomes sparse under the strictest definition because only $n = 17$ observations remain. As a result, exact stability in recommendation content does not provide a sufficiently large sample to estimate the gas multiplier precisely. We therefore retain the baseline definition for the headline counterfactual.

1.16 Transition model

We modelled future rating movement using a five-state transition matrix over the band set $\mathcal{B} = \{A/B,\ C,\ D,\ E,\ F/G\}$. For each specification, we first counted observed consecutive certificate transitions from previous band a to current band b, giving cell counts N_{ab}. We then applied Laplace, or add-one, smoothing before row-normalisation. Specifically, with $K = 5$ rating states, the transition probability from band a to band b was estimated as

$$T_{ab} = \frac{N_{ab} + 1}{\sum_{b' \in \mathcal{B}} N_{ab'} + K}. \tag{S3}$$

Equivalently, each cell of the count matrix is initialized with pseudo-counts before row-normalisation, which prevents zero transition probabilities in sparse rows. This full-band five-state construction

Table S11: Fuel-specific multiplier under progressively stricter no-recorded-works definitions. Median ARpre/ARpost is reported separately for each primary heating fuel. Values in parentheses denote the calibration sample size (n). Successive rows apply increasingly restrictive definitions of the no-recorded-works sample, beginning with the baseline criterion (Δrec $= 0$, ΔFA $< 2\%$) and additionally requiring unchanged heating fuel and air-conditioning status, unchanged payback composition, and finally an identical recommendation-code set between assessments. Note that counts represents the no-recorded-works cohort with bracketed fuel counts representing UPRNs entering each fuel-specific multiplier. In the baseline row the cohort total exceeds the fuel-column sum by one (2,443 versus $1{,}578 + 746 + 118 = 2{,}442$) because one UPRN records a non-positive Asset Rating on its later certificate and is removed by the positive-rating filter. All fuel-specific and threshold-model analyses therefore use 2,442 UPRNs.

No-works definition	Counts	Electric	Gas	Other
Baseline (Δrec$= 0$, ΔFA$< 2\%$)	2,443	1.633 (1,578)	1.197 (746)	1.384 (118)
+ *unchanged fuel and air-conditioning*	1,696	1.584 (1,204)	1.086 (427)	1.489 (65)
+ *unchanged payback mix*	630	1.551 (524)	1.141 (83)	1.489 (23)
+ *identical recommendation-code set*	271	1.511 (235)	1.020 (17)	1.490 (19)

is retained only as a Supplementary sensitivity (Table S18). The preferred specification instead works at the threshold level, with the two states below and at-or-above a threshold (EPC-B at AR $= 50$; F/G at AR $= 125$), because the fixed-accounting bridge does not support exact individual old-basis ratings. Each certificate carries a below-threshold probability with hard (0/1) for ratings already on the relevant basis, and, for post-2022 ratings expressed on the pre-2022 basis, the expected old-basis threshold state $\Pr(\mathrm{AR} \times r > \tau)$ obtained from the empirical distribution of no-works re-pricing ratios $r = \mathrm{AR}_{\mathrm{pre}}/\mathrm{AR}_{\mathrm{post}}$ within fuel. For a reassessment pair with previous and current below-threshold probabilities p_0 and p_1, the expected transition mass is $(1 - p_0)(1 - p_1)$, $(1 - p_0)p_1$, $p_0(1 - p_1)$ and $p_0 p_1$ across the four above/below cells; these are summed over pairs and row-normalised with Laplace smoothing. Let s_0 denote the starting below-threshold share and T the two-state threshold transition matrix. For a cumulative reassessment-throughput scenario p, the projected below-threshold share is

$$s(p;\ T) = (1 - p)\, s_0 + p\,(s_0 T_{\text{below}\to\text{below}} + (1 - s_0) T_{\text{above}\to\text{below}}), \tag{S4}$$

so a share $1-p$ of the stock keeps its current threshold state while a share p undergoes one transition. We evaluated $p \in \{0.10,\ 0.20,\ 0.35,\ 0.50,\ 0.75,\ 1.00\}$, with the observed one-year reassessment coverage is approximately 0.157. Sensitivity assumes time-homogeneous transition probabilities, at most one reassessment step per building, and exogenous reassessment throughput. We validate it in three ways: first, a back-test that estimates the transition matrix on an earlier training period and checks whether it reproduces the later observed threshold share within a single accounting basis (Table S14); next, a transition-pair bootstrap that places a structural interval on the gap (Table S15); and third with a transition-sample robustness sweep across pair-selection rules, gap caps, fuels and sectors (Table S16). Matrix diagnostics confirming that each scenario is internally commensurate are in Table S17.

Fuel- and air-conditioning-exposure diagnostic are reported in Table S19, presenting the mean within-UPRN asset-rating change for straddling repeat-certified UPRNs by primary heating fuel and air-conditioning status at the first certificate. Electrically heated UPRNs exhibit substantially larger rating improvements than gas-heated UPRNs within both air-conditioning strata, indicating that the

Table S12: Preferred threshold-level, expected-state reassessment-throughput sensitivity, with commensurate accounting bases. Projected below-threshold share of the latest-certificate frame as the cumulative reassessed share p rises, for two internally commensurate scenarios (Table S13) at the EPC-B and F/G thresholds. The observed / current-accounting scenario uses the observed starting share and a transition matrix from observed post-2022 reassessments; the constant-accounting scenario expresses every starting and transition rating on the pre-2022 basis, mapping post-2022 ratings to expected old-basis threshold states through the empirical no-works re-pricing ratio distribution. States are below versus at-or-above the threshold; no hard full-band classification is used. The gap persists and widens across the throughput range. Note that it should be interpreted as illustrative only and not as a forecast of legal compliance, national stock, or physical retrofit.

	EPC-B threshold			**F/G threshold**		
Coverage p	**Obs. (%)**	**Const. (%)**	**Gap (pp)**	**Obs. (%)**	**Const. (%)**	**Gap (pp)**
0.10	76.1	84.4	8.3	3.7	13.4	9.7
0.20	74.5	83.8	9.3	3.5	13.5	10.0
0.35	72.2	83.0	10.9	3.2	13.7	10.5
0.50	69.8	82.3	12.5	2.8	13.8	11.0
0.75	65.9	81.0	15.1	2.3	14.0	11.8
1.00	61.9	79.6	17.7	1.7	14.3	12.6

Table S13: Accounting basis of each Markov scenario. In each scenario the starting distribution and the transition matrix are on the same accounting basis, so band states are commensurate. The earlier version, which paired a constant-accounting starting distribution with an observed-basis transition matrix, is not used.

Scenario	**Starting distribution**	**Transition matrix**
Observed/Current-accounting	Observed ratings	Observed post-2022 reassessments
Constant-accounting	Pre-2022 basis (expected state)	Pre-2022 basis (expected-state, all pairs)

Table S14: Markov back-test. A threshold transition matrix estimated on an earlier training period is used to predict the later observed below-threshold share among reassessed UPRNs, within a single accounting basis using a median split of the pairs by current-certificate date and without accounting break inside a split. Reproduction is strong at the F/G threshold on the post-2022 basis and weaker at the EPC-B threshold, consistent with non-stationary transition dynamics.

Basis	**Threshold**	**Counts train**	**Counts test**	**Predicted (%)**	**Observed (%)**
Pre-2022	*EPC-B*	68,280	68,422	85.18	80.21
	F/G	68,280	68,422	10.18	2.03
Post-2022	*EPC-B*	20,498	20,510	56.38	49.35
	F/G	20,498	20,510	1.66	1.53

electric–gas contrast is not explained by differences in air-conditioning prevalence. Air conditioning does not introduce a monotonic shift in the estimated discontinuity as among gas-heated UPRNs,

Table S15: Structural uncertainty of the below-threshold gap. Transition-pair bootstrap of the preferred threshold-level expected-state model. The 95% interval for the below-EPC-B gap excludes zero across the entire coverage grid, so the separation between the observed-basis and constant-accounting trajectories is not an artefact of transition-sample noise.

Threshold	**Coverage** p	**Gap (pp)**	**95% low**	**95% high**
EPC-B	0.10	8.3	8.2	8.3
	0.50	12.5	12.2	12.7
	1.00	17.7	17.3	18.2
F/G	0.10	9.7	9.7	9.7
	0.50	11.0	10.9	11.0
	1.00	12.6	12.4	12.7

Table S16: Transition-sample robustness of the below-threshold gap. The preferred threshold-level expected-state gap at 10% and 100% cumulative reassessment coverage, recomputed under alternative transition-pair samples. A positive gap persists in every specification and at both thresholds, so the conclusion does not depend on any single pair-selection rule.

Transition sample	**Threshold**	n **pairs**	**Gap** p=0.10 **(pp)**	**Gap** p=1.00 **(pp)**	**Gap persists**
All repeat pairs	EPC-B	284,000	8.3	17.7	yes
Valid ratings only ($0 < \text{AR} \leq 200$)	EPC-B	278,072	8.2	17.5	yes
Excluding construction transactions	EPC-B	271,214	8.5	20.1	yes
Nearest reassessment pairs only	EPC-B	196,766	8.3	18.1	yes
Post-2022 pairs only	EPC-B	41,008	7.9	14.5	yes
Max gap					
$\leq$ 24 months	EPC-B	125,396	8.4	19.2	yes
$\leq$ 36 months	EPC-B	143,295	8.4	18.9	yes
$\leq$ 60 months	EPC-B	173,998	8.3	18.5	yes
Fuel-specific					
Electric	EPC-B	149,784	8.7	22.5	yes
Gas	EPC-B	113,144	7.8	13.1	yes
Other	EPC-B	21,072	7.8	12.9	yes
Sector					
Retail/Financial services	EPC-B	88,300	8.3	18.3	yes
Offices/Workshops	EPC-B	77,993	8.5	20.7	yes
Restaurants/Cafes/Drinking	EPC-B	33,021	8.1	16.6	yes
Storage/Distribution	EPC-B	16,268	8.1	16.0	yes

those with air conditioning experience a larger mean improvement than those without (−22.4 versus −17.3 points), whereas the opposite pattern is observed for electrically heated UPRNs (−37.2 versus −42.6 points). As a result, air conditioning acts as a secondary modifier of electricity exposure, while the dominant source of the discontinuity is the carbon-factor re-pricing associated with electric heating.

Table S17: Markov matrix diagnostics. Every transition matrix in the preferred threshold-level specification, with its state space, the accounting basis of the starting distribution and of the previous/current transition ratings, the number of pairs, row sums, non-empty cells, the smallest expected cell mass, and the smoothing rule. Within each scenario the starting distribution and transition matrix share one basis, so the states are commensurate.

Threshold	Scenario	Basis (start / transitions)	n pairs	Row sums	Min cell	Start below (%)
EPC-B	Observed	Observed / observed post-2022	41,008	1.0	4,411	77.7
	Constant-accounting	Pre-2022 / pre-2022 expected	284,000	1.0	17,743	84.9
F/G	Observed	Observed / observed post-2022	41,008	1.0	162	4.0
	Constant-accounting	Pre-2022 / pre-2022 expected	284,000	1.0	11,175	13.3

Table S18: Supplementary hard full-band sensitivity (superseded by the preferred version). The below-EPC-B gap under the hard five-band transition construction, in which post-2022 ratings are converted with the point fuel multiplier rather than mapped to expected threshold states.

Coverage p	Observed-basis (%)	Constant-accounting (%)	Gap (pp)
0.10	76.2	86.8	10.6
0.20	74.7	86.4	11.7
0.35	72.4	85.7	13.3
0.50	70.2	85.1	14.9
0.75	66.4	84.0	17.6
1.00	62.6	83.0	20.3

1.17 Capital-expenditure equivalent

To express the scale of accounting-driven threshold reclassification in investment units, we define a capital expenditure equivalent. This measure is not an estimate of realised retrofit spending, nor is it an estimate of the actual cost required to bring each UPRN into compliance. It is a scale conversion. It asks what notional capital expenditure would correspond to the floor area whose apparent policy-threshold status is conferred by the accounting revision alone. Let R_i^{obs} denote the latest observed EPC rating for UPRN i, let R_i^{cc} denote the corresponding constant-accounting counterfactual rating, and let A_i denote the building's recorded floor area. We evaluate two policy-

Table S19: Asset-rating change by primary heating fuel and air-conditioning status. Mean within-UPRN asset-rating change for straddling repeat-certified UPRNs, stratified by primary heating fuel and air-conditioning status at the first certificate. Counts denote the number of UPRNs in each subgroup.

Category	Specification	Counts	Mean Δ AR (pts)
Electric	*Without* Conditioning	42,568	−42.6
	With Air Conditioning	13,679	−37.2
Gas	*Without* Conditioning	35,760	−17.3
	With Air Conditioning	8,612	−22.4

relevant thresholds. The first is the legal F/G floor, for which a UPRN satisfies the threshold if its rating lies in $\mathcal{T}_{\text{floor}} = \{A, B, C, D, E\}$. The second is the EPC-B target, for which a UPRN satisfies the threshold if its rating lies in $\mathcal{T}_B = \{\text{A}, \text{B}\}$. For threshold t, we define an accounting-conferred threshold crossing as:

$$\text{D}_{it} = \begin{cases} 1 & \text{if } \text{R}_i^{\text{obs}} \in \text{T}_t \text{ and } \text{R}_i^{\text{cc}} \notin \text{T}_t, \\ 0 & \text{otherwise.} \end{cases} \tag{S5}$$

Thus, for the legal floor, $\text{D}_{it} = 1$ identifies UPRNs that are observed as A-E but would be F/G under constant accounting. For the EPC-B target, $\text{D}_{it} = 1$ identifies UPRNs that are observed as A/B but would fall below B under constant accounting. For each threshold, the number of accounting-conferred crossings is $\text{N}_t = \sum_i \text{D}_{it}$ and the corresponding floor area is $\text{M}_t = \sum_i \text{A}_i \times \text{D}_{it}$. The capital expenditure equivalent is then:

$$\text{CEE}_t(c) = c \times \text{M}_t \tag{S6}$$

where $c \in \{25, 50, 100, 200\}$ £/m^2. We use £100/m^2 as the central sensitivity value and report the remaining values as a unit-cost sensitivity grid. Since unit cost is fixed within each sensitivity, variation in the capital-expenditure equivalent is driven only by the floor area whose threshold status changes under the accounting counterfactual. The EPC-B and legal-floor sets are disjoint by construction, because a UPRN observed at $\text{AR} \leq 50$ cannot reach the F/G threshold ($\text{AR}^{\text{const}} > 125$) under the largest fuel-specific multiplier (≈ 1.63); we verified an empty intersection. Their two threshold-specific values can therefore be summed to a unique union without double-counting, though the union remains a floor-area scale conversion rather than an observed investment value. The estimate does not model measure-specific retrofit costs, premises-level technology choices, financing costs, VAT, disruption costs, avoided compliance costs, social welfare, or differences in achievable savings across UPRN types. The public EPC register also does not identify the legally regulated private-rented subset. Results should therefore be interpreted as the investment-scale equivalent of accounting-driven reclassification in the latest-stock register, rather than as legal-compliance cost estimates. The reported headline scale in the main text is not based on the hard fuel-only crossing sets above, which we retain only as a transparent baseline; it is based on the preferred aggregate threshold-probability model of Section 1.18, applied as an expected threshold-crossing count and expected affected floor area summed over eligible register entries on the unweighted latest-register frame, evaluated over the same unit-cost grid.

1.18 Held-out validation of the constant-accounting bridge and the expenditure-equivalent scale

The constant-accounting counterfactual is an empirical bridge that predicts a building's old-basis Asset Rating from its post-2022 rating using no-recorded-works calibration, not an SBEM recomputation. We validated it on the 2,442 no-recorded-works calibration UPRNs, whose actual pre-2022 rating is a direct observation of the old-basis rating, using deterministic five-fold UPRN-level cross-fitting stratified by fuel × observed post-revision band, so every case is held out exactly once and no test-fold outcome enters model fitting. We searched for identical-input SBEM 2013/2021 rerun pairs and found none available, so an external SBEM validation is not currently possible.

Definition of the preferred threshold-probability model. For each policy threshold $\tau \in \{50, 125\}$, corresponding to the EPC-B and F/G boundaries, the model estimates the probability that a UPRN currently meeting the threshold would have fallen below it under pre-revision accounting. For each calibration UPRN, the binary outcome indicates whether its observed pre-revision rating lay above the relevant policy threshold. The predictor is the signed distance of the observed post-revision rating from that boundary, scaled in 25-point EPC-band units as $x_i = (\mathrm{AR}_i^{\mathrm{post}} - \tau)/25$. A separate logistic regression is estimated for each main-heating-fuel group $f \in \{\mathrm{electric}, \mathrm{gas}, \mathrm{other}\}$,

$$\Pr(y_i = 1 \mid x_i, f_i = f) = \sigma(\beta_{0f} + \beta_{1f} x_i)\,, \tag{S7}$$

where $\sigma(z) = 1/\{1 + \exp(-z)\}$. Fuel enters through group-specific intercepts and slopes, while the post-revision rating enters through its distance from the relevant threshold. The model is fitted by penalised maximum likelihood with a ridge penalty of one applied to the slope and no penalty applied to the intercept. Its outputs are therefore logistic-model probabilities rather than empirical bin proportions or probabilities derived from rating-ratio distributions. No additional descriptors such as floor area, sector or air-conditioning status enter this specification. This deliberately low-dimensional structure limits sparse-cell problems in the calibration sample. Where a fuel group contains only one observed outcome class, the model uses its Laplace-smoothed outcome rate and sets the slope to zero. For eligible register entry j, let $\widehat{p}_j$ denote the predicted probability that its old-basis rating exceeded threshold τ. The expected crossing count and expected affected floor area are

$$\widehat{N}_\tau = \sum_{j \in \mathcal{E}_\tau} \widehat{p}_j, \qquad \widehat{A}_\tau = \sum_{j \in \mathcal{E}_\tau} A_j \widehat{p}_j, \tag{S8}$$

where $\mathcal{E}_\tau$ contains entries currently meeting the relevant threshold and A_j is recorded floor area. Entries outside this eligible set cannot lose threshold status and therefore contribute to neither functional. A hard probability cutoff is selected within the training data using balanced accuracy, but it is used only for classification diagnostics. The aggregate count and area estimates always use the complete predicted probabilities. Validation uses deterministic five-fold partitions at the UPRN level, stratified by heating fuel and observed post-revision EPC band. Each UPRN is held out exactly once, and every reported validation quantity is calculated from out-of-fold probabilities. Following validation, the model is refitted on the complete calibration sample and applied to the eligible register frame. The final register estimates therefore use all available calibration observations, while the reported validation performance remains fully out of sample.

The fuel-only multiplicative bridge does not support individually calibrated predictions of old-basis ratings. Its out-of-fold mean absolute error is 25.8 rating points, approximately the width of one EPC band, and its calibration slope of 0.48 shows that predictions are strongly compressed towards the fuel-group mean. The near-zero pooled bias does not indicate accurate individual prediction. Fuel-by-true-band errors offset one another by approximately 85% before aggregation, masking substantial errors within individual groups (Table S20). Some of this error reflects irreducible variation in repeated assessment rather than bridge misspecification alone. Among stable-descriptor certificate pairs assessed under the same methodology within six months, an identity prediction already produces mean absolute errors of 13.4 rating points for pre-revision pairs and 9.8 points for post-revision pairs. A bridge based on the same public register fields cannot recover information that varies even between closely spaced, same-method assessments.

The best-supported specification is therefore the cross-fitted, fuel-specific threshold-probability model. It is the only candidate to satisfy the prespecified out-of-fold validation criteria at both thresholds, comprising balanced accuracy of at least 0.65 and an aggregate probability-sum prevalence error within one percentage point. Its balanced accuracies are 0.73 at the B/C boundary and 0.71 at the E/F boundary, while aggregate prevalence errors remain within 0.02

percentage points (Table S20). These results support aggregate threshold-crossing estimation, but not exact reconstruction of individual old-basis ratings. Calibration is also less uniform within subgroups, where the largest observed probability error is 18.1 percentage points. We therefore interpret the model as supporting a register-conditional aggregate sensitivity rather than precise subgroup estimates or individual reclassification assignments.

Crossing counts and affected floor area require separate validation since they are different estimands. The model was selected partly on its ability to recover aggregate crossing prevalence, whereas the expenditure-equivalent calculation depends on the floor-area-weighted probability sum. A model can reproduce the number of crossings while remaining poorly calibrated among the small number of large premises that dominate affected area. Out-of-fold area-weighted validation confirms that the area estimand is less precisely calibrated than the count estimand (1.19). We consequently report expected crossings and expected affected floor area as aggregate quantities, present expenditure only as an area-to-unit-cost scale conversion, and retain the external-composition and alternative-model results as structural sensitivities (Table S38). We do not interpret these estimates as exact old-basis ratings, individually identified crossings, legal non-compliance, physical-UPRN totals or realised retrofit costs.

Table S20: Out-of-fold validation of the constant-accounting bridge specifications. Deterministic five-fold UPRN-level cross-fitting on the 2,442 no-recorded-works calibration UPRNs. MAE and calibration slope score the predicted old-basis rating; balanced accuracy and aggregate probability error score the EPC-B (B/C) and F/G (E/F) threshold classifications. The fuel-only bridge is a transparent baseline, whereas the threshold-probability model is the best-supported aggregate specification but not individually calibrated.

Specification	MAE	Cal. slope	B/C bal. acc.	E/F bal. acc.	B/C prob. err (pp)	E/F prob. err (pp)
Fuel-only multiplicative (baseline)	25.8	0.48	0.66	0.68	−3.3	1.9
Fuel-only additive (bound)	22.8	0.69	0.55	0.62	6.3	−3.3
Threshold-probability (best-supported)	–	–	0.73	0.71	−0.01	0.01

1.19 Area-weighted validation of the expenditure-equivalent scale

The preferred model is selected on aggregate crossing counts, but the expenditure-equivalent scale multiplies a unit cost by expected affected floor area. Note that such quantities are not equivalent estimands as a model can reproduce crossing prevalence while being poorly calibrated among the few very large premises that dominate total floor area. We therefore validate the area estimand directly, out of fold, on the calibration UPRNs (Table S21). Although the out-of-fold count error is negligible (≤ 0.01 percentage points of the sample at both thresholds), the model over-predicts affected floor area by 25.4% at the EPC-B threshold and by 112% at the F/G threshold. The over-prediction is concentrated in the largest premises: excluding the single largest premise cuts the EPC-B area error to 19.1%, and excluding the largest 25 cuts it to 6.5% (Table S22). By floor-area band, the model is well calibrated below 1,000 m^2 but over-predicts area by 17.9% in the 1,000–5,000 m^2 band and by 57.0% above 5,000 m^2 at the EPC-B threshold (Table S23).

The calibration sample is also not representative of the application population on the dimensions that drive area (Table S24). By floor area, the register application is 56% gas-heated whereas the calibration sample is 65% electrically heated; the calibration sample over-represents offices (37% versus 24% of application floor area) and under-represents warehouses and factories (10% and 6% versus 20% and 11%); and it over-represents better post-revision bands. As a result of the affected-area over-prediction and this transport gap both pointing in the direction of an overstated full-frame area, we pursue two independent corrections. First, we develop better-calibrated affected-area models under leakage-free cross-validation (1.20): enriching the fuel-specific threshold logistic with log floor area, sector and fuel interactions lowers the out-of-fold EPC-B area error from 25.4% to 8.9%, preserves the near-zero count error, improves the Brier score, removes the reversal on excluding the largest premises, and passes a joint count-and-area validation gate. Applied to the register, this gate-passing model lowers the expected affected area from 78.29 to 65.43 million m^2 at the EPC-B threshold and from 14.87 to 8.9 million m^2 at F/G. Second, reweighting the register composition to independent ND-NEED non-domestic floor-area margins lowers it to 50.24 million m^2 at EPC-B (11.56 at F/G). These diagnostics establish Model C as the leading untransported outcome-model candidate. They do not determine the final aggregate register-conditional estimator, which additionally requires transport from the calibration sample and validation of the residual correction. Final estimator selection and the frozen primary estimate are reported in 1.21.

Table S21: Out-of-fold calibration of crossing counts and affected floor area. Observed crossings and predicted expected crossings are aggregated across the held-out calibration folds for the EPC-B and F/G thresholds. The model reproduces aggregate crossing counts closely but overpredicts affected floor area, particularly at the F/G threshold. Area error is calculated relative to the observed affected area.

Threshold	Obs. count	Pred. count	Obs. area (Mm^2)	Pred. area (Mm^2)	Area error
EPC-B (B/C)	2,174	2,173.7	1.359	1.703	+25.4%
Floor (E/F)	260	260.3	0.060	0.127	+112%

Table S22: Large-premise influence on the out-of-fold area error. Out-of-fold predicted-vs-observed affected-area error after excluding the largest k premises by floor area with over-prediction is driven by a small number of large premises.

Threshold	Exclude 0	Exclude 1	Exclude 5	Exclude 10	Exclude 25
EPC-B (B/C)	+25.4%	+19.1%	+11.2%	+9.8%	+6.5%
Floor (E/F)	+112%	+104%	+94%	+85%	+66%

Table S23: Calibration by floor-area band (EPC-B threshold, out of fold). Observed and predicted crossing prevalence and the affected-area error within each floor-area band. Calibration is good below 1,000 m^2 and deteriorates in the largest bands, which dominate total floor area.

Floor-area band (m^2)	Counts	Obs. prev. (%)	Pred. prev. (%)	Area error
0–100	672	92.7	91.9	+0.3%
>100–500	1,150	92.4	89.8	−2.6%
>500–1,000	273	89.0	87.8	−0.6%
>1,000–5,000	279	71.0	82.1	+17.9%
>5,000	68	69.1	81.8	+57.0%

Table S24: Calibration-vs-application transport check. Share of the calibration sample and of the register application across heating fuel and post-revision band. The calibration sample under-represents the gas-heated and better-band UPRNs that dominate application floor area.

Dimension	Level	Calib. (%)	Appl. count (%)	Appl. area (%)
Fuel	Electric	64.6	56.6	37.2
	Gas	30.6	36.7	56.4
	Other	4.8	6.7	6.5
Post band	A/B	45.5	22.3	37.7
	C	34.6	33.7	31.7
	D	14.7	26.7	20.0
	E	5.0	13.4	8.4
	F/G	0.3	4.0	2.3

1.20 Better-calibrated affected-area models

The baseline model calibrates counts but over-predicts area, so we developed candidate affected-area models under the same deterministic five-fold UPRN cross-validation, selecting on a joint gate rather than on count calibration alone. All candidate models predict whether the pre-revision rating exceeded the relevant threshold using only predictors available at application. Outcomes for held-out cases are excluded from model fitting, preventing information leakage. Model A is the baseline fuel-specific threshold logistic. Model B adds log floor area and sector to the pooled logistic; Model C adds fuel-by-rating and fuel-by-size interactions; Model D replaces the linear terms in post-revision rating and log floor area with natural cubic splines. The joint gate requires out-of-fold area error within about 10%, no reversal of sign on excluding the largest 25 premises, and an over-5,000 m^2 error within about 30%. To keep the entire selection procedure out of sample, we use fully nested, repeated cross-validation: within each outer training fold, inner cross-validation selects both the model architecture (A–D) and the ridge penalty by inner-fold area error, and only the inner-selected model is scored on the untouched outer fold; the procedure is repeated over 25 UPRN-level fold assignments, reporting medians and 2.5–97.5 percentiles across seeds (Table S25). Under this protocol the whole selection procedure achieves an out-of-fold area error of 8.0% (2.5–97.5 percentile 6.5–9.6%) at the EPC-B threshold with essentially zero count error, and it selects an interaction model on every seed—Model C on 68% of outer folds and the spline Model D on 32%, never the simpler A or B. We exclude any model failing the count-calibration and balanced-accuracy gates, then minimize absolute affected-area error, breaking ties by Brier score. As a result of all four models passing the count and balanced-accuracy gates, selection reduces to area error and returns the interaction/spline models. Evaluated as fixed models, Model C has the lowest median area error (8.2%, 7.4–9.4%) and the best Brier score, so we treat Model C as the leading parsimonious primary candidate at the EPC-B threshold and retain Model D as architecture uncertainty; the transport (1.21) and bootstrap analyses reselect between them rather than fixing one. Applied to the full register (fitted on the complete calibration set), Model C gives an expected 95,074 EPC-B-threshold crossings over 65.43 million m^2 of affected floor area (£6.5 billion at £100/m^2); this expected count supersedes the baseline model's 94,334.

At the F/G threshold we separate the count and area estimands, because they validate very differently. The expected F/G crossing count is well calibrated out of fold for every model (count error near zero), so we retain 36,684 as the preferred expected F/G count. The F/G affected area, by contrast, fails the gate for every candidate under nested CV: the selection procedure's out-of-fold area error is 13.8% (11.0–16.4%) and no architecture reaches the 10% target. We therefore do not present a single preferred F/G area or a single F/G monetary conversion of equal standing to the EPC-B figure. Instead we report the F/G affected area as a less-reliable structural range across the gate-informed candidate models, 8.9–14.9 million m^2 (£0.9–1.5 billion at £100/m^2; Table S26), and flag it explicitly as not gate-passing. We continue to report the composition-reweighted and full-frame values as a lower alternative and an upper reference. These affected-area and expenditure quantities remain provisional pending the transport and full-pipeline-bootstrap analyses.

Table S25: Fully nested, repeated cross-validation of candidate affected-area models. Medians over 25 UPRN-level fold assignments; within each outer training fold inner CV selects both architecture and penalty, and only the inner-selected model is scored on the untouched outer fold. Area error is the out-of-fold predicted-vs-observed affected floor area. Excl. top 25 repeats it after removing the largest 25 premises. Gate is whether the median passes the joint count-and-area gate. Across seeds the selection procedure picks Model C on 68% and Model D on 32% of EPC-B outer folds, with a selection-procedure out-of-fold area error of 8.0% (6.5–9.6%). Model A is the baseline, whereas Models B–D add log floor area, sector and (C) fuel interactions or (D) splines.

Threshold	Model	Count Error (pp)	Area Error	Excl. top 25	Brier	Gate
EPC-B	A baseline	0.00	+25.1%	+6.5%	0.085	no
	B +logFA+sector	+0.02	+9.5%	−2.2%	0.079	yes
	C +interactions	**+0.02**	**+8.2% [7.4, 9.4]**	**−1.8%**	**0.078**	**yes**
	D splines	+0.02	+8.5% [7.1,10.1]	−2.1%	0.080	yes
F/G	A baseline	0.00	+110%	+65%	0.083	no
	B +logFA+sector	0.00	+18.7%	+8.9%	0.082	no
	C +interactions	−0.01	+13.8% [12.1,15.9]	+6.7%	0.081	no
	D splines	0.00	+19.0%	+6.2%	0.081	no

Table S26: Register affected area and expenditure-equivalent scale by area model. Each candidate is fitted on the full calibration set and applied to the latest-register frame. The enriched models converge to roughly 65 million m^2 at EPC-B, about 17% below the over-predicting baseline.

Threshold	Model	Affected area (Mm^2)	£bn (£100/m^2)
EPC-B	A baseline	78.35	7.83
	B +logFA+sector	64.41	6.44
	C +interactions	**65.43**	**6.54**
	D splines	64.80	6.48
F/G	A baseline	15.02	1.50
	enriched B–D	8.9–9.9	0.9–1.0

The preferred expected F/G count is 36,684 (count calibration passes out of fold); the F/G area row is a less-reliable sensitivity range because no model passes the area gate at F/G, and the two are not combined into a single preferred F/G estimate.

1.21 Transport from the calibration sample to the register

Internal cross-validation evaluates performance among held-out calibration cases, but it does not establish whether the estimator transports to the EPC-register application frame, whose composition differs from that of the calibration sample. We therefore distinguish two stages. Stage 1 transports the stable-descriptor calibration sample to the eligible EPC-register frame and defines the primary register-conditional estimand. Stage 2 standardises the register frame to external non-domestic premises benchmarks, including ND-NEED [29] and BEES, to examine sensitivity to the composition of the wider stock of buildings. The Stage 2 estimates are external-composition sensitivities and are not interpreted as resolving Stage 1 transport.

Transport is estimated separately for each policy threshold, since only entries currently meeting a threshold can lose that status under pre-revision accounting. At the EPC-B threshold, the eligible population comprises post-update latest-register entries with an asset rating of 50 or lower. This gives 1,110 eligible calibration UPRNs and 118,865 eligible register entries. Let Z_i indicate whether record i belongs to the calibration sample, with $Z_i = 1$ for calibration cases and $Z_i = 0$ for entries in the eligible application frame. We estimate the probability of calibration-sample membership as $e(x) = \Pr(Z = 1 \mid X = x)$, where X includes main heating fuel, post-update asset rating, log floor area, sector and air-conditioning status. Missing values in categorical variables are retained as separate levels. The estimated membership probability is used to construct an inverse-odds transport weight for each calibration case, $w_i \propto \frac{1-e(x_i)}{e(x_i)}$, providing, calibration cases whose characteristics are more common in the application frame than in the calibration sample, a greater influence. We trim the weights at the 1st and 99th percentiles and normalise them to have mean one.

Coverage of the eligible application frame by the calibration sample is high. Only 0.18% of eligible register entries fall outside the range of membership probabilities observed in the calibration sample, and the weighted calibration sample has a Kish effective sample size of 857 from 1,110 cases. Inverse-odds weighting also removes most observed compositional differences, with standardised mean differences of no more than 0.03 for post-update rating, log floor area and heating fuel after weighting (Table S28). First-moment entropy balancing over the full categorical design was also examined, but collinear constraints caused the effective sample size to collapse. We therefore use inverse-odds weighting for Stage 1 transport. For a policy threshold τ, let $y_i = \mathbf{1}\{AR_{i,\text{pre}} > \tau\}$ indicate whether an eligible calibration case would fall below the threshold under pre-revision accounting. At the EPC-B threshold, $\tau = 50$. We estimate the crossing probability $\hat{p}(x) = \Pr(y = 1 \mid X = x)$ from the calibration sample using the transport weights w_i. Let $\mathcal{T}$ denote the eligible application frame, let N_T be its number of entries and let A_j denote the floor area of application entry j. The outcome-regression estimates of the expected crossing count and affected floor area are

$$\widehat{N}_{\text{OR}} = \sum_{j\in\mathcal{T}} \hat{p}(x_j), \quad \widehat{A}_{\text{OR}} = \sum_{j\in\mathcal{T}} A_j \hat{p}(x_j). \tag{S9}$$

These expressions sum the predicted crossing probabilities across eligible register entries, either unweighted for the expected count or weighted by floor area for the expected affected area. The finite-target doubly robust estimator supplements these application-frame predictions with a correction based on weighted residuals in the calibration sample. Let $\mathcal{S}$ denote the calibration source. The estimators are

$$\widehat{N}_{\mathrm{DR}} = \sum_{j \in \mathcal{T}} \hat{p}(x_j) + \frac{N_T}{\sum_{i \in \mathcal{S}} w_i} \sum_{i \in \mathcal{S}} w_i \{y_i - \hat{p}(x_i)\},$$

and

$$\widehat{A}_{\mathrm{DR}} = \sum_{j \in \mathcal{T}} A_j \hat{p}(x_j) + \frac{N_T}{\sum_{i \in \mathcal{S}} w_i} \sum_{i \in \mathcal{S}} w_i A_i \{y_i - \hat{p}(x_i)\}.$$

The first term in each expression is the application-frame outcome prediction. The second term corrects systematic prediction error using the weighted difference between observed and predicted outcomes in the calibration sample. The factor $N_T / \sum_i w_i$ places this correction on the scale of the eligible target population. Both the crossing model $\hat{p}$ and the membership model e are cross-fitted, so each source case is evaluated using models estimated without that case.

Numerical checks confirmed the expected double-robustness property. The estimator recovers the target count and area when the crossing model is correctly specified. It also recovers these quantities through the weighted residual correction when the crossing model is misspecified but the membership model is correct. Mean-one normalisation and normalisation of the weights to sum to N_T produce identical estimates. The doubly robust estimator is therefore carried forward into the estimator comparison rather than selected in advance.

Transport is evaluated through out-of-sample predictive performance rather than covariate balance alone. We repeated the 25-partition outer validation with both the membership and crossing models refitted within each training fold. Each held-out fold was then scored using its estimated transport weights. Inverse-odds weighting reduced the target-weighted affected-area error from 13.0% to 9.4% for Model C and from 19.5% to 5.9% for Model D (Table S27).

The ordering of the two architectures changes after transport. Model D has a lower median transported affected-area error of 5.9%, with a 2.5th to 97.5th percentile range of 2.3% to 10.0%. The corresponding error for Model C is 9.4%, with a range of 7.1% to 12.3%. The intervals overlap, so these transported point estimates do not by themselves justify fixing one architecture. Applied to the eligible register frame, Stage 1 transport leaves the expected EPC-B crossing count essentially unchanged at about 94,900. The transported affected-area estimate is 61.3 million m^2 under Model C and 65.2 million m^2 under Model D. At £100/m^2, these values correspond to expenditure-equivalent scales of approximately £6.1 billion and £6.5 billion. For comparison, the untransported Model C estimate is 65.4 million m^2, while the ND-NEED Stage 2 external-composition sensitivity is 50.2 million m^2. BEES is retained as a separate external-composition benchmark rather than treated as another Stage 1 transport estimate. The primary Stage 1 estimator is selected below using a prespecified rule based on threshold accuracy, count calibration, affected-area performance, common support, stability and parsimony.

Since a small share of entries can represent a larger share of area, we therefore report common-support failure as 0.18% of eligible register entries, 0.59% of eligible register floor area and 0.46% of the untransported predicted crossing mass fall outside the calibration membership-probability range. The 209 unsupported entries are large Education and Hospitality premises (largest 87,862 m^2), sectors thin in the office/retail-dominated calibration sample. The transported estimate is robust to the weight-handling rule as across no trimming, 0.5/99.5 and 1/99 percentile trimming and weight caps of 10, 20 and 50, the effective sample size stays between 851 and 1,109, the maximum

Table S27: Out-of-sample validation of transported estimators (EPC-B eligible population). Median [2.5–97.5 percentile] over 25 outer partitions; membership and outcome models are refit inside each fold and the held-out fold is scored under its transport weights. Target-weighted area error weights held-out cases by their transport weights to represent the application population. Inverse-odds weighting improves held-out prediction for both models, with Model D transports yielding the highest results.

Model	Estimator	Held-out area err	Target-weighted
C	unweighted	+11.1%	+13.0% [11.4,15.7]
C	inverse-odds	+8.3%	+9.4% [7.1,12.3]
D	unweighted	+15.5%	+19.5% [16.8,24.7]
D	inverse-odds	+1.3%	+5.9% [2.3,10.0]

Table S28: Stage-1 transport diagnostics and estimates (EPC-B eligible population). Standardized mean differences (calibration vs register) before and after inverse-odds weighting, and the transported expected count and affected area. Entropy balancing is omitted as numerically unstable. Common support 99.8%; Kish effective sample size 857/1,110.

Covariate SMD	Raw	Inverse-odds
Post-update rating	+0.08	+0.00
Log floor area	+0.13	+0.03
Fuel		
Electric	−0.09	+0.00
Gas	+0.13	−0.01
EPC-B estimate	**Count**	**Area (Mm^2)**
Untransported (Model C)	94,770	64.85
Stage-1 transported (Model C)	94,920	61.27
Stage-1 transported (Model D)	96,015	65.21

standardized mean difference between 0.03 and 0.12, and the Model C transported area between 61.0 and 64.4 million m^2 (Model D 65.0–65.9), so the trimming choice does not drive the result (Table S32). A reduced-rank calibration/entropy-balancing specification attains weaker continuous-covariate balance than inverse-odds weighting and a consistent area of 62.9 million m^2. Inverse-odds weighting therefore remains the usable Stage-1 method under the prespecified balance ($|\text{SMD}| \leq 0.10$) and effective-sample-size rules, with reduced-rank calibration a consistent but less well balanced alternative.

We select the Stage-1 estimator by rejecting estimators failing count-calibration or balanced-accuracy; reject weighting failing support, balance or minimum effective sample size; among the rest, take the lowest median absolute target-weighted area error subject to subgroup and large-premises stability; break negligible ties by Brier score, log loss and parsimony. Applying this rule requires the out-of-sample performance of the doubly robust estimator, whose application estimate is materially below the weighted-outcome-regression estimates and so cannot be treated as a mere robustness check. We validate it under a strict leakage-free pseudo-target design where within each outer partition the held-out fold is treated as an unlabelled pseudo-target; the outcome model is fit on the training cases only; the membership model uses the covariates of the held-out fold but never its outcomes; the doubly robust residual correction sums over the training cases with cross-fitted source predictions. The held-out outcomes are revealed only for scoring. Automated leakage tests confirm the design: the doubly robust estimate is unchanged when the held-out outcomes are zeroed or permuted before estimation, and the residual-correction index set is exactly the training set with empty intersection with the held-out fold. Under this design (Table S30), the outcome models over-predict the held-out affected area (median absolute error 11.1% for Model C and 15.5% for Model D), and the weighted-outcome-regression does not fix this (9.8% for C; 22.8% for D), whereas the doubly robust estimator reduces it to 0.7% (C) and 2.4% (D) and wins every one of the 25 partitions against both the plug-in and the weighted-outcome-regression, with a stable cross-fitted residual correction of −11% (C) and −17% (D); count calibration is within 0.1 percentage points for all estimators. The doubly robust advantage therefore survives the leakage audit and strengthens under cross-fitting, so under the selection rule it is the primary estimator; Models C and D are practically equivalent under it, and we fix the parsimonious cross-fitted doubly robust Model C as primary with doubly robust Model D as architecture sensitivity.

The frozen Stage-1 register-conditional result can be found in Table S31. On the complete eligible EPC-B frame, the primary cross-fitted doubly robust estimate is an expected 94,911 crossings over 57.2 million m^2 of affected floor area (£5.7 billion at £100/m^2), with an architecture sensitivity of 54.9 million m^2 (Model D). The weighted-outcome-regression (61.3 million m^2) and unweighted (64.9 million m^2) estimates are retained as estimator sensitivities that over-predict area out of fold. On the policy-proxy subset (8,180 entries) the estimate is of order 5,000 crossings over 17 million m^2. We distinguish throughout the untransported estimate (64.9), the estimator sensitivities (weighted-outcome-regression 61.3), the selected doubly robust Stage-1 estimate (57.2), the architecture sensitivity (54.9), and the external Stage-2 sensitivities reported below.

For the F/G threshold, we transport the expected crossing count with the selected doubly robust procedure. The preferred Model C estimate is 36,847, with Model D providing an architecture sensitivity of 37,783. The F/G affected area is reported only as the non-gate-passing range across the transported estimators, 9.05–9.63 million m^2. A conceptual clarification governs the external sensitivities. The cross-fitted outcome model supplies per-entry plug-in probabilities $\hat{p}(x_j)$. The doubly robust estimator is an aggregate functional, denoted as:

$$\Theta_{\mathrm{DR}}(g) = \sum_{j \in T} g_j \hat{p}(x_j) + \frac{N_T}{\sum_i w_i} \sum_{i \in S} w_i g_i \{y_i - \hat{p}(x_i)\} \tag{S10}$$

with $g_i = 1$ for the expected count, $g_i = A_i$ for the affected area and $g_i = A_i c_i$ for an expenditure functional. Note that the residual correction is an aggregate term and is not allocated across target entries. For a Stage-2 target with application weights r_j, the functional is recomputed as:

$$\Theta_{\mathrm{DR}}(r, g) = \sum_{j} r_j g_j \hat{p}(x_j) + \frac{\sum_j r_j}{\sum_i w_i^{(r)}} \sum_{i} w_i^{(r)} g_i \{y_i - \hat{p}(x_i)\} \tag{S11}$$

where $w_i^{(r)}$ are refit to transport the calibration sample to the same r-weighted target.

1.22 Stage-2 external-composition sensitivities

Stage-2 asks how the register-conditional estimand would shift if the eligible application frame had the composition of the wider non-domestic stock, keeping the Stage-1 register-conditional value as the principal result.

Full-frame construction. ND-NEED margins describe the whole non-domestic stock, not the subset currently certified at EPC-B or better. Raking the eligible subset directly to whole-stock margins would implicitly assume that EPC-B eligibility is unrelated to sector and size, which is indefensible. We therefore (i) construct external-composition weights on the complete latest-certificate register frame by iterative proportional fitting to compatible ND-NEED margins; (ii) inherit those full-frame weights to the entries in the EPC-B-eligible application subset; (iii) treat the resulting weighted distribution as the externally standardized eligible target; (iv) refit the source-membership weights so the labelled calibration cases represent that same weighted eligible target; and (v) recompute the doubly robust count and area functionals. Both functionals are evaluated only over entries currently rated EPC-B or better. Entries below EPC-B contribute zero, and threshold-crossing probabilities are defined within this eligible subset.

Margin ladder and the role of floor area. A certificate-use-only rake is insufficient for an affected-area estimand when compatible size margins exist. We therefore report a ladder of compatible ND-NEED margin sets (Table S29) using sector/use only, floor-area size only, sector+size, and sector+size+region. ND-NEED's Table 2 [29] size margin is a floor-area band, not a business/employment size. Note that ND-NEED does not supply a compatible business-size margin, so no employment-size rake is attempted. Sector-only reweighting raises affected area to 65.8 million m^2 as it up-weights large-premise sectors, whereas adding the floor-area margin, turns the eligible floor-area median from 208 to 123 m^2 and the area down to 43.9 million m^2. The preferred ND-NEED sensitivity leverages sector+size+region yielding 98,600 expected crossings over 43.9 million m^2 of affected floor area, against the Stage-1 register-conditional 94,911 crossings and 57.2 million m^2. All raked margins converge to their targets to within 10^{-4}. The raw floor-area standardized difference under reweighting is small (-0.08).

Two clearly separated interpretations. We report only the composition-standardized sensitivity as the inherited eligible weights are normalized to sum to the observed eligible denominator $N_T = 118{,}750$, isolating composition while preserving the register denominator. Note that as ND-NEED's unit of observation is the building, it cannot be used for reconciling record-for-record with EPC-register UPRNs.

Weight diagnostics. Table S29 reports, for each margin set, the sum of full-frame weights, the inherited eligible weight sum before renormalization, the externally weighted share of the stock currently at EPC-B or better, Kish effective sample size for the full frame and the eligible subset, and the weight share in the largest 1/5/10% of eligible entries. Under the richest set the eligible effective sample size is 44,480 and the largest 1% of entries carry 10% of the weight. Achieved margin balance is exact by construction; residual imbalance is therefore confined to non-raked features and is reflected in the small floor-area standardized difference above.

Region-unmatched rows. Raking to region margins requires a region match. As a result, 170 of 842,673 latest-frame rows (0.02%) are region-unmatched and excluded from the region rake, constituting predominantly certificates labelled to Dumfries and Galloway, a Scottish local authority outside the England-and-Wales frame, and total 0.25 million m^2 of floor area. Reintroducing them under an *unmatched* category changes the area estimate by less than 0.4%. The sector- and size-only rakes, which do not require a region match, retain the full matched frame.

Table S29: Stage-2 external-composition sensitivities (ND-NEED), full-frame construction. Weights are constructed on the complete latest-certificate frame by iterative proportional fitting to compatible ND-NEED margins, inherited to the EPC-B-eligible subset, and register-standardized to $N_T = 118{,}750$. Doubly robust (DR) count and area functionals with cross-fitted Model C nuisances. The register row reruns the Stage-1 functional on the region-matched frame of 842,503 entries and 118,750 eligible entries. It therefore differs from the frozen full-frame estimate of 94,911 crossings and 57.24 million m^2, based on 842,673 entries and 118,865 eligible entries. Effective sample sizes are Kish; top-k% is the eligible weight share in the largest k% of entries.

Margin Set	DR Count	DR area (Mm^2)	Corr. (Mm^2)	ESS (full)	ESS (elig.)	Top 1%
Register (Stage-1)	94,815	57.6	−7.2	842,503	118,750	0.01
ND sector only	93,705	65.8	−12.5	470,746	67,113	0.08
ND size only	98,585	42.3	−2.7	683,796	96,485	0.02
ND sector + size	98,730	43.9	−2.8	331,387	46,578	0.09
ND sector + size + region	98,600	43.9	−2.8	328,482	44,480	0.10

BEES. We include as part of the study the Building Energy Efficiency Survey (BEES) to be used as a qualitative composition benchmark. Its unit of observation is the surveyed premises/business; its sector and floor-area categories and its rented/owned and survey-weight definitions are not directly commensurate with the eligible EPC application frame. As forcing a full BEES reweighting would conflate incompatible denominators, we report BEES as a partial-composition comparison on the compatible sector margins only, document which margins are and are not compatible, and note that applying only the compatible sector margins moves the affected area in the same direction as the ND-NEED sector-only rake.

The transported sector-cost recomputation as separate doubly robust cost functionals (1.24) and the full-pipeline bootstrap with its sanity checks (1.25) are reported below. Note that the main-text expenditure figure is held at the register-conditional area-calibrated value, with the external-composition, architecture, transport, unit-cost and unmatched-sector ranges reported as separate structural sensitivities rather than folded into a single interval.

Table S30: Leakage-audited pseudo-target validation of the transport estimators. Median [2.5–97.5 percentile] absolute error in recovering each held-out fold's true affected area over 25 partitions, strict pseudo-target design with cross-fitted source residuals. Tests confirm the doubly robust estimate is invariant to zeroing or permuting held-out outcomes and that the residual correction indexes only training cases.

Estimator	Model C	Model D
Plug-in (unweighted) |area err|	11.1%	15.5%
Weighted-outcome (IO) |area err|	9.8%	22.8%
Doubly robust |area err|	0.7% [0.0,2.6]	2.4% [0.1,6.1]
Paired |DR| − |plug-in| (% win)	−10.2 pp (100%)	−13.2 pp (100%)
Paired |DR| − |IO| (% win)	−8.8 pp (100%)	−19.9 pp (100%)
DR cross-fitted residual correction	−10.7%	−17.3%

Table S31: Frozen Stage-1 register-conditional estimate. Primary estimator is a cross-fitted doubly robust Model C (materially better validated area calibration; Table S30). Model D-DR is architecture sensitivity; the weighted-outcome-regression (inverse-odds, *IO*) and unweighted estimates are estimator sensitivities that over-predict area. Areas are reported in million m^2.

Frame (N_T)	Unweighted	IO	DR primary (C)	Arch. (D-DR)
EPC-B eligible (118,865) count	94,770	94,919	94,911	95,866
EPC-B eligible area	64.85	61.27	**57.24**	54.94
Policy-proxy (8,180) area	17.30	17.21	∼17	—
F/G eligible (316,293) count	37,052	36,690	36,847	37,783
F/G area (not gate-passing)	9.05	9.63	9.10	9.26

Table S32: Weight-trimming sensitivity of the Stage-1 transported estimate. Inverse-odds weights under prespecified trimming/capping rules. ESS is Kish effective sample size; max SMD is the largest standardized mean difference over post-rating and log floor area. Areas are the transported expected EPC-B affected area reported in Mm^2 under Models C and D.

Rule	Max wt	ESS	Max SMD	C area	D area
No trimming	97.9	851	0.03	61.0	65.1
0.5/99.5 pct	78.2	855	0.03	61.2	65.1
1/99 pct (default)	75.7	857	0.03	61.3	65.2
Cap 10	10.0	1109	0.12	64.4	65.9
Cap 20	20.0	1078	0.09	62.8	65.0
Cap 50	50.0	929	0.05	61.5	65.1

1.23 Sector-specific cost schedule

The generic £25–200/m^2 grid applies one rate to all floor area. Where a defensible source exists we replace it with a sector-specific schedule. The Investment Property Forum costing study [32] models six commercial archetypes at Q2 2017 prices with costs including materials, labour, builder's work, preliminaries, overheads, contingencies and profit, and exclude VAT, design, Building Control, survey, legal and finance fees. Its approximate EPC-B package costs are £200–260/m^2 for the four office archetypes (F→B or E→B), £80–120/m^2 for the industrial/storage warehouse (G→B), and £100/m^2 (lower package) to £350/m^2 (deeper package) for the retail warehouse (D→B), hailed from different starting ratings and packages and are not averaged into a single rate (Table S33). We rebase Q2 2017 costs to 2024-25 prices using the HM Treasury GDP deflator (June 2026 release; index 74.4665 for 2017-18 and 96.6694 for 2024-25, multiplier 1.298) [33]. Deflator values and schedule are stored as machine-readable inputs and the original-price values are retained.

We then map register sectors to archetypes conservatively (Table S34) with offices to the office archetype range, and factories, warehouses and storage/distribution to the industrial archetype. Expenditure is computed at entry level as $\sum_i A_i \, \hat{p}_i \, c_{s(i)}$ over eligible register entries, with $\hat{p}_i$ the preferred-model crossing probability, A_i the floor area and $c_{s(i)}$ the sector-specific rebased unit cost.

Two bases must be distinguished. First, under the primary doubly robust area functional, the three sectors with comparable IPF archetypes account for 30.15 million m^2, or 52.6% of the 57.24 million m^2 affected area, and the remaining 27.11 million m^2 (47.4%) is concentrated in shops, hospitality and other unmatched sectors, representing quantities carried into Figure 3c and Section 1.24. Earlier plug-in decomposition below assigns 56% of the larger, over-predicting 64.85 million m^2 plug-in area to matched sectors and is retained only as an estimator sensitivity, not as the primary basis.

On that plug-in basis, at the EPC-B threshold, sectors with a directly comparable archetype account for 36.59 million m^2 (56% of the plug-in affected area) and imply an expenditure of £6.0–8.3 billion (central £7.1 billion, 2024-25 prices); the office-dominated composition means the generic £100/m^2 rate materially understates this covered cost. The remaining 28.85 million m^2 (44%) is unmatched; applying the generic £25–200/m^2 grid to it adds £0.7–5.8 billion, so the full-register figure is £6.7–14.0 billion only under the explicit assumption that unmatched sectors cost within the generic range (Table S35). As an aggregate plausibility check, the government's non-domestic private-rented-sector impact assessment reports central EPC-B modelled costs of about £4.6 billion capital and installation and £6.1 billion total present-value cost in 2018 prices [22].

Table S33: IPF (2017) archetype EPC-B package costs and GDP-deflator rebasing. Q2 2017 costs from the IPF commercial-building costing study [32], rebased to 2024-25 prices (multiplier 1.298; HM Treasury GDP deflator, June 2026 [33]). Costs include contractor materials, labour, overheads and profit and exclude VAT, design and professional fees.

Archetype	Start → B	2017 (£/m^2)	2024-25 (£/m^2)
Office (four archetypes)	F/E → B	200–260	260–338
Retail warehouse	D → B	100–350	130–454
Industrial/storage	G → B	80–120	104–156

Table S34: Register sector to IPF archetype crosswalk. Only offices and industrial/storage sectors have a directly comparable archetype; all other sectors are left unmatched rather than assigned office costs. Central 2024-25 unit costs are shown for matched sectors.

Register sector	Archetype	Central £/m^2 (2024-25)
Offices	Office archetypes	299
Factories	Industrial/storage	130
Warehouses	Industrial/storage	130
Others†	—	—

† *Shops*, *Hospitality*, *Education*, *Health*, *Arts*, *Community*, *Leisure*, and *Emergency* could not be matched, and therefore no archetype could be assigned.

1.24 Sector cost as separate doubly robust functionals

The entry-level scale conversion above is a plug-in sum. To carry the same doubly robust correction used for the count and area estimands into the covered-sector costs, we treat each covered sector as a separate functional $g_i^{(s)} = A_i c_s \mathbf{1}\{\text{sector}_i = s\}$ with c_s the rebased IPF central unit cost, and evaluate $\Theta_{\text{DR}}(g^{(s)})$ on the register-conditional (Stage-1) target using the same cross-fitted Model C, and, as an architecture sensitivity, Model D, nuisance and the same inverse-odds source weights (Table S36). The covered sectors are Offices, Factories and Warehouses. Measured on the same doubly robust basis as the cost functionals themselves, they hold 30.15 million m^2, or 52.6%, of the 57.24 million m^2 primary affected area. The remaining 27.11 million m^2 (47.4%) is concentrated in shops and hospitality and has no comparable IPF archetype, and is excluded from the doubly robust

Table S35: Sector-specific expenditure-equivalent scale (2024-25 prices). Entry-level expenditure $\sum_i A_i \hat{p}_i c_{s(i)}$ using the preferred area model and the crosswalk. Covered sectors use IPF archetype costs. Unmatched sectors are reported separately, and the full-register figure applies the generic £25–200/m^2 grid to unmatched area under an explicit assumption.

Component	EPC-B (£bn)	F/G (£bn)
Covered sectors (56% of plug-in area)	6.0–8.3 †	1.0–1.3 ‡
Unmatched area, generic £25–200/m^2	0.7–5.8	0.1–0.8
Full register (explicit assumption)	6.7–14.0	1.1–2.1

† Represents values relative to the central $7, 1$. ‡ Represents values relative to the central $1, 2$.

cost estimate rather than assigned office costs. The corresponding plug-in decomposition, which assigns 56% of the over-predicting 64.85 million m^2 plug-in area (13.7, 5.4 and 17.2 million m^2) to the same sectors, is reported only as an estimator sensitivity.

The per-sector functionals reproduce the covered-total cost functional to numerical tolerance ($< 10^{-12}$ £bn) under both architectures, giving a covered-sector doubly robust cost of £6.46 billion (Model C) and £6.14 billion (Model D), against a covered plug-in of £7.03 billion, namely the same downward correction that takes the plug-in affected area (64.85 million m^2) to the doubly robust 57.2 million m^2. The correction is not uniform across sectors, and two of the three covered sectors have limited labelled support. Offices are well supported (415 labelled source pairs, effective sample size 393): the doubly robust cost is £4.50 billion (Model C) / £4.51 billion (Model D), no single source case contributes more than 8% of the correction, and removing the ten largest residual contributors moves the estimate by under 1%. Factory support is sparse (45 source pairs; effective sample size 41), and one case contributes 29% of the correction. We therefore report the factory plug-in value of £0.70 billion as the stable central value. Warehouses are adequately counted (104 pairs, effective sample size 97) but yielded large and negative (£-0.93 billion Model C, £-1.33 billion Model D) and top-heavy. As a result, excluding the ten largest source residual contributors moves the estimate from £1.31 to £2.42 billion. We therefore report the plug-in £2.23 billion as the stable central value with the doubly robust correction as a downward sensitivity. Taken together, the covered-sector expenditure is £6.1–7.0 billion — the plug-in £7.0 billion, the fully-corrected doubly robust £6.1–6.5 billion, with the difference driven almost entirely by the large but unstable warehouse correction. As factories and warehouses fail the stable-correction criterion, we do not present sector-specific doubly robust point costs for them.

1.25 Functional sanity checks and full-pipeline bootstrap

Prior to quantifying sampling uncertainty we confirm that every doubly robust functional is admissible and not driven by a few source cases. At the EPC-B threshold the register-conditional count is 94,937, inside $[0, N_T]$ with $N_T = 118{,}865$; the largest single source residual contributes 1.0% of the count correction, and removing the ten largest contributors moves the count by under 2% (to 96,839). The affected area is 57.29 million m^2, inside [0, 106.3] (total eligible area); all covered-sector costs are non-negative; and the per-sector cost functionals sum to the covered-total functional to within 10^{-12} £bn (1.24). The area correction is more concentrated than the count correction, as expected when a few large gas-heated premises drive the over-prediction it corrects: the single largest source residual is 6.6% of the area correction, and removing the ten largest contributors moves the area to 65.8 million m^2 (about +15%, back toward the plug-in 64.85 million m^2).

We then propagate the calibration-sample sampling uncertainty with a full-pipeline bootstrap

Table S36: Covered-sector cost as separate doubly robust functionals $\Theta_{\mathrm{DR}}(g^{(s)})$, register-conditional (Stage-1) target, 2024-25 prices. Unit costs are IPF central archetype costs rebased by the GDP deflator. Source counts and effective sample size (ESS, Kish) are within-sector labelled support. The largest single source residual contribution as a fraction of the sector correction is denoted by *max share*. The estimate after removing the ten largest residual contributors is denoted by *drop-10*. The per-sector functionals sum to the covered-total functional exactly. Sectors flagged sparse or unstable are reported at their plug-in value in the text.

Model	Sector	Source Counts	ESS	Plug-in (£bn)	Corr. (£bn)	DR (£bn)	Drop-10 (£bn)	Max share
C	Offices	415	393	4.10	+0.40	4.50	4.54	0.08
	Factories†	45	41	0.70	−0.05	0.65	0.72	0.29
	Warehouses‡	104	97	2.23	−0.93	1.31	2.42	0.19
	Covered total	—	—	**7.03**	−0.57	**6.46**	—	—
D	Offices	415	393	4.25	+0.26	4.51	4.61	0.09
	Factories†	45	41	0.72	−0.07	0.65	0.73	0.23
	Warehouses‡	104	97	2.31	−1.33	0.98	2.51	0.20
	Covered total	—	—	**7.03**	−0.89	**6.14**	—	—

† Sparse labelled support (ESS 41): plug-in reported. ‡ Unstable, top-heavy correction: plug-in reported as central, doubly robust as downward sensitivity.

Table S37: Full-pipeline bootstrap sampling intervals. Replicates resample labelled calibration source with replacement and re-runs the cross-fitted outcome model, the refitted inverse-odds transport weights and the functional-specific residual corrections. Register application frame is held fixed and costs represent the 2024-25 prices, reflecting sampling variability and being reported independently from structural sensitivities.

Functional	Median	95% interval	SD
Expected count (n)	94,880	94,432–95,358	245
Affected area (Mm^2)	57.8	50.7–62.5	2.9
Covered-sector cost (£bn)	6.54	5.59–7.24	0.42
Office cost (£bn)	4.52	4.12–4.92	0.21

with the application frame being fixed, the labelled source sample being resampled with replacement, and each replicate re-running the entire estimator. In particular, we cross-fitted outcome model, refitted inverse-odds transport weights, and the functional-specific residual corrections for the count, the affected area, the covered-sector cost total and the office cost, all over 1,000 replicates.

Sampling intervals are narrow relative to the structural ranges (Table S37). The count is sampling-stable within $\pm 0.5\%$ (95% interval 94,432–95,358). The affected area has a sampling standard deviation of 2.9 million m^2 (95% interval 50.7–62.5 million m^2, about $\pm 10\%$), smaller than the spread between the register-conditional (57.2), external-composition (43.9) and unweighted (64.9) area values, confirming that the affected-area uncertainty is dominated by structural composition and calibration choices rather than by sampling. The covered-sector cost has a 95% sampling interval of £5.6–7.2 billion, with a median of £6.54 billion. Office costs, covered sector with stable support, yields the most precisely determined at £4.12–4.92 billion.

Table S38: Expected aggregate threshold-crossing counts, expected affected floor area and expenditure-equivalent scale from the preferred model. Expected threshold-crossing counts (probability sums over eligible register entries) and expected affected floor area from the cross-fitted fuel-specific threshold-probability model, with the expenditure-equivalent scale obtained by applying the £25–£200/m^2 unit-cost grid to the expected affected floor area (£100/m^2 central column). Because out-of-fold area-weighted validation shows the unweighted full-register area is over-predicted and driven by a few large premises (1.19), we take the gate-passing area-calibrated model (Model C, adding log floor area, sector and fuel interactions; 1.20) as the preferred central value, report the composition-reweighted estimate as a lower alternative, and the unweighted full-register baseline as an over-predicting upper reference; the transparent fixed-accounting bridge and the policy-proxy subset are shown alongside. All figures are register-conditional area-to-cost scale conversions, not observed cost, avoided investment, savings, compliance cost, welfare or social net present value.

Frame	Threshold	Expected count	Area (Mm^2)	£bn (£100/m^2)	£bn range
Fixed-accounting bridge (baseline)	EPC-B (B/C)	80,565	55.0	5.50	1.38–11.00
	Floor (E/F)	51,342	17.7	1.77	0.44–3.55
Preferred, area-calibrated model	EPC-B (B/C)	95,074	65.43	**6.54**	1.64–13.09
(Model C; central)	Floor (E/F)	36,684	8.90	**0.89**	0.22–1.78
Composition-reweighted (lower alt.)	EPC-B (B/C)	86,811	50.24	5.02	1.26–10.05
(ND-NEED margins)	Floor (E/F)	41,894	11.56	1.16	0.29–2.31
Full register, baseline (upper ref.)	EPC-B (B/C)	94,334	78.29	7.83	1.96–15.66
(unweighted, over-predicts area)	Floor (E/F)	36,441	14.87	1.49	0.37–2.97
policy-proxy subset	EPC-B (B/C)	6,434	21.31	2.13	0.53–4.26
	Floor (E/F)	1,209	3.43	0.34	0.09–0.69

2 France DPE 2026 coefficient reform

The France analysis provides an external same-certificate test of factor-induced label reclassification under a rating architecture distinct from the England and Wales non-domestic EPC. All quantities derived from the ADEME microdata are certificate-flow counts and shares for records issued within the analysis window. The register contains lodgements generated by sales, rentals, reassessments and administrative updates and does not expose a stable dwelling identifier. The results therefore describe certificate flows rather than unique dwellings or the standing national stock. National dwelling-scale quantities are cited from official SDES/ONRE and government sources rather than inferred by scaling the ADEME flow.

2.1 Inclusion criteria and quality control

We use ADEME's open register of DPE certificates for existing dwellings under the post-July-2021 residential DPE methodology [24]. The analysed window contains certificates issued between 1 July 2025 and 30 June 2026, centred on the 1 January 2026 coefficient reform. The raw window contains 2,809,777 records. The France counterfactual requires a certificate-level decomposition of final energy by carrier, since the reform is implemented by changing only the electricity primary-energy conversion coefficient. Prior to recomputing the labels, we impose a carrier-reconciliation screen to ensure that the reported electricity component and the reported carrier slots were internally consistent with total final energy. For certificate i, let EF_i denote total final energy, $\mathrm{EF}_i^{\mathrm{elec}}$ the electricity final-energy component, and EF_{ik} the final-energy value reported in carrier slot k. We allowed a small reconciliation tolerance, $\tau_i = \max\{1.0 \text{ kWh},\ 0.005 \times \text{EF}_i\}$. The fixed 1.0 kWh component is deliberately conservative relative to the 0.1 kWh reporting precision in the ADEME data and prevents certificates from being rejected because of harmless rounding. The proportional component allows the same rule to remain scale-appropriate for high-consumption dwellings, where small percentage discrepancies can translate into larger absolute differences. A certificate was retained only if the electricity component was physically plausible, $0 \leq \text{EF}_i^{\text{elec}} \leq \text{EF}_i + \tau_i$, and if the sum of the reported carrier slots reconciled with total final energy,

$$\left| \sum_{k=1}^{3} \text{EF}_{ik} - \text{EF}_i \right| \leq \tau_i. \tag{S12}$$

The first condition rules out cases in which the electricity component alone exceeds the certificate's total final energy, up to the permitted rounding tolerance. The second condition is broader. It verifies that the carrier-level decomposition used to isolate electricity is consistent with the certificate-level total on which the DPE calculation is based. Table S39 reports the effect of the screen. The electricity-exceeds-total failures are empirically nested within the broader carrier-slot reconciliation failures. Consequently, although 1,549 records fail the electricity plausibility check, the unique number of excluded records is 74,790 rather than the sum of the two rows. The final recomputation sample contains 2,734,987 certificates, corresponding to 97.34% of the downloaded analysis window.

This exclusion has negligible bearing on the substantive estimates. Headline F/G shares and F/G exit rates are unchanged to one decimal place before and after the quality-control screen, removing implausible transition cells generated by inconsistent energy decompositions, thereby ensuring that movements across DPE classes are attributable to the electricity primary-energy coefficient change rather than to malformed carrier data. One residual F-to-D transition remains in the cleaned transition matrix. This certificate passes the carrier-reconciliation screen but has an inconsistent published primary-energy value, representing a single record among 2.73 million retained certificates and has no material effect on any reported estimate.

Table S39: Quality-control screen for the France DPE counterfactual sample. Before recomputing DPE classes, certificates are screened for internal consistency between reported total final energy and the carrier-specific final-energy fields used to isolate electricity. The screen removes records for which the same-certificate coefficient counterfactual is not well-defined, while retaining certificates whose discrepancies fall within a conservative rounding tolerance. Electricity-total plausibility failures are nested within the broader carrier-slot reconciliation failures, yielding a unique exclusion count of 74,790 certificates. All headline and robustness estimates use the resulting QC-cleaned sample.

Quantity	Records	Share of window (%)
Downloaded analysis window	2,809,777	100
Certificates retained for same-certificate recomputation	2,734,987	97.34
Records failing the carrier-slot reconciliation check	74,790	2.66
Unique records excluded after overlap	74,790	2.66
Records failing the electricity-total plausibility check	1,549	0.06

2.2 Data, legal change and analysis window

After applying the quality-control criteria in Section 2.1, the analytical sample contains 2,734,987 certificates. The legal intervention is narrowly defined. From 1 January 2026, the electricity final-to-primary-energy conversion coefficient used in the DPE falls from 2.3 to 1.9 [19–21]. We implement this change as a same-certificate counterfactual rather than a before versus after comparison across different dwellings. For each ADEME record, final energy, greenhouse-gas emissions, dwelling surface area and all other reported diagnostic inputs are held fixed, while only the electricity primary-energy coefficient changes. The analysis therefore isolates the mechanical effect of the coefficient revision on the DPE label.

Microdata results remain certificate-flow quantities. For national dwelling-scale quantities, we use official stock-reweighted estimates indicating that approximately 700,000–850,000 dwellings leave *passoire thermique* status under the reform [20, 21, 34]. We use ADEME's microdata to verify the same-certificate mechanism and to characterise the fuel, surface-area, temporal and geographical composition of the resulting reclassification.

2.3 Energy decomposition and same-certificate counterfactual

The ADEME register reports total five-use final energy, total five-use primary energy, and up to three final-energy carrier slots. Let EF_i denote total final energy for certificate i, and let EF_{ik} denote final energy in carrier slot $k \in \{1, 2, 3\}$. Electricity final energy is defined as

$$\mathrm{EF}_i^{\mathrm{elec}} = \sum_{k=1}^{3} \mathrm{EF}_{ik}\, \mathbf{1}\{s_{ik} = \text{electricity}\}, \tag{S13}$$

where s_{ik} is the reported carrier in slot k; this electricity final energy is denoted E_i^{elec} in the main text. We verified that pure-electric certificates in the window satisfy $P_i/\mathrm{EF}_i = 1.9$, confirming that the published ADEME primary-energy field P_i is already expressed using the post-reform coefficient. The pre-reform counterfactual is therefore obtained by adding back the coefficient difference on electricity:

$$P_i^{2.3} = P_i^{1.9} + 0.4\, E_i^{\mathrm{elec}}. \tag{S14}$$

This is the same certificate under two accounting regimes. Final energy, dwelling surface A_i, greenhouse-gas emissions and all diagnostic inputs are held fixed; only the electricity primary-energy coefficient changes. Primary-energy intensity is $p_i^r = P_i^r/A_i$ for $r \in \{1.9, 2.3\}$. The residential DPE label is assigned by the double-threshold rule. The final label is the worse of the primary-energy class and the greenhouse-gas class. We use the standard residential thresholds in Table S40 for dwellings of at least 40 m^2 and the official surface-dependent thresholds in Table S41 for dwellings below 40 m^2 [19]. A certificate is classified as a factor-revision exit if it is labelled F or G under the counterfactual 2.3 coefficient and E or better under the published 1.9 coefficient. Since the reform changes only the primary-energy coefficient for electricity, it can only move the energy axis. A certificate can therefore leave F/G only if the greenhouse-gas axis is not itself F or G. This implication is tested directly in Section 2.9.

2.4 Official small-surface thresholds and the reproduction gain

Dwellings with a reference surface below 40 m^2 located below 800 m altitude are assigned classes using surface-dependent cut-points rather than the standard thresholds, following the Arrêté du 25 mars 2024 [19]. The regulation tabulates both the primary-energy (CEP) and greenhouse-gas

Table S40: Standard residential DPE class thresholds used in the France counterfactual. The table reports the standard primary-energy and greenhouse-gas cut-points used to assign residential DPE classes for dwellings of at least 40 m^2. Each certificate receives a class on both axes, and the final DPE label is the worse of the two. Dwellings below 40 m^2 use the official surface-dependent cut-points in Table S41. Because the reform lowers only the electricity primary-energy coefficient, the same-certificate counterfactual can change the primary-energy axis but leaves the greenhouse-gas axis fixed. A certificate can therefore exit F/G only when its greenhouse-gas class is already E or better. GES is an abbreviation for greenhouse-gas (2.9).

Class	Primary energy, EP (kWh/m^2yr)	GES (kgCO_2eq/m^2yr)
A	EP $<$ 70	GES $<$ 6
B	70 $\leq$ EP $<$ 110	6 $\leq$ GES $<$ 11
C	110 $\leq$ EP $<$ 180	11 $\leq$ GES $<$ 30
D	180 $\leq$ EP $<$ 250	30 $\leq$ GES $<$ 50
E	250 $\leq$ EP $<$ 330	50 $\leq$ GES $<$ 70
F	330 $\leq$ EP $<$ 420	70 $\leq$ GES $<$ 100
G	EP $\geq$ 420	GES $\geq$ 100

(EGES) class boundaries at reference surfaces of 8, 9, 10, 15, 20, 25, 30 and 35 m^2, converging to the standard thresholds at 40 m^2, with linear interpolation for intermediate surfaces (Table S41). We implement this rule exactly: each certificate below 40 m^2 is classified against thresholds interpolated at its own surface, on both axes and under both coefficients. The register does not expose altitude, so we apply the below-800 m schedule that governs the overwhelming majority of the dwelling stock; this is the only material assumption in the small-surface implementation. Implementing the correction raises published-label reproduction for dwellings below 40 m^2 from 72.49% under standard thresholds to 98.46%, lifting overall reproduction from 93.27% to 97.93% while leaving the at-least-40 m^2 reproduction unchanged at 97.81% (Table S42). Since reproduction is now uniformly high across surface groups, the headline uses the full certificate flow rather than a surface subset.

2.5 Validation against ADEME-published labels

As an internal validation, we recomputed the post-reform 1.9 label and compared it with ADEME's published DPE label. With the official surface-dependent thresholds implemented (2.4), the recomputed 1.9 label matches the published label for 97.93% of QC-cleaned certificates, and for at least 97.8% within every surface group: 97.81% at 40 m^2 or above, 98.46% below 40 m^2 and 97.96% where surface is unrecorded (Table S42). Reproduction is also stable across the reform date, at 98.05% for certificates issued before 1 January 2026 and 97.81% for those issued from that date, and within each surface-by-issue-date cell (all $\geq$ 97.6%; Table S43). The residual 2.07% of mismatches is dominated by the greenhouse-gas axis and by near-threshold rounding rather than by the energy channel that drives exits (Table S44), and restricting to exactly reproduced certificates leaves the headline essentially unchanged (2.14). As validation resulted uniformly high, the headline certificate-flow result is reported for the full QC-cleaned window.

Table S41: Official surface-dependent DPE class thresholds for dwellings below 40 m^2 (Arrêté du 25 mars 2024, below 800 m altitude) [19]. Entries represent the upper bounds of each class on the primary-energy (CEP, kWh/m^2yr) and greenhouse-gas (EGES, kgCO_2eq/m^2yr) axes at each tabulated reference surface. Intermediate surfaces use linear interpolation, and at 40 m^2 the schedule coincides with the standard thresholds in Table S40. Smaller dwellings receive proportionally higher thresholds, principally because fixed domestic-hot-water demand is spread over less floor area.

	CEP upper bound (kWh/m^2yr)						**EGES upper bound (kgCO_2/m^2yr)**					
Surface (m^2)	**A**	**B**	**C**	**D**	**E**	**F**	**A**	**B**	**C**	**D**	**E**	**F**
≤ 8	146	186	386	505	622	739	11	16	44	68	90	122
9	134	174	355	464	574	685	11	16	42	65	87	118
10	124	164	329	428	533	640	10	15	40	62	84	115
15	100	140	263	333	421	514	8	13	36	56	76	107
20	88	128	230	300	385	476	8	13	34	54	74	104
25	81	121	210	280	363	454	7	12	32	52	73	103
30	76	116	197	267	349	439	7	12	32	52	72	102
35	73	113	188	258	338	428	7	12	31	51	71	101
40	70	110	180	250	330	420	6	11	30	50	70	100

2.6 Headline reclassification

Table S45 reports the core same-certificate counterfactual, using the official thresholds including the small-surface correction. Under the pre-reform coefficient of 2.3, 361,984 certificates are labelled F or G. Under the published 1.9 coefficient, the same certificates, with final energy and diagnostic inputs held fixed, contain 258,899 F/G certificates. The coefficient change therefore moves 103,085 certificates out of F/G in the full QC-cleaned flow, equal to 28.48% of pre-reform *passoires*. No certificates enter F/G. Because the recomputation is a deterministic transformation

Table S42: Validation of the France same-certificate label recomputation. The DPE class recomputed with the post-reform 1.9 electricity primary-energy coefficient is compared with ADEME's published DPE class for the same QC-cleaned certificates. The final column reports the match rate once the official surface-dependent small-surface thresholds are implemented (Table S41); the penultimate column reports the match rate under standard thresholds only. Surface-stratified rows are restricted to certificates with usable surface-area information. Implementing the official small-surface rule raises reproduction below 40 m^2 from 72.49% to 98.46% and makes reproduction uniformly high across surface groups.

Sample	**Certificates**	**Match rate standard (%)**	**Match rate small-surface (%)**
All QC-cleaned certificates	2,734,987	93.27	97.93
Surface area < 40 m^2	491,526	72.49	98.46
Surface area ≥ 40 m^2	2,133,311	97.81	97.81
Surface missing/unrecorded	110,150	97.96	97.96

Table S43: France label reproduction by surface group and issue-date group. Each cell reports the share of QC-cleaned certificates whose recomputed post-reform label (1.9), using the official surface-dependent thresholds, matches ADEME's published label. Reproduction is uniformly high ($\geq$ 97.6%) across every surface-by-issue-date cell, indicating the small-surface implementation and the 2.3 reconstruction behave consistently on both sides of the reform date.

Surface group	**Before 2026-01-01**	**From 2026-01-01**	**All**
< 40 m^2	98.34%	98.58%	98.46%
≥ 40 m^2	97.97%	97.64%	97.81%
Missing surface	98.23%	97.68%	97.96%
All	98.05%	97.81%	97.93%

Table S44: Audit of residual label mismatches in the France recomputation. Each QC-cleaned certificate whose recomputed post-reform label differs from the published label is assigned the first applicable reason. Missing-surface cases are recorded first, then greenhouse-gas-axis disagreements, then energy or greenhouse-gas intensities within 2% of a class boundary, then residual sub-40 m^2 cases, then unresolved energy-axis cases. The residual disagreement is dominated by the greenhouse-gas axis and by near-threshold rounding, neither of which is the electricity-driven energy channel that produces exits.

Mismatch reason	**Counts**	**% of mismatches**	**% of valid**
Greenhouse-gas (EGES) axis mismatch	41,898	74.17	1.532
Rounding/margin to threshold ($\leq$ 2%)	8,875	15.71	0.324
Missing/inconsistent surface	2,249	3.98	0.082
Energy axis / unresolved	2,105	3.73	0.077
Small-surface residual ($<$ 40 m^2)	1,361	2.41	0.050
Total mismatches	56,488	100.00	2.065

of the whole certificate flow rather than a sample from a population, and the register exposes no stable dwelling identifier, we report these counts as certificate-flow quantities and characterise uncertainty structurally, through the surface-subset, reproduced-label, issue-date, duplicate-record and threshold-margin analyses (2.14), rather than through a sampling confidence interval. A split by certificate issue date reproduces the published label at 98.05% before and 97.81% from 1 January 2026, indicating the published primary-energy field applies the post-reform coefficient consistently across issue dates.

Restricting to the 97.93% of certificates whose published label we reproduce exactly leaves 102,004 exits, 98.95% of the full-flow total, confirming the estimate is not an artefact of unreproduced labels. Excluding certificates explicitly observed below 40 m^2 yields 63,852 exits (24.55%), and restricting to certificates explicitly recorded at least 40 m^2 gives 57,681 exits (23.95%). Surface subsets and a conservative proxy deduplication bound the estimate from below without changing the mechanism, and are collected in the structural sensitivity table (Table S56).

2.7 Reproduction and exits by certificate issue date

The reform is effective for certificates issued from 1 January 2026, and official guidance states that certificates issued before 2026 remain valid and may be updated through a free attestation using the new coefficient [35]. If the published ADEME primary-energy field mixed pre-reform (2.3) and post-reform (1.9) bases by issue date, our reconstruction of the 2.3 counterfactual would be invalid for part of the window. Table S46 tests this directly. Splitting the QC-cleaned valid sample by issue date, our recomputation reproduces the published post-reform label for 98.05% of certificates issued before 1 January 2026 and 97.81% of those issued from that date. The similar reproduction rates before and after 1 January 2026 reduce concern that the result is driven by a discontinuity in how the published primary-energy field is constructed across the analysis window; the available fields do not establish provenance directly. Table S47, representing a finer monthly split, shows reproduction between 97.75% and 98.22% in every month, with the electric share and small-surface share of certificates stable throughout. Exit rate among passoires steps up modestly for certificates issued from January 2026, consistent with a certificate flow whose composition shifts slightly towards electrically heated dwellings after the reform rather than with a change in field construction. On field provenance, the register exposes a single published primary-energy field, which pure-electric certificates confirm is expressed at the 1.9 coefficient, together with the electricity final-energy carriers used to reconstruct the 2.3 counterfactual and a method-version field. It further contains no attestation, replacement or superseded-certificate flag and no stable dwelling identifier, only department- and region-level geography. Provenance is therefore assessed indirectly through the issue-date reproduction rates rather than established directly. Exits divide as 44,126 among pre-2026 certificates and 58,959 among post-2026 certificates, consistent with a certificate flow whose passoire composition is somewhat more electric after the reform date.

2.8 Old-label to new-label transition matrix

Table S48 gives the old-label to new-label transition matrix. Rows are labels under the counterfactual 2.3 coefficient and columns are labels under the published 1.9 coefficient. Because the reform lowers the electricity primary-energy coefficient, it cannot make a certificate worse. The matrix is therefore lower triangular, and the few multi-band improvements are immaterial. Nearly all F/G exits are F-to-E transitions, with G certificates rarely exiting due to the 0.4 coefficient reduction being insufficient to lift G records above the F/G threshold. The transition matrix policy-relevant effect is concentrated at the F/E boundary, where electrically exposed F certificates fall below the F threshold once electricity is converted at 1.9 rather than 2.3.

Table S45: Headline France DPE factor-revision counterfactual after quality control. The same QC-cleaned certificates are classified under the pre-reform 2.3 electricity primary-energy coefficient and the post-reform 1.9 coefficient. Differences between the two columns therefore isolate the mechanical effect of the coefficient revision, holding the underlying certificate record fixed.

Quantity	**Coefficient 2.3**	**Coefficient 1.9**
F/G certificates	361,984	258,899
F/G share of QC-cleaned certificate flow	13.24%	9.47%
Certificates exiting F/G		103,085
Exit share of pre-reform *passoires*		28.48%
Certificates entering F/G		0

Table S46: France DPE reproduction and exits by certificate issue date. QC-cleaned valid certificates split by issue date relative to the 1 January 2026 reform date. Published-label reproduction is the share for which our recomputed post-reform (1.9) label matches the register's published label; a similar rate on both sides indicates the published field applies the post-reform coefficient consistently, so the reconstructed 2.3 counterfactual is like-for-like. Exits are certificates moving from F/G under 2.3 to E or better under 1.9.

Issue period	Counts	Pre-reform F/G	Exits	Label reproduction
Issued before 2026-01-01	1,428,742	172,691	44,126	98.05%
Issued from 2026-01-01	1,306,245	189,293	58,959	97.81%

Table S47: France reproduction, exits and composition by calendar month of issue. QC-cleaned valid certificates binned by month of date of DPE establishment. Reproduction is the surface-adjusted published-label match rate. Exit rate represents exits as a share of pre-reform *passoires*, 40 m^2 (%) representing the share of certificates being electrically heated.

Month	Counts	Reprod. (%)	Passoires	Exits	Exit/*passoires* (%)	Electric (%)	<40 m^2 (%)	Unrec. surf. (%)
2025-07	268,918	98.04	33,196	8,278	24.94	40.5	20.4	3.5
2025-08	181,753	97.89	20,671	5,253	25.41	39.6	19.2	3.4
2025-09	266,666	98.22	31,280	7,930	25.35	39.9	17.4	3.8
2025-10	262,878	97.99	32,667	8,464	25.91	39.8	17.3	4.1
2025-11	237,742	98.16	29,948	7,901	26.38	37.8	17.0	4.2
2025-12	210,785	97.93	24,929	6,300	25.27	36.8	18.2	4.6
2026-01	247,723	97.75	33,818	11,106	32.84	42.4	17.5	4.2
2026-02	242,198	97.85	34,851	10,820	31.05	41.8	18.0	4.6
2026-03	259,557	97.76	37,748	11,654	30.87	43.6	17.0	4.2
2026-04	238,580	97.83	35,947	10,935	30.42	43.8	17.7	4.0
2026-05	204,327	97.88	30,329	9,344	30.81	42.9	18.4	3.8
2026-06	113,860	97.83	16,600	5,100	30.72	45.3	18.0	3.7

2.9 Greenhouse-gas axis binding

The residential DPE label is the worse of the primary-energy and greenhouse-gas classes. Since the coefficient change affects only the primary-energy axis, any certificate exiting F/G must be one whose F/G status was caused by the energy axis, not by the emissions axis. Table S49 verifies this implication directly, and Table S50 states it as a binding-criterion partition. With the surface-adjusted greenhouse-gas (GES) thresholds implemented, every one of the 103,085 exiters has a non-binding GES class (A–E), and all 103,085 exits are energy-limited with none occuring where the emissions axis binds. Conversely, 126,751 pre-reform *passoires* remain F/G precisely because their GES axis is F or G, which the coefficient revision cannot touch. Table S51 reports the published GES class of exiters as an independent face-validity check with 99.7% carry a published GES class of E or better, and the residual 289 certificates (0.28% of exiters) with a published GES-F label reflect minor boundary differences between our surface-adjusted GES classification and the published one, not emissions-axis exits. The pattern is shows how lowering the electricity primary-energy coefficient can improve the primary-energy class but cannot improve the emissions class. As a result, exits occur only where the primary-energy criterion is relaxed and the greenhouse-gas criterion is not binding.

2.10 Window robustness

Table S52 repeats the same-certificate counterfactual in narrower symmetric windows around 1 January 2026. The full analysis window is (± 6) months. The robustness checks use (± 3) months and (± 1.5) months. The F/G exit share among pre-reform *passoires* remains between 28.48% and 29.21%, and the electric share of exits remains approximately 89–90% in all windows. The stability of Table S52 is important because the ADEME register is a certificate flow. Narrowing the window changes the composition and size of the sample, but fails to alter the mechanism. Roughly 28–29% of pre-reform *passoires* in the certificate flow exit F/G, remaining overwhelmingly electric.

2.11 Heating-energy decomposition and negative controls

Table 3 in the main text decomposes exits by the reported main heating energy. Electrically heated certificates are 41.0% of the QC-cleaned certificate flow but account for 89.1% of all exits. Their exit rate among pre-reform *passoires* is 48.36%, compared with 7.74% for gas, 12.33% for wood or biomass, 15.04% for heat networks and 1.91% for oil and other fossil-heated certificates. This is the

Table S48: Transition matrix for the France DPE factor-revision counterfactual. Rows classify each QC-cleaned certificate under the pre-reform 2.3 electricity primary-energy coefficient; columns classify the same certificate under the post-reform 1.9 coefficient. Entries are certificate counts. Diagonal cells are certificates whose DPE class is unchanged, while off-diagonal cells to the left of the diagonal are certificates mechanically reclassified into a better class when only the electricity primary-energy coefficient is lowered. The absence of rightward off-diagonal mass reflects the monotonic structure of the reform: holding the certificate record fixed, the coefficient revision can improve or leave unchanged the energy axis, but cannot worsen it.

Label under 2.3	**A**	**B**	**C**	**D**	**E**	**F**	**G**
A	35,787	0	0	0	0	0	0
B	19,890	103,941	0	0	0	0	0
C	0	64,631	811,516	0	0	0	0
D	0	0	208,063	631,604	0	0	0
E	0	0	0	193,911	303,660	0	0
F	0	0	0	1	103,033	122,164	0
G	0	0	0	0	51	47,491	89,244

Table S49: F/G exits by greenhouse-gas binding status in the France DPE factor-revision counterfactual. Pre-reform F/G certificates are partitioned according to whether the greenhouse-gas axis permits an exit from F/G. Certificates with a GES class of A–E are F/G because of the primary-energy axis and can therefore leave F/G when the electricity primary-energy coefficient is lowered from 2.3 to 1.9. Certificates with a GES class of F or G remain constrained by the emissions axis, so the same coefficient revision cannot move them out of the final F/G category.

GES binding status	**Pre-reform F/G certificates**	**Exits**	**Exit rate (%)**
A–E (primary-energy axis binding)	235,233	103,085	43.82
F/G (emissions axis binding or co-binding)	126,751	0	0

Table S50: Binding-criterion partition of pre-reform France *passoires*. Pre-reform *passoires* (2.3) are partitioned by which axis places them in F/G, using the surface-adjusted classes. Energy-limited *passoires* are F/G on the primary-energy axis with a non-binding greenhouse-gas axis, representing the only certificates that can exit when the electricity coefficient falls. Every exit falls in this row, and no certificate whose greenhouse-gas axis binds exits, confirming that exits occur where the primary-energy criterion is relaxed and the greenhouse-gas criterion is not binding.

Binding axis (2.3 basis)	**Passoires**	**Exits**	**Exit rate (%)**
Energy-limited (energy F/G, GES not)	235,233	103,085	43.82
GES-limited (GES F/G, energy not)	27,574	0	0.00
Both axes F/G	99,177	0	0.00

Table S51: Published greenhouse-gas class among certificates exiting F/G in the France DPE counterfactual. Rows report ADEME's published GES class for certificates that move from F/G under the pre-reform 2.3 electricity primary-energy coefficient to E or better under the post-reform 1.9 coefficient. Because the counterfactual changes only the electricity primary-energy coefficient, the greenhouse-gas axis is held fixed. Exits should therefore occur only when the emissions class is E or better under the applicable surface-adjusted thresholds. The small residual published-F count reflects boundary differences between the reconstructed surface-adjusted GES class and the published class.

Published GES class	**Exiting certificates**	**Share of exiters (%)**
A	2	0.00
B	15,085	14.63
C	79,045	76.68
D	2,847	2.76
E	5,817	5.64
F	289	0.28
G	0	0

Table S52: Robustness of France DPE F/G exits to alternative certificate windows around the 1 January 2026 reform. Each row repeats the same QC-cleaned same-certificate recomputation on a symmetric calendar window centred on the reform date. Certificates are classified once under the pre-reform 2.3 electricity primary-energy coefficient and once under the post-reform 1.9 coefficient. Exits are records moving from F/G to E or better when only the coefficient is changed. The stability of exits as a share of the retained certificate flow, exits among pre-reform F/G certificates, and the electric-exit share shows that the headline estimate is not driven by the particular one-year window used in the main specification.

Window (months)	**Start**	**End**	**Certificates**	**F/G 2.3**	**F/G 1.9**	**Exits**	**Exits/certificates**	**Exits/*passoires***	**Electric exits**
± 6	2025-07-01	2026-06-30	2,734,987	13.24%	9.47%	103,085	3.769%	28.48%	89.1%
± 3	2025-10-01	2026-03-31	1,460,883	13.28%	9.43%	56,245	3.850%	29.00%	89.2%
± 1.5	2025-11-15	2026-02-15	709,591	13.04%	9.23%	27,019	3.808%	29.21%	90.0%

French analogue of the England and Wales electricity-exposure measure, using directly observed DPE energy-carrier data.

We complement the decomposition with three falsification checks (Table S53). First, we confirm that non-electrically heated *passoires* exit at much lower rates than electric ones if the mechanism is the electricity coefficient (6.54% versus 48.36%). Second, *passoires* whose greenhouse-gas axis binds should essentially never exit. Third, *passoires* far from the F/E primary-energy boundary should be less sensitive than those near it as the exit rate among *passoires* within 20 kWh/m^{2}yr of the surface-adjusted F/E cut-point is 55.55%, against 22.27% for those further away.

Table S53: Negative controls for the France factor-revision mechanism. Each pair contrasts a group the electricity-coefficient mechanism predicts should exit at a high rate with a group it predicts should exit at a low rate or not at all. The near/far split uses distance from the surface-adjusted F/E primary-energy cut-point under the 2.3 coefficient. Exit rate is exits as a share of the row's pre-reform *passoires*.

Control	Passoires	Exits	Exit rate (%)
(A) Electric-heated *passoires*	189,920	91,836	48.36
Non-electric *passoires* (expect low)	172,064	11,249	6.54
(B) GES not binding (can exit)	235,233	103,085	43.82
GES-binding *passoires* (expect ∼0)	126,751	0	0.00
(C) Near F/E boundary ≤ 20 (expect high)	67,492	37,493	55.55
Far from F/E boundary > 20 (expect low)	294,492	65,592	22.27

2.12 Surface robustness

Table S54 reports surface-band heterogeneity. Because the official small-surface thresholds are now implemented and validated (98.46% reproduction below 40 m^2; Table S42), the sub-40 m^2 group is a validated part of the headline, shown separately to document surface heterogeneity. Small dwellings exit at a higher rate (38.49% of their pre-reform *passoires*) due to being disproportionately electrically heated, and the exit rate falls monotonically across the larger surface bands. As lower bounds that deliberately discard the smallest dwellings, excluding certificates explicitly observed below 40 m^2 gives 63,852 exits (24.55% of pre-reform *passoires*) and restricting to certificates explicitly recorded at least 40 m^2 gives 57,681 exits (23.95%), with both retaining the same electricity-driven mechanism.

2.13 Department-level heterogeneity

Department-level results are used only as a heterogeneity check. They should not be interpreted as stock-reweighted regional estimates, because the ADEME register is a certificate flow. Table S55 reports the twelve departments with the largest number of exits, showing that the result is not confined to a single department or region. Large exit counts appear in departments with large certificate volumes and high electric-heating exposure, while the electricity-driven composition remains visible across the largest-exit departments.

2.14 Structural sensitivity and estimate hierarchy

Because the recomputation is deterministic over the certificate flow and the register exposes no stable dwelling identifier, we characterise uncertainty structurally rather than through sampling confidence intervals. Table S56 collects the headline and every sensitivity estimate. The main estimate is

Table S54: Surface-band robustness of F/G exits in the France DPE factor-revision counterfactual. Rows report QC-cleaned certificate-flow quantities classified under the pre-reform 2.3 and post-reform 1.9 electricity primary-energy coefficients. Pre-reform F/G certificates are certificates classified as F or G under the 2.3 coefficient. Exits are certificates reclassified to E or better under the 1.9 coefficient, using the official surface-adjusted thresholds throughout. The lower-bound line excludes certificates explicitly observed below 40 m^2; certificates with missing or unclassified surface area are retained rather than imputed. Rows below the midrule decompose the observed at-least-40 m^2 subset.

Surface group	Certificates	Pre-reform *passoires*	Exits	Exit rate among *passoires*	Share of all exits
Full QC-cleaned sample	2,734,987	361,984	103,085	28.48%	100.00%
Observed (<40 m^2)	491,526	101,922	39,233	38.49%	38.06%
Excluding observed (<40 m^2)	2,243,461	260,062	63,852	24.55%	61.94%
Observed ($\geq$ 40 m^2)	2,133,311	240,803	57,681	23.95%	55.95%
Surface missing or unclassified	110,150	19,259	6,171	32.04%	5.99%
40–60 m^2	588,639	70,276	22,447	31.94%	21.78%
60–90 m^2	927,125	83,745	19,502	23.29%	18.92%
90–130 m^2	397,929	55,495	10,687	19.26%	10.37%
($\geq$ 130) m^2	219,618	31,287	5,045	16.12%	4.89%

Table S55: Departments contributing most to F/G exits in the France DPE factor-revision counterfactual. Departments are ranked by the number of QC-cleaned certificate records moving from F/G under the pre-reform 2.3 electricity primary-energy coefficient to E or better under the post-reform 1.9 coefficient. Counts are certificate-flow quantities from the downloaded register window, not stock-reweighted estimates of departmental dwelling stocks. Exit rate is the share of pre-reform F/G certificates in each department that leave F/G. Share of all exits uses the full-flow denominator of 103,085 exits. Electric exits report the share of each department's exits among certificates recorded as electrically heated.

Department	Certificates	Pre-reform *passoires*	Exits	Exit rate among *passoires*	Share of all exits	Electric exits
75	143,609	41,223	13,278	32.21%	12.88%	91.3%
92	79,420	13,574	3,255	23.98%	3.16%	84.6%
59	91,108	10,646	3,060	28.74%	2.97%	84.3%
69	81,896	8,762	2,930	33.44%	2.84%	88.3%
93	84,272	10,780	2,455	22.77%	2.38%	86.2%
13	86,322	4,824	2,372	49.17%	2.30%	94.7%
94	57,622	10,015	2,311	23.08%	2.24%	85.7%
38	48,841	6,792	2,104	30.98%	2.04%	89.9%
76	42,410	6,773	1,993	29.43%	1.93%	87.0%
44	52,810	5,643	1,942	34.41%	1.88%	91.8%
74	32,343	7,298	1,903	26.08%	1.85%	92.7%
78	53,419	7,121	1,819	25.54%	1.76%	90.2%

the full certificate flow (103,085 exits, 28.48% of pre-reform *passoires*). Restricting to certificates whose published label we reproduce exactly retains 102,004 exits (98.95% of the main estimate), so the headline is not an artefact of the residual 2% of unreproduced labels. Surface subsets that deliberately discard the smallest dwellings ($\geq$ 40 m^2: 57,681; excluding observed < 40 m^2: 63,852) and the proxy-deduplicated flow (84,407) yield lower counts than the full-flow estimate, while the narrower issue-date windows return very similar exit rates. The superseded standard-threshold recomputation, which mis-scores small dwellings, is shown for reference. Across all specifications the exit rate among pre-reform *passoires* stays in a narrow 22.9–29.2% band and the electricity-exposure pattern remains qualitatively stable.

Table S56: France structural sensitivity and estimate hierarchy. Each row applies a different structural restriction to the same-certificate counterfactual. Exit rate is exits as a share of that row's pre-reform *passoires*. The main estimate represents the full certificate flow with official surface-adjusted thresholds. Robustness rows provide lower-bound or alternative-flow checks, and the standard-threshold row is superseded and shown only for comparison.

Estimate	Role	Certificates	Passoires	Exits	Exit rate (%)
Full certificate flow (surface-adjusted)	Main	2,734,987	361,984	103,085	28.48
Reproduced-label-only certificates	Robustness	2,678,499	354,035	102,004	28.81
Excluding observed < 40 m^2	Robustness	2,243,461	260,062	63,852	24.55
Observed ≥ 40 m^2 only	Robustness	2,133,311	240,803	57,681	23.95
Proxy-deduplicated flow	Robustness	1,903,380	320,391	84,407	26.34
$\pm$3-month window	Robustness	1,460,883	193,961	56,245	29.00
$\pm$1.5-month window	Robustness	709,591	92,514	27,019	29.21
Standard thresholds (superseded)	Reference	2,734,987	399,363	112,495	28.17

Threshold-margin and rounding robustness. A concern for any threshold-crossing count is that it is driven by certificates sitting arbitrarily close to a class boundary. Table S57 excludes *passoires* whose primary-energy intensity lies within 1, 2, 5 or 10 kWh/m^{2}yr of the surface-adjusted F/E cut-point on either the 1.9 or the 2.3 basis, and separately reclassifies after rounding every intensity to the nearest integer. Excluding a $\pm$1 kWh/m^{2}yr band around the boundary retains 94.9% of exits, and even a wide $\pm$10 kWh/m^{2}yr exclusion retains 62.3%; integer rounding changes the count by 0.6%. The result is therefore not a knife-edge rounding artefact.

Table S57: France threshold-margin and rounding sensitivity. Passoires within the stated margin of the surface-adjusted F/E primary-energy cut-point are excluded before exits are recomputed. Exits retained (%) is calculated relative to the full-flow count of 103,085.

Rule	Excluded	Exits	Exits retained (%)
No exclusion (full flow)	0	103,085	100.00
Exclude within 1 kWh/m^{2}yr of F/E	14,308	97,841	94.91
Exclude within 2 kWh/m^{2}yr of F/E	26,845	93,332	90.54
Exclude within 5 kWh/m^{2}yr of F/E	60,532	81,379	78.94
Exclude within 10 kWh/m^{2}yr of F/E	112,622	64,219	62.30
Round intensities to nearest integer	0	102,420	99.35

Duplicate and repeated-dwelling sensitivity. The ADEME open window provides only a per-certificate identifier (`numero_dpe`) and department-level geography, with no stable dwelling identifier, so exact dwelling-level deduplication is not possible and the France analysis is a certificate-flow analysis. As a conservative proxy we collapse certificates that share department, rounded surface, rounded primary-energy and greenhouse-gas intensities, heating energy and published label, keeping the latest certificate per proxy dwelling (and, as a check, one at random). This proxy treats 30.4% of certificates as potential repeats, an upper bound because distinct dwellings can share all six attributes. Even so (Table S58), exits fall only to 84,407 (keep-latest) or 84,359 (keep-random), the exit rate among *passoires* stays at 26.3%, and the share of exits arising in duplicated proxies (28.7%) is close to the overall duplicate share, indicating exits are not concentrated in repeat lodgements.

Proxy deduplication lowers the exit count but leaves the exit rate and electricity-exposure pattern broadly similar, indicating that repeated proxy records do not drive the qualitative result.

Table S58: France duplicate / repeated-dwelling sensitivity. The register has no stable dwelling identifier; rows apply a conservative proxy deduplication on department, rounded surface, rounded primary-energy and greenhouse-gas intensities, heating energy and published label. Exit rate is exits as a share of the row's pre-reform *passoires*.

Basis	Counts	Exits	Exit rate (%)
Certificate flow (all certificates)	2,734,987	103,085	28.48
Proxy-dedup (keep latest per proxy)	1,903,380	84,407	26.34
Proxy-dedup (keep one random per proxy)	1,903,380	84,359	26.33
Proxy-dedup + $\geq$ 40 m^2	1,509,432	50,894	22.85

Pre-2026 reconstruction validity. For certificates issued before 1 January 2026, the published primary-energy field is the same 1.9-based field, and the 2.3 counterfactual is reconstructed from the same electricity final-energy carriers used throughout; there is no separately archived 2.3 primary-energy field to compare against. We further recomputed the published 1.9 label for pre-2026 certificates as an additional validity check, yielding reproduction is 97.97% for dwellings at least 40 m^2, 98.34% below 40 m^2 and 98.23% with missing surface, and 96.9–99.3% across heating fuels, and the reconstructed old-basis (2.3) F/G set contains the published F/G set. The pre-2026 reconstruction is therefore a like-for-like counterfactual.

2.15 National-scale interpretation

The France certificate-flow counterfactual should not be scaled directly into a national dwelling-stock estimate. In the QC-cleaned ADEME window, 28.48% of pre-reform *passoires* exit F/G, whereas official stock-reweighted estimates imply a lower national exit rate. This difference is expected because recent certificate flows over-represent the electrically heated dwellings most exposed to the coefficient change. We therefore use official national sources for dwelling headcounts. These sources report approximately 700,000 fewer *passoires* in the 1 January 2025 main-residence stock and approximately 850,000 under an earlier government estimate based on the larger 1 January 2023 stock [20, 21, 34]. The role of the ADEME microdata analysis is instead to establish the mechanism at certificate level: holding final energy and all diagnostic inputs fixed, changing only the electricity primary-energy coefficient moves a large number of certificates across the F/G threshold, overwhelmingly among electricity-exposed certificates whose greenhouse-gas axis is not binding.