

An Open Firm-Level Microsimulation of the UK VAT Registration Threshold^{*}

Vahid Ahmadi[†]

PolicyEngine

Preliminary and Incomplete. Comments are welcome.

July 2026

Abstract

I present an open, reproducible firm-level microsimulation for costing UK VAT registration-threshold reforms, released with its synthetic-data generator. The threshold is a *notch*: once annual turnover crosses £85,000, VAT falls due on a firm’s whole base, not the excess. I build synthetic firm records from ONS registered-business counts, calibrate them to 2023–24 HMRC statistics—with the near-threshold density pinned to the within-side geometry of the OBR’s published £1,000-band bunching data—and cost *level*, *shape* (a taper) and *rate* (a reduced-rate band) reforms statically on a common £184.8bn base: raising the threshold to £100,000 costs £753m, a graduated taper over £85,000–£105,000 costs £520m, and 10% and 15% reduced-rate bands cost £484m and £242m. A static costing of the April 2024 rise to £90,000 is consistent with HMRC’s published profile under a voluntary-retention convention (£184m against £185m in 2025–26) and reproduces its narrowing and 2028–29 sign change; the microdata are calibrated to HMRC aggregates, so this is a consistency check, not validation. I characterise the notch by its *dominated region*—the exact £21,250 band above the threshold in which no firm profits from locating—and how each reform changes it: a level rise relocates it, a reduced rate splits it into two bands of essentially unchanged total width, a taper removes it. Conditional on an assumed turnover elasticity, intensive-margin responses barely move these costings: a pure level rise has exactly zero intensive-margin offset, and reduced-rate offsets stay under 6%. The synthetic near-threshold shape is target-inherited, and provably so: the estimator reads back the excess mass the OBR targets put in at £85,000, a placebo that regenerates without those targets returns zero, and the uncalibrated £90,000 vintage shows a band-edge artifact that lands within 10% of the

^{*}**Acknowledgements:** I thank Max Ghenis and Seyed Mahdi Hosseini for their great suggestions and helpful discussions.

[†]Research Associate at PolicyEngine. Email: vahid@policyengine.org

administrative bunching estimate—a coincidence of band geometry, unstable across specifications—so band-calibrated synthetic data support no bunching inference in either direction.

Keywords: value-added tax, microsimulation, tax notch, registration threshold

1 Introduction

The United Kingdom’s Value Added Tax (VAT) applies above a registration threshold that is the highest in the OECD alongside Switzerland (Seely, 2024; HM Revenue and Customs, 2024). The threshold is a *notch* rather than a kink: once a firm’s annual taxable turnover crosses it, VAT falls due on the firm’s whole base, not only on the part above the threshold, so registration produces a discrete jump in liability rather than a marginal one. The threshold falls in a dense part of the firm distribution. Figure 1(a) shows that VAT- and PAYE-registered UK businesses are concentrated at the lower end of the turnover distribution, and Figure 1(b) shows that the VAT-registered population per £1,000 of turnover is densest at and below the threshold, so the notch applies to a large number of firms.

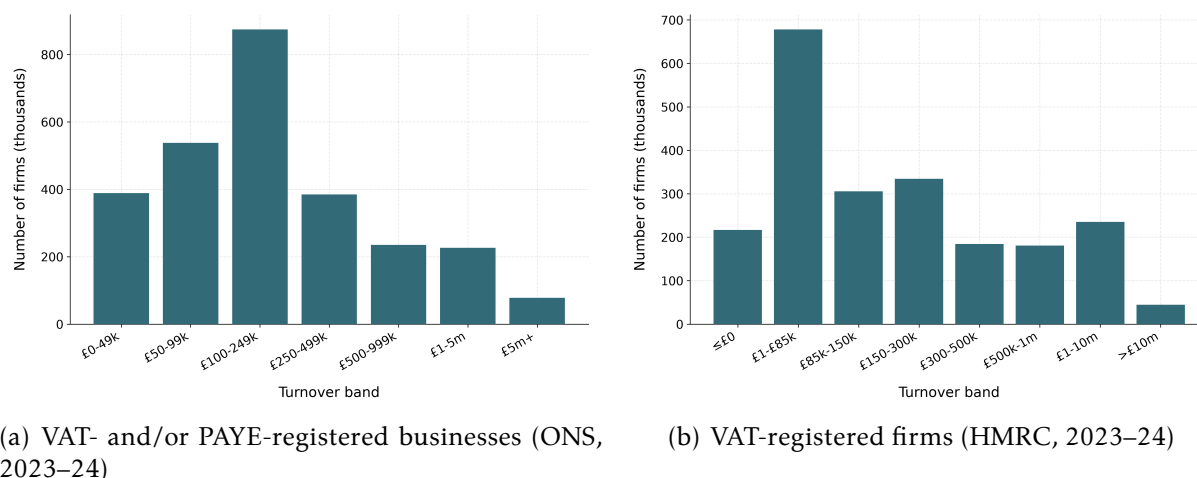


Figure 1: The UK firm-size distribution and VAT registration, 2023–24 (threshold £85,000). Panel (a): number of businesses by turnover band in the ONS business register, which covers businesses registered for VAT and/or PAYE (about 2.7 million); the full business population including unregistered businesses is roughly twice as large (DBT business population estimates). Panel (b): number of VAT-registered firms by turnover band. Band widths are unequal, so cross-band count comparisons should be read per £1,000 of turnover. Source: panel (a), ONS, *UK Business: Activity, Size and Location* (business counts by turnover band); panel (b), HMRC, *Value Added Tax (VAT) Annual Statistics* (VAT-registered population by turnover band).

Firms respond to the notch by locating below the threshold. Tabulations of HMRC turnover data show the number of firms rising towards £84,000 and then dropping sharply at the £85,000 threshold, leaving excess mass just below it (Office for Budget Responsibility, 2023). Liu et al. (2021) document this bunching in UK administrative firm data over 2004–05 to 2014–15, when the threshold rose from £58,000 to £81,000, and Liu et al. (2024) find that firms slow their turnover growth as they approach the

threshold. The notch therefore distorts the size distribution and the growth decisions of firms near the margin, which places the threshold under recurrent pressure for reform.

Reform proposals change the threshold in one of three ways: its *level* (raising or lowering it), its *shape* (replacing the notch with a schedule that phases the rate in over a band), or its *rate* (a reduced rate for firms in a band above the threshold). The three are not equivalent. Changing the level moves the notch along the turnover axis; changing the shape or the rate alters the size of the discrete jump itself. Costing the three on a common basis is the practical difficulty this paper addresses. A change in the level can be costed by mechanical accounting, but a change in the shape of the schedule cannot be evaluated with the reduced-form bunching methods used to study the threshold, because a continuous schedule leaves no single point of excess mass to exploit. The firm-level data needed to study turnover near the threshold is, in addition, not available in open form.

In this paper I build an open, reproducible firm-level microsimulation for costing UK VAT threshold reforms on a common static basis. I construct synthetic firm microdata calibrated to published official aggregates—HMRC registered-firm counts by turnover band and by sector, net VAT liability by turnover band, and ONS business counts and employment bands—for the 2023–24 tax year, when the threshold was £85,000, and separately for 2024–25, when it rose to £90,000. Each firm’s net VAT liability is the standard rate applied to its value added. The synthetic data can be released openly and regenerated under any counterfactual schedule.

On these data I do four things. First, I cost level reforms and schedule reforms—a graduated taper and a reduced-rate band—statically, by applying each counterfactual schedule mechanically to the firm distribution. Second, I characterise the notch’s distortion through its *dominated region*—the exact, elasticity-free range of turnover above the threshold in which no firm profits from locating—and how each reform changes it. Third, I price the level and rate reforms behaviourally: each firm re-optimises its turnover within its schedule region, conditional on an assumed turnover elasticity e that the data do not identify, reported as an e -sensitivity range that nests the static costing exactly in its $e \rightarrow 0$ limit. Fourth, I show that the synthetic data’s near-threshold structure is target-inherited and cannot support bunching inference: the £85,000 profile is calibrated to the OBR’s published £1,000-band data and the estimator reads exactly that back, a placebo without the fine targets returns zero, and the uncalibrated £90,000 vintage shows a spurious band-edge signal.

Four findings follow. First, the static pipeline prices all the reforms on a common basis, and its anchor costing tracks the Government’s published number once the deregistration convention is stated. My static costing of the April 2024 increase from £85,000 to £90,000 depends on what released firms are assumed to do: if every firm in the band deregisters, the 2025–26 cost is £323m, overshooting HMRC’s published £185m (HM Treasury, 2024) in line with the official costing’s phased registration response; if the roughly 43% voluntary-registration share that Liu et al. (2021) document is retained, it is £184m, in line with the published figure—agreement finer than the model’s liability-level calibration error supports reading as precision (Section 4). Both conventions re-

produce the official profile’s narrowing and its 2028–29 sign change. Forward moves from the £90,000 base range from a gain of about £1,040m at £70,000 to a cost of about £1,450m at £120,000, against a base of about £200.9bn in 2025–26.

Second, the shape and rate reforms act on the notch’s distortion in a way that a change in level does not. The notch creates a *dominated region* of width $a = T^*\tau/(1-\tau) = £21,250$ above the threshold (£85,000–£106,250), a range of turnover in which no firm can locate without being strictly worse off than at the threshold. This region is an exact property of the schedule: under the value-added formulation of Section 7 its width is the same for every firm with positive value added, whatever its input share, and it requires no behavioural calibration. It also spans a populous part of the firm distribution—about 148,000 weighted firms in the calibrated data—so its width is an economically meaningful distortion index. Raising the threshold relocates this region without removing it. A reduced rate does *not* remove it: the jump at the threshold falls (from £17,000 to £12,750 at a 15% band), but the band reverts to the standard rate at its £105,000 top, introducing a second notch and a second dominated region, so the total dominated width is essentially unchanged—£21,562 at 15% and £22,569 at 10%, both marginally above the £21,250 baseline. Only the graduated taper, being continuous, removes the dominated region. On the common £85,000 data-year base, raising the threshold to £100,000 costs £753m and removes none of the dominated region; the taper removes all of it for £520m; and the 10% and 15% reduced rates cost £484m and £242m while splitting the dominated band in two. On the dominated-region metric the level move is the most expensive option per unit of distortion removed—a comparison that indexes the misallocation distortion only and omits the compliance-cost saving a higher threshold delivers (Section 8).

Third, conditional on the assumed elasticity, intensive-margin responses barely move these costings. A pure level rise has *exactly* zero intensive-margin revenue offset at every e : released firms leave the VAT base, so their expansion is untaxed, and no other firm’s effective rate changes—a result the simulator verifies to £0.1m. For the reduced-rate bands the offset is second order: at $e = 0.17$ the 10% band’s cost falls from £484m to £475m and the 15% band’s from £242m to £235m, under 6% even at $e = 0.32$. I do not report a behavioural cost for the graduated taper: its rate varies continuously with turnover, so its marginal-rate channel is outside the flat-rate machinery, and only its static cost and dominated-region removal are reported. The sweep values $e \in \{0.05, 0.17, 0.32\}$ are assumptions: the lower end has an external anchor (Kleven and Waseem, 2013), and the published UK VAT-base estimates of Liu et al. (2021) fall between it and the mid-point. The consequence of these results runs opposite to the usual worry: the static costings are robust to the intensive margin, and the behavioural uncertainty that matters—registration and location choice at the notch—is exactly the margin the synthetic data cannot identify.

Fourth, the open data carry a transparency lesson about bunching: every sub-band feature of band-calibrated synthetic data is exactly what its targets put there, and this paper demonstrates the point in three directions. The 2023–24 population’s near-threshold density is calibrated to the OBR’s published £1,000-band profile (Office for Budget Re-

sponsibility, 2023), so it reproduces the administratively observed bunching shape—and the estimator reads that shape back as excess mass ($E = 7,933$), which is target inheritance, not behaviour. A placebo that regenerates the population *without* the fine targets returns $E = 0$: remove the targets and the “bunching” vanishes. And the 2024–25 vintage, for which no fine bands are published, shows what the trap looks like from the other side: the coarse £90,000 band edge alone produces an excess-bunching statistic within 10% of the administrative estimate of Liu et al. (2021)—a coincidence of band geometry, four-fold unstable across estimator specifications, with no behaviour behind it. An earlier draft of this paper reported “reproducing” the administrative step when a liability mis-scaling in the generator manufactured it; the correction removed it, and the OBR targets now supply the shape from published data instead. The lesson is symmetric and general: synthetic bunching is not evidence of behaviour, its absence is not evidence of no behaviour, and any sub-band feature of calibrated data should be traced to its targets before it is read as economics. The behavioural evidence on UK VAT bunching remains Liu et al.’s administrative estimate.

Related literature

The paper draws on four strands of work. The first is the methodology of bunching and notches (Chetty et al., 2011; Saez, 2010; Kleven and Waseem, 2013; Kleven, 2016; Best et al., 2015), which provides the analytic notion of a dominated region; I use the dominated region to characterise a distortion that leaves no single point of excess mass to exploit.

The second is the empirical literature on VAT registration thresholds (Liu et al., 2021, 2024; Onji, 2009; Harju et al., 2019; Asatryan and Peichl, 2017; Nandi and Warwick, 2020; Bashir et al., 2024). Liu et al. (2021) is closest: it estimates bunching at the UK notch on administrative data over 2004–2014 and documents voluntary registration; I take its estimates as the established behavioural facts and cost shape and rate reforms statically on a common basis, which that literature does not do. The international evidence (Harju et al., 2019; Onji, 2009; Asatryan and Peichl, 2017; Nandi and Warwick, 2020) notes that bunching at a VAT threshold can reflect compliance costs, firm-splitting, or reporting rather than real turnover responses.

The third strand questions whether bunching identifies a structural elasticity at all: Blomquist et al. (2021) show that the elasticity is not non-parametrically identified from bunching at a kink or a notch without strong functional-form assumptions, and Bertanha et al. (2023) develop the notch case. This literature motivates my refusal to read an elasticity off the synthetic data: even setting aside the placebo evidence that any synthetic step is mechanical, the mapping from excess mass to an elasticity requires assumptions I am unwilling to impose. The fourth strand is the structural literature on size-based thresholds (Garicano et al., 2016; Gourio and Roys, 2014), which identifies policy-invariant primitives at the French 50-employee threshold and computes welfare from them; the present paper makes no such structural claim.

This paper makes three contributions. First, it provides an open, reproducible firm-

level microsimulation for costing UK VAT threshold reforms, where the existing firm-level evidence rests on restricted administrative data; the released pipeline regenerates every number in the paper from public inputs. Second, it characterises the notch’s distortion through its exact dominated region and costs two schedule reforms—a graduated taper and a reduced-rate band—on the same static basis as level reforms. Third, it demonstrates, by construction and by correction, that band-calibrated synthetic data carry no sub-band behavioural information beyond their targets: the paper documents how its own earlier build manufactured a plausible-looking density step, how the correction removed it, how calibrating to the OBR’s published fine bands restores the administrative shape as an explicit target, and how the same band-edge mechanism manufactures a spurious signal where no fine targets exist. Identifying the turnover elasticity from administrative micro-data, a welfare accounting, revenue-neutral designs, and an explicit model of voluntary registration are left for future work.

The remainder of the paper proceeds as follows. Section 2 describes the institutional background of the UK VAT threshold. Section 3 sets out the synthetic firm microdata and its calibration to official aggregates. Section 4 sets out the static costing of threshold changes and the schedule reforms. Section 5 shows that the synthetic data contain no usable bunching signal and validates the estimator’s power on injected signals. Section 6 develops the notch and its exact dominated region, and shows how each reform changes it. Section 7 prices the reforms behaviourally, conditional on an assumed elasticity, and derives the level-rise invariance. Section 8 concludes.

2 Institutional background

Registration and the notch. The United Kingdom levies a value-added tax at a standard rate of 20% on most goods and services. A business must register once its taxable turnover over any rolling twelve-month period exceeds the registration threshold, after which it charges VAT on its taxable sales and reclaims VAT paid on inputs. Because registration makes VAT due on the firm’s *entire* taxable turnover rather than only the amount above the cutoff, the threshold is a *notch*—a discrete jump in liability—and not a kink. The cost of crossing is largest for consumer-facing firms with limited pass-through and small reclaimable input shares, for whom the incentive is to restrain turnover and locate just below the threshold (Liu et al., 2021; Kleven and Waseem, 2013).

Value added, input reclaim, and voluntary registration. Real VAT is a tax on *value added*, not on turnover: a registered firm charges output VAT on its sales but reclaims input VAT on its purchases. The model reflects this: each synthetic firm’s net liability is the standard rate applied to its value added (Section 3), and under the value-added formulation—in which unregistered firms bear no unreclaimed input VAT—the notch’s dominated region is invariant to the firm’s input share (Section 6). Two features of the real system remain outside the model’s scope. A large share of below-threshold firms register *voluntarily*—roughly 43% of eligible below-threshold firms (Liu et al., 2021), driven by input-VAT reclaim and business-to-business sales—and for them the

threshold carries no bunching incentive; voluntary registration enters the model only as a registration-rate parameter and a retention sensitivity, not as a choice. And net-repayment traders (zero-rated exporters and similar, HMRC’s negative-liability column) are not represented, because the public aggregates do not resolve rated structure by turnover band.

Threshold history. The registration threshold was held at £85,000 from 1 April 2017 to 31 March 2024—a seven-year nominal freeze, the longest in the history of the tax—before being raised to £90,000 on 1 April 2024, with the deregistration threshold rising from £83,000 to £88,000 (HM Revenue and Customs, 2024; Seely, 2024). Because nominal turnover grows with inflation and real activity, freezing the threshold lowers it in real terms and draws an increasing number of firms towards the registration margin; this fiscal-drag mechanism was a central motivation for the April 2024 increase. My data year is 2023–24, so the prevailing statutory threshold throughout my analysis is £85,000. Figure 2 shows the pattern in the turnover distribution: the number of firms rises towards £84,000 and falls sharply above £85,000, and the projected 2025–26 distribution is more sharply bunched than the 2019–20 outturn as fiscal drag draws more firms towards the frozen threshold.

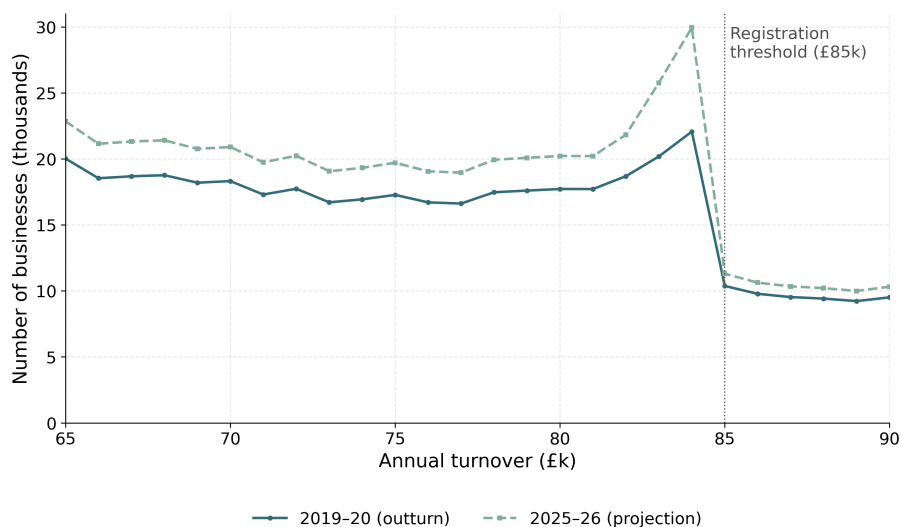


Figure 2: Bunching below the £85,000 VAT registration threshold: number of businesses by £1,000 turnover band. Source: OBR, *Economic and Fiscal Outlook March 2023*, Chart C (Open Government Licence).

The reform debate. The UK threshold is the highest in the OECD alongside Switzerland (HM Revenue and Customs, 2024), and there is debate over whether to raise it—to reduce the compliance burden on small firms—or lower it, to broaden the base and reduce the growth distortion at the margin. Three families of reform feature, and I evaluate all three: a change in the threshold’s *level* (the £85,000 to £90,000 increase of April 2024 and further moves around it); a change in its *shape* (a *graduated taper* that

phases the effective rate in over a band, replacing the notch with a continuous profile); and a change in its *rate* (a *reduced rate* for small firms in a band above the threshold, with the standard 20% rate retained above). The taper and reduced-rate designs follow proposals in the UK policy debate: [Office of Tax Simplification \(2017\)](#) canvassed a financial incentive or reduced effective rate to smooth the cliff-edge at the threshold, and HM Treasury’s call for evidence on the registration threshold examined smoothing mechanisms in the same family ([HM Treasury, 2018](#)). The European Union ran a closely related “graduated relief” scheme for small enterprises and abolished it from January 2025 on complexity grounds, in favour of a simple exemption with a transition tolerance ([Council of the European Union, 2020](#))—relevant precedent for the administrability of taper-style designs. A banded reduced rate introduces a smaller *secondary* notch at the band top, which the continuous taper avoids. I cost all three by the same machinery in Section 4.

3 Data

Studying firm responses at the threshold needs firm-level turnover near the cutoff, which the official statistics publish only in coarse turnover bands. I therefore build a synthetic firm population calibrated to those aggregates: it reproduces the official totals while resolving the individual firm that the authorities do not release, opening the analysis to anyone. The same calibrate-to-published-aggregates approach underlies open household-level tax-benefit microsimulation ([PolicyEngine, 2025](#)); here I apply it at the firm level. The generator takes the registration threshold as a parameter and is documented in Appendix A.1; I build one population for 2023–24 (threshold £85,000), the paper’s primary data year, and one for 2024–25 (threshold £90,000), used for the forward sweep of Section 4.

The archived CSV inputs in this repository remain the reproduction source for the paper’s reported numbers; a pinned migration snapshot also represents the same numeric target surface through PolicyEngine Ledger and the experimental Populace firm generator, and Appendix A.1 reports the parity check between that shared source-of-truth path and the paper inputs.

Construction. Index firms by i . For each sector s and ONS turnover band $b = [\underline{b}, \bar{b}]$, the *UK Business: Activity, Size and Location* table gives a business count $N_{s,b}$, and I draw $N_{s,b}$ firms with within-band turnover

$$y_i = \underline{b} + (\bar{b} - \underline{b}) u_i + \varepsilon_i, \quad u_i \sim \mathcal{U}(0, 1), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_b^2),$$

truncated below at zero, where ε_i smooths the within-band distribution. The ONS register covers businesses registered for VAT and/or PAYE—about 2.7 million—so the synthetic population represents that registered-business frame, not the full business population of roughly 5.6 million in the DBT business population estimates; the difference is concentrated below the threshold and matters for threshold *cuts*, which I flag where

relevant. Each firm is given input expenditure $x_i = \rho_i y_i$, with the input–output ratio ρ_i drawn from a rescaled Beta distribution with sector-specific shifts, clamped to $\rho_i \in [0.1, 0.95]$ (mean about 0.6, so value added averages about 40% of turnover, in line with the UK non-financial business economy). Its net VAT liability is the standard rate applied to its value added,

$$v_i = \tau (y_i - x_i), \quad \tau = 0.20,$$

so every firm’s net rate $v_i/y_i = \tau(1 - \rho_i)$ lies in $[1\%, 18\%]$, strictly positive and below the statutory cap of 20%. The model therefore abstracts from net-repayment traders (firms with zero-rated outputs whose input reclaim exceeds output VAT—HMRC’s negative-liability column); representing them requires rated-structure detail (zero, reduced, and exempt output shares) that the public aggregates do not resolve by band. Employment is assigned from the ONS employment-band shares.¹

The base population reproduces the ONS structure but not the HMRC VAT-registered totals, so I re-weight it. Each firm receives a strictly positive weight $w_i = e^{\theta_i}$, and for a vector of official targets $T = (T_k)_{k=1}^K$ the weighted synthetic counterpart of target k is

$$\hat{T}_k(\theta) = \sum_i a_{ki} w_i,$$

where a_{ki} is firm i ’s contribution to target k : an indicator of band or sector membership for a count target, and the firm’s net liability v_i for a VAT-liability target. I choose θ to minimise a symmetric relative-error loss with per-target importance weights λ_k ,

$$L(\theta) = \frac{1}{K} \sum_k \lambda_k \min\{(\hat{T}_k/T_k - 1)^2, (T_k/\hat{T}_k - 1)^2\} + \frac{\alpha}{N} \sum_i |\theta_i|,$$

the final term a mean absolute penalty on the log-weights. The targets are the HMRC VAT-registered counts by turnover band and by sector, the HMRC net VAT liability by turnover band, the ONS employment-band totals, and—for the 2023–24 vintage—the OBR’s published near-threshold profile: the March 2023 *Economic and Fiscal Outlook* Chart C data give HMRC counts of businesses by £1,000 turnover band over £65,000–£90,000 ([Office for Budget Responsibility, 2023](#)), which I interpolate to the 2023–24 data year between the 2019–20 outturn and the OBR’s 2025–26 frozen-threshold projection (the chart’s final row, the £90,000 bin, lies beyond the fine window and is not used). All twenty-five bins enter as *shape* targets, applied side-consistently: each side of the threshold takes the OBR profile’s within-side shape, normalised over its own window (£65,000–£85,000 below; £85,000–£90,000 above) and scaled to the synthetic population’s own mass there. The chart counts HMRC traders—a different unit and, below

¹An earlier draft of this generator set $v_i = y_i - x_i$ with $\rho_i \in [0.6, 1.5]$ —omitting the factor τ , so per-firm liabilities were about five times too large and a large minority of firms carried negative value added. The weight optimiser compensated on the band totals, leaving aggregate calibration scores similar but distorting the near-threshold density; Section 5 documents the consequence and the correction.

the threshold, a substantially narrower universe than the ONS business frame—so its levels are not imported on either side. An earlier build that mixed conventions (direct counts above the threshold, frame-scaled shape below) placed more mass just above the threshold than just below it, inverting the administrative profile’s cross-threshold ordering; under side-consistent scaling the cross-threshold step comes from the frame’s own band structure and the OBR data supply only the within-side geometry—the rise into the threshold and the decline beyond it. These fine targets pin the near-threshold density to the OBR profile’s within-side geometry at frame-implied levels—the region that coarse bands alone leave to the weight optimiser (Section 5); the resulting cross-threshold step (−25%) is about half the OBR chart’s own (−60%), because the chart’s below-threshold universe is narrower than the frame. No fine bands are published for the £90,000 threshold era, so the 2024–25 vintage keeps coarse bands only. Turnover-band and near-threshold targets carry the largest λ_k , followed by VAT liability by band. Two liability dimensions are reported as informational diagnostics but excluded from the optimizer. VAT liability by sector is excluded because the generator draws a common input-share distribution with sector shifts rather than calibrating the input/output tax structure by sector. The below-threshold liability band (£1–threshold, £1.46bn in 2023–24) is excluded because that total is remitted by voluntary registrants—input-reclaim traders averaging about £2,150 of net remittance—whom the model’s standard-rate-on-value-added liability does not represent (Section 2); imposing it on the whole below-threshold population would force the optimiser to suppress weights in the £50,000–£85,000 region, distorting the near-threshold shape the fine targets exist to pin. The problem is solved by gradient descent, and the symmetric loss keeps targets of very different magnitudes on a common scale.

A firm is registered with certainty when its data-year turnover exceeds the threshold T^* . This is an approximation to the statutory test, which is a rolling 12-month test on *taxable* turnover with a roughly two-month registration lag, a forward-looking 30-day test, and an exception for temporary breaches (Section 2); the model also registers firms whose turnover is largely exempt, whom the law would not require to register. For a single-year cross-section calibrated to registered-population totals these timing and composition wedges are second order for the aggregates, but they are worth naming: the HMRC band tables classify traders by total declared outputs (VAT-return Box 6), and the ONS turnover measure is likewise total turnover, so the running variable throughout is total rather than taxable turnover. Below the threshold a firm registers voluntarily with probability $r = M/n_<$, where M is the HMRC count of registered firms in the £1–threshold band and $n_<$ is the unweighted number of synthetic below-threshold firm rows. This sets the expected row-level voluntary-registration rate; I do not treat below-threshold voluntary registration as a separate weighted calibration target. The resulting dataset comprises approximately 2.94 million firm records per vintage, weighted to represent roughly 2.5 million registered businesses. Checked against the official sources along every calibrated dimension, the vintages attain overall calibration accuracies of 89.2% (2023–24) and 92.8% (2024–25), with dimension scores ranging from 77.9% (2023–24 employment bands) to 94.2%; Appendix A.1 reports the full ta-

ble. The headline is the mean of five calibrated-dimension scores, each clipped at zero as $\max\{0, 1 - |\hat{T} - T|/|T|\}$; it excludes the VAT-liability-by-sector diagnostic described above.

The registration threshold in the data. My primary data year is 2023–24, when the statutory threshold was £85,000. The HMRC band targets step down across the threshold—678,350 firms in the £1–£85,000 band against 305,320 in £85,000–£150,000, or roughly 8,000 against 4,700 firms per £1,000 of band width—and the calibrated population necessarily inherits that band-average step. The *local* shape within a few thousand pounds of the threshold is pinned by the OBR near-threshold targets described above: the calibrated density rises into the threshold (about 13,800 firms per £1,000 in the £84,000–£85,000 bin), steps down across it (about 10,300 just above), and declines through the band—the administratively observed bunching geometry ([Office for Budget Responsibility, 2023](#); [Liu et al., 2021](#)), present in the file because it is targeted, not because any synthetic firm chooses it. In an earlier build without these targets the local shape was an unpinned optimiser equilibrium (and, under the pre-correction liability scale, a spurious bunching look-alike); the central data caveat of the paper stands in all three configurations: no sub-band feature of band-calibrated synthetic data carries behavioural information beyond what its targets put there (Section 5 quantifies this with placebo and recovery tests). Counterfactual policies—including the £85,000 to £90,000 reform and the forward threshold sweep of Section 4—are evaluated on these firms by re-applying the registration rule to the microdata for the static costings, holding the empirical turnover distribution fixed. A behavioural layer that re-prices firms under counterfactual schedules, conditional on an assumed turnover elasticity, is developed in Section 7.

4 Static costing of threshold reforms

A static costing measures the first-order revenue effect of a reform by re-applying the tax rule to a fixed firm population, with no behavioural response. Let firm i have turnover y_i , calibrated net VAT liability $v_i = \tau(y_i - x_i)$ (Section 3), and weight w_i . A firm remits VAT only once registered, which the model equates with turnover above the threshold T , so total VAT revenue at threshold T is the weighted sum of liabilities over registered firms,

$$R(T) = \sum_i w_i v_i \mathbf{1}\{y_i \geq T\}.$$

Two scope conventions follow from this formula and are worth stating. First, below-threshold *voluntary* registrants’ remittances (a £1.46bn HMRC target in 2023–24) are outside $R(T)$, so a threshold cut credits the full liability of newly covered firms even though roughly 43% of below-threshold firms would already be registered voluntarily ([Liu et al., 2021](#)). Second, and symmetrically, the headline treats every firm released by a threshold rise as deregistering; the anchor comparison below reports the voluntary-retention sensitivity. Because turnover is held fixed, moving the threshold from T_0 to T

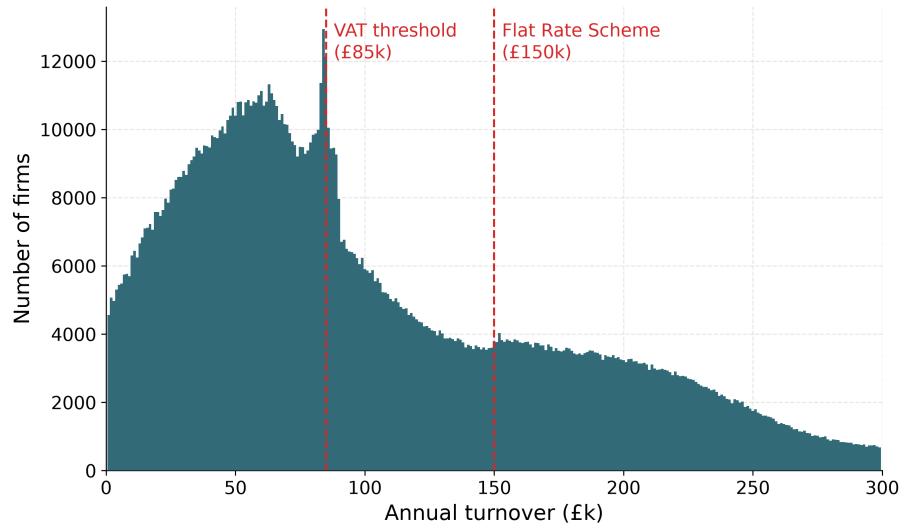


Figure 3: Synthetic firm turnover distribution (weighted firm counts, £1,000 bins), 2023–24. The near-threshold profile—rising into the £85,000 threshold and stepping down across it—is calibrated to the OBR’s published £1,000-band counts and is therefore target-inherited, not behavioural (Section 5). Levels reflect the ONS VAT/PAYE-registered frame: unregistered businesses—most of the below-threshold mass, and most of the cross-threshold cliff, in all-business administrative charts—are outside it, so level comparisons to such charts are not like-for-like. The steps at £90,000 (about 9,100 firms per £1,000 in the last OBR-pinned bin against about 6,800 in the first unpinned bin) and £150,000 are calibration-target boundaries—the top of the OBR fine-band window and an HMRC band edge respectively—not features of the underlying firm distribution. The £90,000 boundary lies inside the bunching estimator’s fit range; Appendix A.2 measures its effect on the headline statistics.

changes revenue only through the firms whose registration status flips:

$$\Delta R = R(T) - R(T_0) = -\text{sign}(T - T_0) \sum_i w_i v_i \mathbf{1}\{T_{\min} \leq y_i < T_{\max}\},$$

where $T_{\min} = \min(T, T_0)$ and $T_{\max} = \max(T, T_0)$: lowering the threshold draws firms into the net and raises revenue, while raising it releases firms and loses revenue. The estimate is the mechanical reclassification effect and excludes the behavioural response of Section 7. I age the microdata to a given fiscal year by a cumulative nominal-growth factor; band membership is evaluated on data-year turnover against each year’s nominal thresholds, and only liabilities are aged—an approximation stated because a fully aged band would shift membership toward the pinned near-threshold density. At the £90,000 threshold this puts total VAT revenue at £200.9bn in 2025–26 and £206.1bn in 2026–27.²

Anchor comparison against HMRC’s published costing. I compare the static pipeline against HMRC’s costing of the April 2024 increase from £85,000 to £90,000, announced and costed at Spring Budget 2024 ([HM Treasury, 2024](#)). The comparison requires the costing baseline convention: the official pre-measure baseline is not a frozen £85,000 threshold but a counterfactual path in which the freeze legislated to March 2026 ends and the threshold resumes uprating (85, 85, 87, 89, 92 thousand pounds over 2024–25 to 2028–29 in the model’s implementation), so the measure’s impact shrinks over the forecast and changes sign in 2028–29, when the counterfactual threshold overtakes £90,000. The official profile is –150, –185, –125, –50, +65 £m. My mechanical costing depends on the deregistration convention. If every released firm deregisters, the model gives –317, –323, –195, –66, +103: it reproduces the sign, the narrowing profile, and the 2028–29 turning point, but overshoots the official magnitudes throughout (–317 against –150 in 2024–25; –323 against –185 in 2025–26). If instead the [Liu et al. \(2021\)](#) voluntary-registration share (43%) of released-firm liability is retained, the model gives –180, –184, –111, –38, +59: above the official number in the first year, close to it in 2025–26 (–184.1 against –185), and modestly below it thereafter (Figure 4). That is the honest reading of the anchor: the official costing embeds a phased registration response (HMRC expected roughly 28,000 fewer registrants in the first year, against about 48,000 firms in the released band here, now pinned by the OBR fine-band targets), so the full-deregistration convention overshoots throughout, and the retention convention crosses

²These aggregates are weighted sums of the synthetic registered population’s net VAT liabilities aged forward—a theoretical-liability (VTTL-style) base, not the published cash-receipts line (around £169bn in 2023–24; OBR forecast near £180bn for 2025–26), which is reduced by refunds, the VAT gap, and collection timing. The unaged bases also sit about 5.8% above the HMRC liability-by-band target sums they are calibrated toward (£184.8bn against £174.6bn above-threshold in 2023–24; £190.7bn against £180.4bn in 2024–25, and by 9.8% in the £85,000–£150,000 band specifically): the multi-target optimiser trades the liability totals off against the count targets, and the liability-by-band calibration scores (91.3% and 92.0%) bound the shortfall. Band-to-band *differences*, which drive every costing in the paper, are less exposed than the level. The growth factors are cumulative from 2023–24 and are applied to both vintages; to the extent the 2024–25 vintage already embodies 2024–25 nominal levels, the aged 2024–25 aggregates overstate nominal growth by up to one year’s factor ($\approx 3\%$), within the calibration error just described.

the official path—above it in the first year, below it from 2026–27. Because my microdata are calibrated to HMRC VAT aggregates, even a close match would confirm internal consistency rather than out-of-sample accuracy. The sub-£1m agreement in 2025–26 is, in particular, finer than the model’s own liability level supports: the £85,000–£150,000 band’s calibrated liability sits 9.8% above its HMRC target total (footnote above), and rescaling the released band to target puts the retention figure near –£168m—the year-2 agreement is a coincidence within that error band, not model precision. The retention share is itself a convention: LLAT’s roughly 43% is a *firm* share among all below-threshold eligibles, applied here as a liability share among just-released firms.

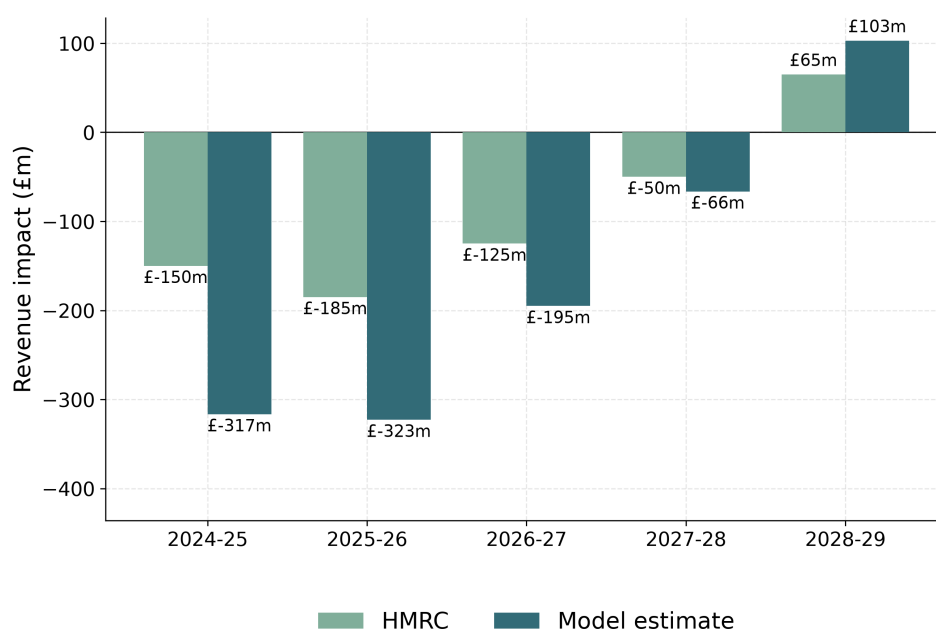


Figure 4: Anchor reform: revenue impact of raising the VAT registration threshold from £85,000 to £90,000 (the April 2024 change), my estimate versus HMRC’s published costing, 2024–25 to 2028–29. Both series narrow over the forecast and turn positive by 2028–29 because the costing baseline is a counterfactual threshold path that resumes uprating after March 2026 and overtakes £90,000. The mechanical full-deregistration estimate overshoots the official costing while its phased registration response builds; the voluntary-retention convention runs close to the official figure in 2025–26 and modestly below it thereafter; agreement finer than the calibrated liability level supports is coincidental (see text). Because the microdata are calibrated to HMRC and OBR aggregates, this is a consistency check, not an out-of-sample validation.

A menu of threshold locations. I next cost a menu of further threshold locations, taking the post-reform £90,000 threshold as the baseline. This sweep runs on the 2024–25 vintage population (Section 3), whose liability-by-band calibration is the marginally stronger of the two (92.0% against 91.3%; both vintages score about 90% overall). Each

row of Table 1 costs a move of the threshold to a new location for 2025–26, holding the turnover distribution fixed, so the revenue change reflects only the reclassification of firms into or out of the VAT net. Because the optimiser determines the within-band shape near the threshold (Section 3), I evaluate the reclassified band from a smooth extrapolation of the above-threshold profile rather than the raw local mass. Lowering the threshold draws firms in and raises revenue; raising it loses revenue; the threshold-cut rows inherit the registered-business frame, so they omit currently unregistered businesses that a lower threshold would newly cover. The schedule is approximately linear with a mild concavity: each £5,000 step changes revenue by roughly £230m–£265m, slightly less per step at higher thresholds because the number of firms reclassified per step falls (from about 38,000 to about 26,000) faster than per-firm liability rises. A £30,000 rise to £120,000 costs about £1.4bn, 0.7% of the £200.9bn base, because the threshold governs many small firms whose individual liabilities are small. Figure 5 shows the sweep graphically. These are *static* results: they bound the first-order fiscal effect of a threshold change but abstract from the behavioural response of Section 7.

Table 1: Static revenue and VAT-paying-firm effects of moving the VAT registration threshold, 2025–26, relative to the post-reform £90,000 baseline (total VAT revenue £200.9bn; 2024–25 vintage population). Figures are mechanical, holding firm turnover fixed (no behavioural response). This sweep is distinct from the anchor reform of Figure 4. The machine-readable table is `results/static_sweep.txt` in the replication repository.

Threshold	Revenue change (£m)	Change (%)	VAT-paying firms (000s)
£70,000	+1,042.8	+0.52	+145.7
£75,000	+776.2	+0.39	+107.2
£80,000	+513.6	+0.26	+70.1
£85,000	+254.8	+0.13	+34.4
£90,000 (baseline)	0	0	0
£95,000	−250.9	−0.12	−33.0
£100,000	−498.0	−0.25	−64.6
£105,000	−741.1	−0.37	−94.8
£110,000	−980.4	−0.49	−123.7
£115,000	−1,215.7	−0.61	−151.2
£120,000	−1,447.1	−0.72	−177.3

4.1 Costing schedule reforms

The threshold sweep above moves the *location* of a single cutoff. The same static principle—re-applying a tax rule to the fixed turnover distribution, with no behavioural response—also costs reforms that change the *shape* of the schedule rather than merely its position.

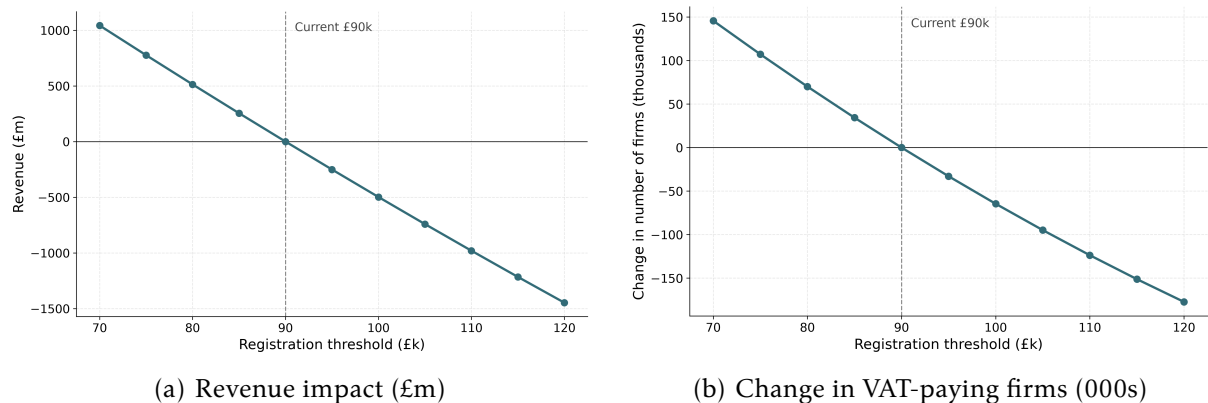


Figure 5: Static effect of moving the VAT registration threshold, 2025–26, relative to the £90,000 baseline (dashed line). Panel (a): revenue impact—lowering the threshold raises revenue, raising it loses revenue, approximately linearly over the £70,000–£120,000 range. Panel (b): the change in the number of VAT-paying firms, tracking the local density of firms on each side of the threshold.

I consider four reforms and report their first-order revenue effects against a common baseline: the £85,000-notch / £184.8bn 2023–24 base. (This 2023–24 data-year base differs from the £200.9bn figure in the sweep above, which runs on the 2024–25 vintage aged to 2025–26 at the post-reform £90,000 threshold; the schedule reforms are costed in the data year to match the bunching and dominated-region analysis. Both bases are theoretical-liability (VTTL-style) aggregates in the sense of the footnote above, not cash-receipts lines.) For each reform I recompute the weighted sum of liabilities $R = \sum_i w_i v_i$ under the new schedule, holding each firm’s turnover fixed, and difference it against the baseline.

Threshold relocation. Raising the threshold from £85,000 to £100,000 releases the firms with turnover in [£85,000, £100,000)—about 111,900 firms—from the VAT net, for a static revenue change of –£753m on the common base. Because the released band lies *above* the data-year threshold, its liabilities are directly calibrated (and its lower reach pinned by the OBR fine-band targets), so the band is summed directly; the sweep’s smooth-counterfactual method, which irons out the near-threshold concentration by construction, gives –£697m and 100,700 firms, 7% and 10% below the direct sum—a gap that now measures the real, target-pinned concentration of firms just above the threshold rather than an optimiser artifact. On this common base the level move carries the largest static cost of the menu and leaves the dominated region in place, relocating it to the new cutoff (Section 6); the taper removes the dominated region at a static cost of –£520m. Which matters more depends on what the reader wants from the reform—the dominated-region metric indexes misallocation only, and Section 8 discusses what it omits.

Graduated taper. A graduated taper replaces the discrete notch with an effective rate that phases linearly from 0% at £85,000 to the full 20% at £105,000. Firms in the band remit a tapered rather than a full rate on their (fixed) turnover; mechanically this reduces revenue by –£520m, with about 140,800 firms in the £85,000–£105,000 band facing a lower effective rate.

Banded reduced rate. A banded reduced rate charges firms in the £85,000–£105,000 band a reduced rate $\tau_{\text{low}} \in \{10\%, 15\%\}$ in place of the standard 20%, with 20% retained above £105,000. Statically this reduces revenue by –£484m at $\tau_{\text{low}} = 10\%$ and by –£242m at $\tau_{\text{low}} = 15\%$, again with about 140,800 firms in the band facing the lower rate.

Table 2: Static, behaviour-free revenue effect of the four reforms relative to the common £85,000-notch / £184.8bn 2023–24 baseline. Each reform is costed mechanically by re-applying its schedule to the fixed turnover distribution, with no behavioural response.

Reform	Lever	Static revenue (£m)
Raise threshold to £100,000	Location	–753
Graduated taper [£85k, £105k]	Shape (phase-in)	–520
Reduced rate 10% [£85k, £105k]	Rate (step)	–484
Reduced rate 15% [£85k, £105k]	Rate (step)	–242

These static figures bound the first-order fiscal effect of each reform but abstract from the behavioural response of Section 7. One feature of the shape reforms is not visible in a static costing: unlike a threshold move, which transports the notch to a new cutoff, the taper and the reduced rate *change the size of the notch itself*—the discrete drop in net revenue at the threshold and the empty band of turnover it induces. Quantifying that change requires the structural object I develop in Section 6, where I derive the notch’s dominated region and show analytically how each of these reforms alters it.

5 Bunching and the limits of band-calibrated synthetic data

This section applies the polynomial counterfactual-density estimator of Chetty et al. (2011) and Kleven and Waseem (2013)—as applied to UK VAT data by Liu et al. (2021)—to the synthetic firm population. Its finding is about the data, not about firm behaviour: every sub-band feature of a band-calibrated synthetic population is exactly what its targets put there, and this section demonstrates the point in three directions. The 2023–24 population’s near-threshold density is calibrated to the OBR’s published £1,000-band profile (Section 3), and the estimator reads that profile back as excess mass—target inheritance, not behaviour. A placebo that regenerates the population *without* those targets returns zero. And the 2024–25 vintage, for which no fine bands are published, pro-

duces a very large *spurious* excess-mass signal sitting exactly on an HMRC calibration-band edge. Together with a power (recovery) test, these results establish the section’s transparency claim: band-calibrated synthetic data support no bunching inference in either direction, and any sub-band feature of such data should be traced to its calibration targets before it is read as economics. The behavioural evidence on UK VAT bunching remains the administrative estimate of [Liu et al. \(2021\)](#).

The section also documents a correction. An earlier draft of this paper reported an excess mass of 8,712 firms at £85,000 that appeared to “reproduce” the density step documented in administrative data. That excess mass was an artifact of a liability mis-scaling in the generator (net liability set to value added itself rather than the standard rate applied to value added, inflating per-firm liabilities roughly fivefold; Section 3, footnote), which distorted how the weight optimiser distributed mass near the threshold. Correcting the scaling removed the step entirely; the near-threshold profile is now supplied by the OBR targets instead, as published data rather than an optimiser side-effect.³ A plausible-looking, literature-adjacent bunching statistic was manufactured by an accounting bug in the data layer and survived until the liability accounting was audited. I report this openly because it is the sharpest available demonstration of the section’s thesis.

5.1 Methodology

Following [Liu et al. \(2021\)](#), I group firms into fine turnover bins of £1,000 and estimate a counterfactual density—the distribution that would obtain absent the registration notch—by polynomial regression on bin counts, excluding a manipulation window around the threshold y^* :

$$c_j = \sum_{l=0}^q \beta_l (y_j)^l + \sum_{i \in [y_-^*, y_+^*]} \gamma_i 1\{j = i\} + \varepsilon_j,$$

where c_j is the number of firms in bin j , y_j is the distance of bin j from the threshold, q is the polynomial order, and $[y_-^*, y_+^*]$ is the excluded window. The fitted polynomial is then projected into the excluded region to recover the counterfactual density $g(y)$, with a mass-conservation restriction that locates the upper edge of the manipulation window endogenously (excess mass below the threshold must equal missing mass above it; Appendix A.2). Excess bunching is the integrated difference between the observed

³The defect and its diagnosis are recorded in the replication repository (issue #15); the near-threshold targets are issue #23. A second correction followed: the below-threshold HMRC liability total, which belongs to voluntary registrants the liability model does not represent, was initially calibrated against the whole below-threshold population, draining weights around the fine-target window; on an intermediate build it also manufactured a spectacular ($E \approx 196,000$), specification-robust artifact at the £90,000 band edge. It is now an informational diagnostic (Section 3). Aggregate calibration scores barely moved under any of these changes—the optimiser hit the band totals every time—which is itself part of the lesson: aggregate fit does not discipline sub-band shape.

and counterfactual densities just below the threshold,

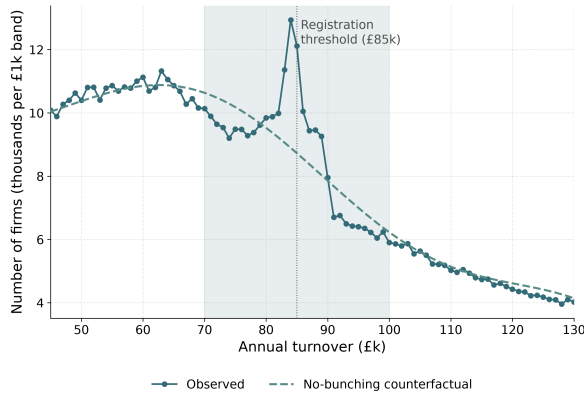
$$B(y^*) = \int_{y^* - \Delta y^*}^{y^*} g(y) dy \approx g(y^*) \Delta y^*,$$

the approximation holding when g is locally smooth. The bunching ratio $b(y^*) = B(y^*)/g(y^*) \approx \Delta y^*/y^*$ is the width of the bunching region expressed as a fraction of the threshold—a normalised *ratio*, not an elasticity. The estimator delivers the raw geometry of any density step at the threshold; whether that step reflects firm behaviour is a separate question. The estimator reports geometry only (b , excess mass E , and the mass-conservation objects): no elasticity is computed from it, both because the identification critique of bunching designs applies (Blomquist et al., 2021; Bertanha et al., 2023) and because on these data there is no behavioural mass to convert.

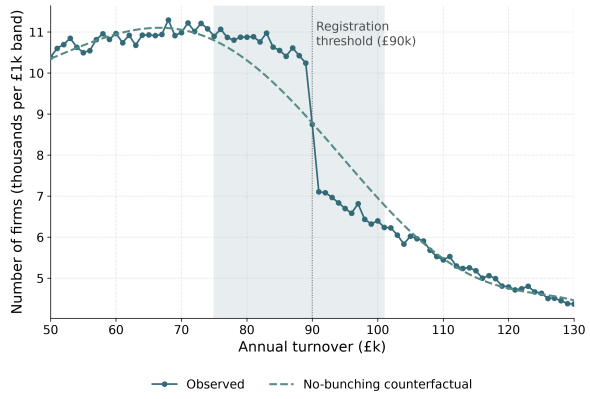
5.2 Results: a target-inherited signal at £85,000, a spurious one at £90,000

Applied to the 2023–24 population at the £85,000 threshold, the estimator returns an excess mass of $E = 7,933$ weighted firms: it reads back the bunching profile that the OBR £1,000-band targets place in the file (Section 3). This is target inheritance, not behaviour—the generator contains no firm location-choice—and the placebo below proves it by deletion. On the Liu et al. (2021) normalisation the target-inherited excess is $b_{\text{LLAT}} = 1.91$, about 1.4 times their administrative 1.361; the two are not comparable estimates of the same object—LLAT’s sample ends in 2014–15, the profile here is interpolated OBR data whose cross-threshold level is frame-implied (Section 3), and the readback varies severalfold across estimator specifications (Appendix A.2)—so the distance from 1.361 carries no information. Mass conservation fails to bind at £85,000: the marginal-buncher walk is censored at the search bound ($y_R = £100\text{k}$) with missing mass $\Delta_R = 5,761 < E$, all of it accruing beyond the £90,000 fine-target boundary—a calibration-band edge inside the estimator’s own window, not administrative-style missing mass (Appendix A.2).

Applied to the 2024–25 population at the £90,000 threshold—which has *no* fine targets—the same estimator returns $E = 11,072$ and $b_{\text{LLAT}} = 1.23$: within ten per cent of Liu et al. (2021)’s administrative 1.361. Nothing behavioural produced it. It is the coarse £90,000 band edge, where the £1–£90,000 and £90,000–£150,000 targets meet, read through the estimator; its magnitude swings four-fold across estimator specifications (E between roughly 8,600 and 36,700 over the degree–window grid, Appendix A.2) and the manipulation-window edge sits at the search bound. The pair of results is the section’s central exhibit (Figure 6): where fine targets exist, the estimator recovers exactly the shape they encode; where they do not, band geometry alone can manufacture a bunching statistic in the publishable range of the literature—a researcher looking to “validate” synthetic data against Liu et al. (2021) would have found agreement here, and it would have meant nothing.



(a) 2023–24 population, £85,000: target-inherited profile



(b) 2024–25 population, £90,000: band-edge artifact

Figure 6: Observed synthetic firm density and polynomial no-bunching counterfactual at each vintage’s registration threshold. Panel (a): the 2023–24 population reproduces the administratively observed bunching profile because it is calibrated to the OBR £1,000-band targets; the estimator reads the target-inherited excess back as $E = 7,933$ ($b_{\text{LLAT}} = 1.91$). Panel (b): the 2024–25 population has no fine targets, and its coarse £90,000 band edge alone yields $E = 11,072$ with $b_{\text{LLAT}} = 1.23$ —within 10% of the administrative estimate, coincidentally and unstably. Neither is behavioural: the generator has no firm location-choice, so all sub-band structure is inherited from the targets and the weight optimiser.

5.3 Placebo test

The placebo diagnoses where any threshold-adjacent structure comes from by manipulating the calibration targets directly. HMRC publishes turnover counts only in coarse bands across the threshold: in 2023–24, 678,350 firms at or below £85,000 (about 7,981 per £1,000) against 305,320 between £85,000 and £150,000 (about 4,697 per £1,000), so a band-average step is present in the target before any firm is generated. In Placebo A I reweight the synthetic firms so the per-£1,000 density follows a single smooth trend across the threshold, holding total mass fixed. In Placebo B I regenerate the entire population and re-calibrate from scratch with *no* fine structure in the targets at all: the OBR near-threshold bins are removed and the two HMRC band counts straddling the threshold are replaced by a split implied by one common density. The results track the targets exactly. The control—calibrated to the OBR profile—returns $E = 7,933$; Placebo A returns $E = 0$ ($b = -0.059$); Placebo B returns $E = 0$ ($b = -0.073$). Deleting the fine targets deletes the “bunching”; and where a coarse step is present in the targets—as at the 2024–25 vintage’s £90,000 band edge—the estimator reads it as (artifactual) mass. Inserting or removing structure in the targets inserts or removes “bunching” in the estimates.

5.4 Power (recovery) test

A null result is only informative if the estimator could have detected a signal. I therefore inject a Kleven–Waseem-consistent behavioural response of known magnitude into the step-free (Placebo A) population: firms relocate to just below the threshold from a donor region extending to a true marginal buncher at £112,000—beyond the estimator’s exclusion window, as with genuine bunching—with relocation probability declining linearly in distance from the threshold. Scored by the headline estimator with no reference to the known null, injected excess masses of 5,000 and 8,000 weighted firms are recovered at 49% and 62% respectively, while a small 2,000-firm injection sits near the detection floor (15%); the estimated marginal-buncher location is attenuated toward the threshold throughout. The attenuation has a measured mechanical source: part of the missing-mass region lies outside the exclusion window, where it pulls the fitted counterfactual down. The estimator therefore *under*-states a true response, more severely for small ones, and does not over-state one—so the target-inherited $E = 7,933$ at £85,000 is, if anything, an attenuated reading of the targeted profile.⁴

5.5 Comparison with the literature

The behavioural fact in the literature is administrative. [Liu et al. \(2021\)](#) (hereafter LLAT) document genuine bunching at the UK VAT registration notch in HMRC micro-data over 2004–05 to 2014–15, during which the threshold rose from £58,000 to £81,000, with

⁴An earlier version of this exercise relocated mass entirely within the exclusion window, which the counterfactual fit cannot see by construction; its reported “~90% recovery” measured deposit bookkeeping, not estimator power. The redesigned test is in `analysis/recovery_bunching.py`.

an excess-bunching ratio of $b = 1.361$ (SE 0.202) around the year-specific threshold and a manipulation window of $-\pounds 14,000$ to $+\pounds 24,000$. My data's $b_{\text{LLAT}} = 1.91$ is not a second estimate of their parameter: it is the OBR target profile read back through their normalisation; and the uncalibrated 2024–25 vintage's $b_{\text{LLAT}} = 1.23$, despite sitting within 10% of their estimate, is not one either—it is band geometry.⁵ What does carry across is institutional context: LLAT report that almost half (roughly 43%) of eligible below-threshold firms register voluntarily, which dampens observed bunching relative to a pure notch and which motivates the voluntary-retention sensitivity of Section 4. For wider order-of-magnitude context, [Office for Budget Responsibility \(2023\)](#) count firms holding turnover below the frozen threshold over a *wide* band (rising from 23,000 in 2017–18 to about 44,000 in 2025–26, with foregone turnover of $\pounds 110\text{m}$ rising to $\pounds 350\text{m}$); this is a different object from a near-threshold spike and is cited only as context.

5.6 Reduced-form evidence and the dominated region

These results sharpen, rather than weaken, the case for grounding the policy analysis in structure. The reduced-form density around the threshold cannot price reforms—it has no single point of excess mass to forward-map onto a change in the *shape* of the schedule—and on band-calibrated synthetic data it cannot be read behaviourally at all, in either direction. What the structure of Section 6 *does* deliver exactly, for every reform considered, is the notch's dominated region and how each schedule change alters it. It is precisely because the reduced-form bunching is non-identifying here that the exact dominated region carries the distortion characterisation. The behavioural layer that re-prices each firm's turnover under counterfactual schedules—conditional on an assumed turnover elasticity—is developed in Section 7.

6 The notch and its dominated region

The behavioural response to the registration notch is documented in administrative data by [Liu et al. \(2021\)](#); the synthetic data contain no bunching signal to exploit (Section 5). To characterise the notch precisely I therefore turn to its analytic structure, which yields an exact distortion measure that needs no behavioural calibration. I set out the firm-level structure following [Kleven and Waseem \(2013\)](#), [Saez \(2010\)](#), and [Kleven \(2016\)](#), whose bunching framework [Liu et al. \(2021\)](#) adapt to UK VAT. The UK registration threshold is a *notch*, not a kink: once turnover crosses the threshold the firm becomes liable for VAT on its *entire* base, not merely on the increment above it. The section's central deliverable is the notch's *dominated region*—an exact, elasticity-free property of the schedule that indexes the distortion the threshold imposes (Section 6.2)—and its analytic response to each schedule reform, complementing the static reform costs of

⁵The earlier draft's $b_{\text{LLAT}} = 0.94$, read against LLAT's 1.361, illustrated how a band-inherited artifact can land in the publishable range of the literature; that artifact's disappearance under the liability correction, and its replacement by an explicit published-data target, are reported above.

Section 4.1. Section 7 prices the intensive-margin response to each reform conditional on an assumed turnover elasticity; nothing in this section depends on that assumption.

Scope. The model prices the registration margin for firms that remit positive net VAT. Two populations are outside its scope and are treated as such throughout: below-threshold *voluntary* registrants (roughly 43% of eligible firms in Liu et al., 2021), who have chosen registration and so face no notch at the threshold, and net-repayment traders (zero-rated exporters and similar), whose remittance is negative for rated-structure reasons the generator does not represent (Section 3).

6.1 The firm's problem

Each firm i chooses turnover y_i to maximise after-tax profit. Let $c_i(y_i)$ denote the cost of producing turnover y_i , T^* the registration threshold, and τ the standard VAT rate. In the simplest statement VAT applies to the firm's entire turnover once it registers, so after-tax revenue is discontinuous and profit is

$$\pi_i(y_i) = \begin{cases} y_i - c_i(y_i), & y_i < T^* \quad (\text{unregistered}), \\ (1 - \tau) y_i - c_i(y_i), & y_i \geq T^* \quad (\text{registered}). \end{cases} \quad (1)$$

At T^* after-tax revenue falls discretely by τT^* : a firm that nudges its turnover from just below to just above the threshold loses τT^* of net revenue on turnover it was already earning. With the standard rate $\tau = 0.20$ and threshold $T^* = \text{£}85,000$ this discrete drop is $\tau T^* = \text{£}17,000$ on whole turnover. Real VAT taxes value added, and the corrected data give each firm net liability $v_i = \tau(y_i - x_i)$ with deductible inputs $x_i = \delta_i y_i$ (Section 3); profit is then $(1 - \delta_i)y_i - \tilde{c}_i(y_i)$ unregistered and $(1 - \delta_i)(1 - \tau)y_i - \tilde{c}_i(y_i)$ registered—the deductible share multiplies *both* branches, so the notch geometry of (1) carries over firm by firm, with the cash size of the jump scaled to $\tau(1 - \delta_i)T^*$ (about $\text{£}6,800$ at the mean value-added share of 40%). This drop—absent from a kink, where only the *marginal* rate changes at the threshold while average liability moves continuously—is the defining feature of a notch: it makes locating just below the cutoff strictly attractive and creates the empty band derived next.

6.2 The dominated region

The notch creates a range of turnover just above T^* that *no* firm chooses, because every point in it yields strictly lower profit than locating at T^* . A firm at the threshold earns net revenue T^* ; a registered firm at $T^* + a$ earns $(1 - \tau)(T^* + a)$. Equating the two—the cost difference across this narrow band is negligible—the upper edge of the dominated region is

$$(1 - \tau)(T^* + a) = T^* \implies a = T^* \frac{\tau}{1 - \tau}. \quad (2)$$

This is the Kleven–Waseem dominated-region width (Kleven and Waseem, 2013). At my parameters it gives $a = £85,000 \times 0.20/0.80 = £21,250$, so the dominated region is $(T^*, T^* + a) = (£85,000, £106,250)$, shown in Figure 7.

Two properties make this band a transparent index of the distortion. First, it is *exact and elasticity-free*, and—less obviously—*invariant to the firm’s input share*. Under the value-added formulation above, the indifference condition at the edge of the region is $(1 - \delta_i)(T^*) = (1 - \delta_i)(1 - \tau)(T^* + a)$: the deductible share $(1 - \delta_i)$ cancels, so every firm with positive value added faces the *same* dominated width $a = T^* \tau / (1 - \tau)$, whatever its input intensity. (A per-firm width built from the net remittance rate, such as $T^* \tau_{oi} / (1 - \tau_{oi})$, corresponds to a different technology in which marginal turnover carries no input cost, and is not used here.) The width depends on τ and T^* alone—not on the cost specification, any behavioural elasticity, or any smoothing parameter. Who escapes the notch is a matter of scope, not of input shares (the invariance is a property of the formulation: unregistered profit carries the same deductible share, so no input VAT goes unreclaimed below the threshold; if deregistered firms instead bore VAT on their inputs, the width would shrink in the input share and vanish for shares above one half—the input-reclaim margin that drives voluntary registration, Section 2): voluntary registrants escape by choice, and net-repayment traders have no positive remittance to jump (the Scope paragraph above). Second, the region is an order of magnitude wider than the $\sim £1k$ region implied by a smoothed *kink* approximation, and it spans a populous part of the firm distribution: about 148,000 weighted firms in the calibrated synthetic population lie in the £85,000–£106,250 band (`results/dominated_region_mass.txt`). Because the generator has no location choice, this counts the firms the data *place* in the band—indexing the size of the population exposed to the distortion—rather than firms that optimally locate there; the count is conditional on the uniform within-band fill inside the coarse HMRC bands (Section 3) and would shift under a different within-band shape. It is the width a , not the count, that is exact.

How each reform changes the dominated region. Because the width $a = T^* \tau / (1 - \tau)$ depends on the threshold and the rate alone, the effect of each schedule reform of Section 4.1 on the dominated region is exact and follows directly from the notch arithmetic of (2) and (3), with no dependence on the cost specification, any behavioural elasticity, or any simulated re-optimisation. Raising the threshold to £100,000 leaves the rate unchanged and therefore *relocates* the region without altering its width: it is the same £21,250 band, transported to the new cutoff. A banded reduced rate does *not* remove the distortion. It reduces the *largest single* notch—the jump at T^* falls from $\tau T^* = £17,000$ to $r T^* = £12,750$ at a 15% band rate—but, because the band reverts to the standard rate τ at its upper edge $T_1 = £105,000$, it introduces a *secondary* notch there, and hence a second dominated region the simple “ a shrinks with τ ” reasoning omits. The primary region at T^* (the band rate r on whole turnover) has the usual width $a = T^* r / (1 - r)$: $£85,000 \times 0.15/0.85 = £15,000$ at 15% and $£85,000 \times 0.10/0.90 = £9,444$ at 10%. The secondary region at the band top—where the schedule steps from r back to τ on whole

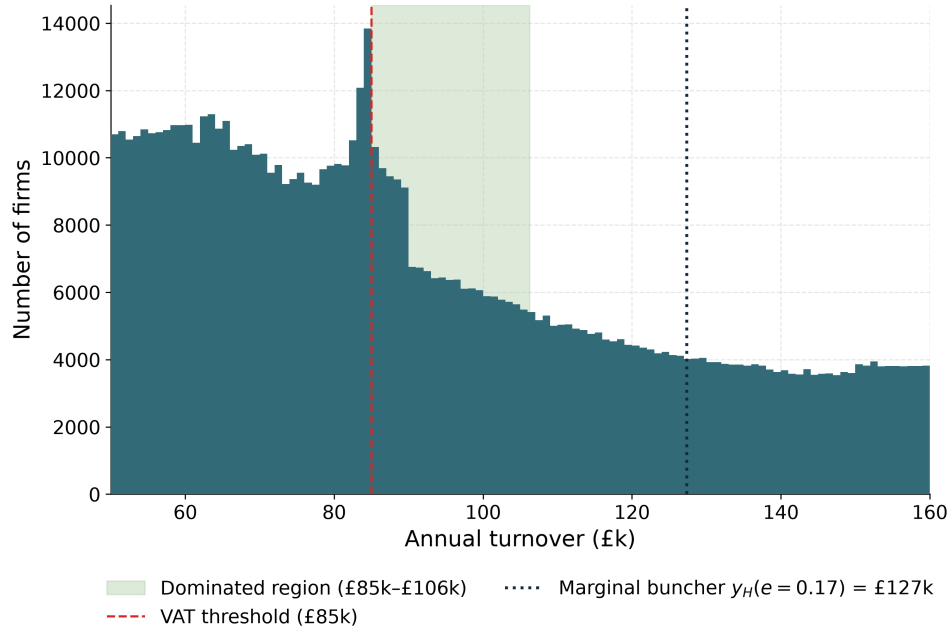


Figure 7: The £85,000 notch and its dominated region. The observed synthetic turnover density is shown with the dominated region (£85,000, £106,250) shaded; no firm optimally locates there. The marginal buncher at an assumed turnover elasticity $e = 0.17$ sits at $n_H = £127,382$ (Section 7). The dominated-region width $a = £21,250$ is exact and independent of any elasticity.

turnover—has width

$$a' = T_1 \frac{\tau - r}{1 - \tau}, \quad (3)$$

giving $\text{£}105,000 \times 0.05 / 0.80 = \text{£}6,562$ at 15% and $\text{£}105,000 \times 0.10 / 0.80 = \text{£}13,125$ at 10%. The *total* dominated-turnover width is therefore $a + a' = \text{£}21,562$ at 15% and $\text{£}22,569$ at 10%—both marginally *above* the $\text{£}21,250$ baseline, not below it. The reduced rate splits one wide empty band into two narrower ones rather than shrinking the total: it lowers the largest single notch (to $\text{£}12,750$ at 15%) at the price of adding a secondary notch of $(\tau - r)T_1 = \text{£}5,250$. The firm counts confirm this (`results/dominated_region_mass.txt`): the primary regions hold about 111,900 (15%) and 77,300 (10%) weighted firms, the secondary regions add about 34,200 and 65,500, and the totals—146,100 and 142,900—sit within 3.3% of the 147,700-firm baseline band. A graduated taper, by contrast, has *no* notch and hence no dominated region: marginal net revenue declines smoothly through the band rather than off a cliff, so the band of dominated turnover—and with it the discrete incentive to suppress turnover—is removed entirely. On the dominated-region metric, then, only the taper removes the distortion; the reduced rate splits it and the level move relocates it. The metric indexes the misallocation distortion only—what it omits, and how the reforms compare on cost, is taken up in Section 8. The dominated region is the paper’s exact, elasticity-free measure of the distortion the notch imposes, and the analytic statement above is exactly how much of that distortion each reform of Section 4.1 removes.

7 Behavioural reform costs

The reforms of Section 4 are costed *statically*: each schedule is re-applied to a fixed turnover distribution, and the revenue change is the mechanical reclassification of firms with no response in turnover. Section 5 shows that the synthetic data contain no behavioural signal from which a response could be estimated, and Section 6 supplies the one object that needs no behavioural calibration at all: the exact dominated region. What is missing from this sequence is the behavioural layer itself—a costing that lets firms re-optimize turnover when a reform changes the effective rate they face. This section supplies that layer for an *assumed* turnover elasticity e , reported as an e -sensitivity range rather than a point estimate, and nesting the static costing exactly in its $e \rightarrow 0$ limit. Its headline is deflationary, and usefully so: on the intensive margin the static costings are robust. A pure threshold rise has *exactly* zero intensive-margin revenue offset at every e , and the reduced-rate bands’ offsets stay under 6% of the static cost across the sweep.

7.1 An iso-elastic simulator on the intensive margin

I adapt the iso-elastic model of Kleven and Waseem (2013), in the sufficient-statistics tradition of Saez (2010), writing the tax on *value added* rather than on turnover. (Liu et al.

(2021) caution that turnover-tax bunching formulae do not carry to VAT unmodified, because input costs intervene; the value-added statement below is the adaptation this requires.) In this statement—formulation A—a firm of ability n_i chooses turnover y to maximise

$$\pi_i(y) = (1 - \delta_i)(1 - \tau f(y))y - C(y; n_i, e), \quad C(y; n, e) = \frac{n}{1 + 1/e} \left(\frac{y}{n}\right)^{1+1/e}, \quad (4)$$

where $\delta_i \in [0, 1)$ is the firm's *deductible-input share* of turnover, so that value added is $(1 - \delta_i)y$; τ is the statutory VAT rate; and $f(y) \in [0, 1]$ is the fraction of τ that the schedule levies on the firm once registered ($f = 0$ below the threshold, $f = 1$ for a fully registered firm under the standard rate). Here $e > 0$ is the elasticity of turnover with respect to the net-of-tax rate.

It is essential to read C correctly. Because the deductible inputs $\delta_i y$ are already netted out of revenue through the $(1 - \delta_i)$ factor, $C(y; n_i, e)$ must be understood as the firm's *own-factor production cost*—the labour and capital effort the firm itself supplies—and *not* its total cost of production. Reading C as total cost would double-count the bought-in inputs, which the value-added factor has already removed. Marginal own-factor cost is $(y/n)^{1/e}$, so the undistorted optimum is $y = n$: ability n is the firm's frictionless turnover. A firm registered under a flat rate ($f \equiv 1$) chooses

$$y^* = n[(1 - \delta)(1 - \tau)]^e, \quad (5)$$

so the retained wedge facing the firm is the product $(1 - \delta)(1 - \tau)$ of the value-added share and the statutory net-of-tax rate, and $d \ln y^* / d \ln[(1 - \delta)(1 - \tau)] = e$: e is exactly the knob that governs the intensive response. Figure 8 traces the formulation-A profit and its optima across the notch and taper schedules for three deductible shares.

The response ratio is free of the deductible share. The reform costing needs each firm's optimal turnover under a counterfactual schedule, *given* that its observed turnover is optimal under the baseline notch. Formulation A makes this a one-line object. Inverting (5) at the baseline (fraction f_0) and re-solving at the reform (fraction f_1 on the firm's region of the schedule), the ability and the deductible share cancel in the ratio:

$$y_i^* = y_{\text{obs},i} \left(\frac{1 - \tau f_1}{1 - \tau f_0} \right)^e. \quad (6)$$

The intensive response depends only on e , the statutory rate, and the schedule fractions—not on δ_i , not on the firm's net remittance rate, and not on the recovered ability level. Ability recovery ($n_i = y_{\text{obs},i}$ below the threshold, $n_i = y_{\text{obs},i} / [(1 - \delta_i)(1 - \tau)]^e$ above) is an *accounting anchor*—it rationalises the observed allocation given e and reproduces the baseline by construction—but no reform number below depends on it.

Region-confined scope. The margin priced here is purely *intensive*: each firm re-optimises within the region of the reform schedule that contains its observed turnover, per (6),

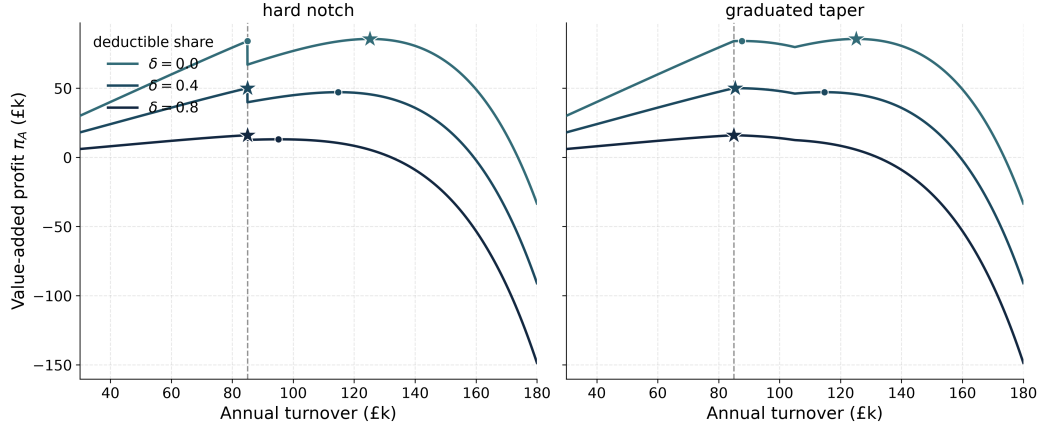


Figure 8: Optima of the formulation-A profit function (4) at $\tau = 0.20$, $n = £130k$, $e = 0.17$, under the registration notch (left) and the graduated taper (right), for deductible-input shares $\delta \in \{0.0, 0.4, 0.8\}$. Dots mark local maxima, the star the global optimum, and the dashed line the £85,000 threshold. A higher δ scales the profit curve down and pulls the registered optimum $y^* = n[(1 - \delta)(1 - \tau)]^e$ toward the threshold, eventually tipping the global optimum to bunching at the notch.

clipped at its region’s edges. Relocation across a notch—bunching below a threshold from above, or de-registering past a band edge—is the *extensive* margin, which is the separate analytic object of Section 6 (the dominated region and the marginal buncher) and is deliberately out of scope: the synthetic data contain no information about it (Section 5), and pricing it would require assumptions about re-bunching dynamics the paper has declined to make. Confining the response to the firm’s own region also makes the $e \rightarrow 0$ limit collapse onto the static costing *exactly*, with no solver tolerance.⁶

Conditional on e . The elasticity e is not identified from these data, so every behavioural magnitude below is conditional on the assumed e , and the sweep is a sensitivity analysis rather than a set of forecasts. The sweep values $e \in \{0.05, 0.17, 0.32\}$ are assumptions: 0.05 is the external anchor—Kleven and Waseem (2013) report lower-bound structural elasticities mostly in $[0.05, 0.15]$ —and 0.17 and 0.32 are chosen to span the range used in the threshold literature; the published UK VAT-base elasticities of Liu et al. (2021) (roughly 0.09–0.14) fall between the anchor and the midpoint. A reader may substitute any preferred value directly. The robust, e -free headline objects remain the static costs of Section 4 and the dominated region of Section 6.

⁶An earlier version of this simulator computed the registered optimum by damped fixed-point iteration on the discontinuous schedule. For any firm whose ability straddles a reform notch that iteration has no fixed point—the iterate oscillates across the notch and the reported turnover is an artifact of the iteration count. The reported behavioural offsets in the earlier draft (for example –£292m against a static –£508m for the raise at $e = 0.17$) were dominated by this artifact. The corrected, region-confined solve is closed form; the replication repository documents the change.

7.2 Results

Table 3 reports the behavioural cost of each of the three flat-rate reforms—the raised threshold and the two reduced-rate bands—measured as the change in revenue relative to the £85,000-notch baseline, alongside the static figure of Table 2, for each e . For context, the marginal buncher $n_H(e)$ —the highest ability that still finds bunching at the threshold optimal, the extensive-margin object held fixed here—sits at £112,795, £127,382 and £143,527 for $e = 0.05, 0.17, 0.32$.

Table 3: Behavioural revenue effect of the three flat-rate reforms relative to the £85,000-notch baseline, by assumed turnover elasticity e , with the static ($e \rightarrow 0$ limit) figure for comparison. Each behavioural figure re-solves every firm’s optimum within its schedule region per (6), holding e fixed. The final column reports the number of firms whose optimal turnover moves by more than £500 at $e = 0.17$; slightly more firms (111,940 and 140,826) see their effective rate change, the difference being firms whose optimum shifts by less than the tolerance. Estimates are conditional on the assumed e (Section 7.1). The graduated taper is excluded; see the paragraph below.

Reform	Static	$e = 0.05$	$e = 0.17$	$e = 0.32$	Firms ($e=0.17$)
Raise threshold to £100,000	–£753m	–£753m	–£753m	–£753m	108,973
Reduced rate 10% [£85k, £105k]	–£484m	–£481m	–£475m	–£467m	137,944
Reduced rate 15% [£85k, £105k]	–£242m	–£240m	–£235m	–£228m	137,944

Note. The graduated taper is omitted: its rate varies continuously with turnover, so its intensive response is outside the region-confined solve, and its marginal-rate channel is not represented by the flat-rate first-order condition. Its e -free figures—the static cost of –£520m and the exact removal of the dominated region—are in Sections 4 and 6.

A level rise has no intensive-margin offset. The raise-to- £100,000 row is identical to its static cost at every e , and this is a result, not a failure to respond. Partition the firms. Those with observed turnover below £85,000 face $f = 0$ before and after; nothing changes. Those released by the raise ($y_{\text{obs}} \in [\text{£85k}, \text{£100k})$) see their rate fall to zero and expand per (6) by the factor $(1 - \tau)^{-e}$ (about +3.9% at $e = 0.17$)—but their expansion is *untaxed*: they have left the VAT base, so the revenue loss is their full baseline remittance, exactly the static figure. Those above £100,000 face $f = 1$ before and after; nothing changes. The only candidates for a taxed response are released firms whose expansion would carry them across the new threshold, and for every such firm the new notch’s bunching condition binds (its ability lies strictly below the new marginal buncher), so it optimally stops at the threshold and remits nothing. Base broadening from a threshold rise is therefore real economic activity but generates no VAT: conditional on this model, the static costing of a level move is not an upper bound that behaviour erodes—it is the behavioural costing on the intensive margin. The simulator asserts this invariance programmatically (to £0.1m at each swept e).

Reduced-rate offsets are second order. A reduced-rate band lowers the rate on band firms from τ to r , so each scales up by (6) with $f_1 = r/\tau$: about +2.0% at the 10% rate and $e = 0.17$, and about +1.0% at 15%. Remittance scales with turnover, so the offset is roughly the scale gain times the band’s (rate-reduced) remittance—£9m against a static £484m at $e = 0.17$ for the 10% band, reaching £17m (3.5%) only at $e = 0.32$. The direction is uniform—a larger e makes the bands slightly cheaper—but the magnitude never approaches the static term. Figure 9 traces the costs across the full range of e : the raise is exactly flat, the bands nearly so.

Validation. Three properties are asserted programmatically in the released code on every run: the baseline solve reproduces each firm’s observed turnover exactly (the accounting anchor is internally consistent); the $e \rightarrow 0$ behavioural cost equals the static cost to £0.1m for every priced reform; and the solved response satisfies $d \ln y^*/d \ln(1 - \tau f) = e$ to machine precision. The static costing is thus the $e = 0$ corner of the same machine, not a separate calculation.

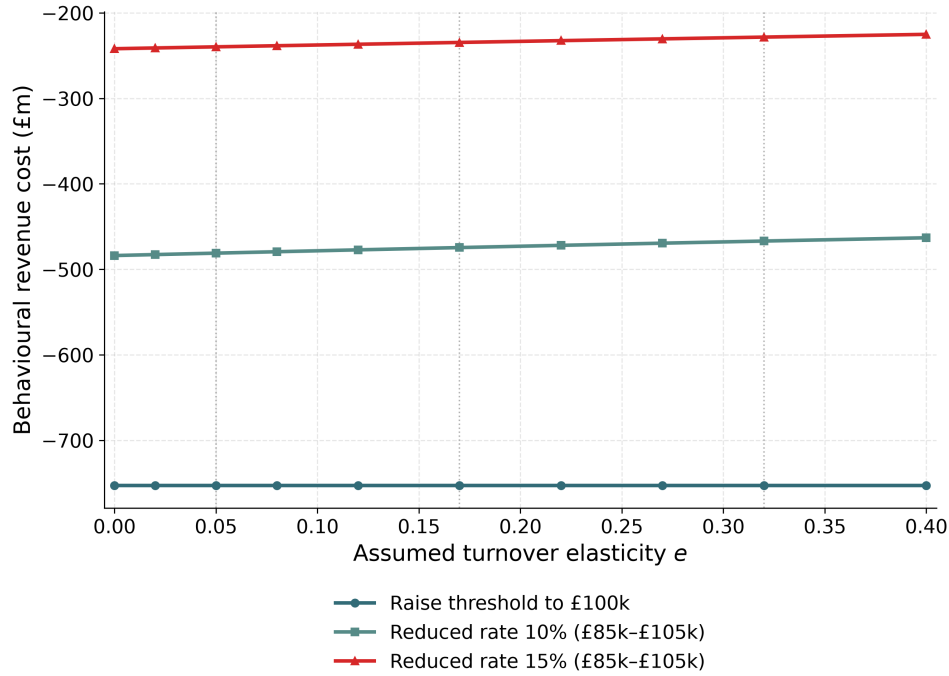


Figure 9: Behavioural reform cost against the assumed turnover elasticity e , for the three flat-rate reforms. Each curve is the change in revenue relative to the £85,000-notch baseline; the intercept at $e = 0$ is the static cost of Table 2. The raise-to-£100,000 line is exactly flat (no intensive-margin offset); the reduced-rate lines decline by under 8% of their static values across the plotted range. Dotted lines mark the swept values $e \in \{0.05, 0.17, 0.32\}$. Estimates are conditional on the assumed e (Section 7.1).

This section completes the costing framework, and its contribution is a robustness statement. Where Section 4 prices each reform mechanically and Section 6 measures

Table 4: Summary of headline estimates, consolidating results defined in the body sections cross-referenced below. Rows 1–2 are 2025–26 figures (the anchor on the 2023–24 population, the sweep on the 2024–25 population at its £90,000 baseline); the remaining rows refer to the £85,000 data-year baseline.

Object	Symbol / definition	Value	Source
Static cost, £85k to £90k anchor	dereg. / retention conventions	–£323m / –£184m	Section 4
Static sweep, £90k baseline	per £5k step	£230m–£265m	Table 1
Dominated region	$a = T^*\tau/(1 - \tau)$	£21,250	Section 6.2
Firms in the dominated band (weighted)	—	$\approx 148,000$	Section 6.2
Raise to £100k, static and behavioural	invariant in e	–£753m	Sections 4, 7
Graduated taper, static	—	–£520m	Section 4.1

the distortion each removes, the intensive-margin layer shows that naming an elasticity barely moves the static numbers: not at all for a level move, and by under 6% for the reduced-rate bands. The behavioural uncertainty that could move these costings materially sits on the *extensive* margin—registration, de-registration, and location choice at the notch—which the synthetic data cannot identify and which this paper prices only through its exact analytic geometry.

8 Conclusion

This paper contributes an open, reproducible firm-level microsimulation for analysing the UK VAT registration threshold, released in full—including its synthetic-data generator—for replication. Its value is as usable policy infrastructure rather than a structural identification result: I make no claim to recover policy-invariant firm primitives, unlike [Garicano et al. \(2016\)](#) and [Gourio and Roys \(2014\)](#). The analysis is anchored to the 2023–24 data year, in which the statutory threshold was £85,000, with a second 2024–25 population (threshold £90,000) for the forward sweep. Table 4 collects the headline objects.

Findings. Four results stand out. First, the static pipeline prices level and schedule reforms on a common £85,000 / £184.8bn base (–£753m to raise the threshold to £100,000, –£520m for a graduated taper, –£484m and –£242m for 10% and 15% banded reduced rates), and its anchor costing of the £85,000 to £90,000 move tracks HMRC’s published figure once the deregistration convention is stated: the full-deregistration convention overshoots (–£323m against the published –£185m in 2025–26, consistent with the official costing’s phased registration response), the retention convention runs close to the published 2025–26 figure (–£184.1m against –£185m; agreement finer than the calibrated liability level supports, Section 4), and both conventions reproduce the official profile’s narrowing and its 2028–29 sign change under the uprated counterfactual threshold path.

Second, the notch’s distortion is captured exactly by its dominated region, the [Kleven](#)

and Waseem (2013) width $a = T^*\tau/(1-\tau) = £21,250$ —an accounting identity rather than a new structural result, but one that is invariant to the firm’s input share under the value-added formulation and spans a populous range of roughly 148,000 weighted firms. Setting cost against distortion on this metric: raising the threshold carries the largest static cost and relocates the region without changing its width; a banded reduced rate lowers the largest single notch (from £17,000 to £12,750 at 15%) but adds a secondary notch at the band top, leaving the total dominated width essentially unchanged (£21,562 at 15%, £22,569 at 10%); only the graduated taper removes the dominated region, at a static cost of –£520m against the 10% band’s –£484m, which removes none of it. The metric indexes misallocation only; the limitations paragraph below states what it omits.

Third, the synthetic data carry a transparency lesson: sub-band structure in band-calibrated microdata is exactly what its targets put there, so it supports no bunching inference in either direction. The 2023–24 population’s near-threshold profile is calibrated to the OBR’s published £1,000-band data, and the estimator reads it back as target-inherited excess mass ($E = 7,933$); a placebo that regenerates without the fine targets returns zero; and the 2024–25 vintage, which has no fine targets, shows the coarse £90,000 band edge alone yielding an excess-bunching statistic within 10% of the administrative estimate—coincidental, specification-unstable, and behaviour-free. A redesigned power test confirms the estimator detects moderate genuine signals (recovering half to two-thirds of 5,000–8,000-firm injections, attenuated, never inflated). An earlier draft of this paper reported “reproducing” the administrative density step at £85,000 when a liability mis-scaling manufactured it; the paper now documents the artifact, its correction, and the replacement of optimiser side-effects with published fine-band targets as the sharpest demonstration of the lesson. The behavioural evidence on threshold bunching remains the administrative estimate of Liu et al. (2021).

Fourth, the behavioural layer’s contribution is a robustness statement: on the intensive margin, naming an elasticity barely moves the static costings. A pure level rise has exactly zero intensive-margin offset at every e —released firms leave the VAT base, so their expansion is untaxed, and boundary-crossers optimally bunch below the new threshold—and the reduced-rate bands’ offsets stay under 6% of the static cost across $e \in \{0.05, 0.17, 0.32\}$ (the 10% band falls from £484m to £475m at $e = 0.17$). The graduated taper is priced only statically: its continuously varying rate is outside the flat-rate machinery. The behavioural uncertainty that could move these costings materially sits on the extensive margin—registration and location choice at the notch—which these data cannot identify and the paper prices only through its exact analytic geometry.

Limitations and future work. The model abstracts from several margins, each stated where it binds. It has no firm location choice and no extensive-margin response; voluntary registration enters only as a registration-rate parameter and a retention sensitivity, not a choice model; net-repayment traders are outside the liability model (Section 3); the statutory rolling registration test is approximated by a data-year turnover test on total rather than taxable turnover; the near-threshold profile rests on two stated assumptions (the 2023–24 interpolation between the OBR’s 2019–20 outturn and 2025–26 frozen-

threshold projection, and the shape-only, side-consistent treatment of the OBR bins, whose chart universe differs from the ONS frame); and the calibrated revenue bases sit about 5.8% above the HMRC liability-by-band target sums, with liability-by-band calibration scores of 91.3% and 92.0% (Section 4, footnote). The dominated-region width indexes the misallocation distortion only and omits the compliance and administrative-cost saving that a *higher* threshold delivers by removing small firms from VAT—the classic rationale for a high registration threshold (Keen and Mintz, 2004)—so the efficiency comparison is partial; it likewise books no compliance saving to the taper or reduced rate, which keep every band firm inside the VAT system, and a taper’s administrability (an effective rate that depends on annual turnover cannot be charged on invoices, so a real design would operate as remittance-side relief) is a design question the costing does not address. The menu reforms are costed only in aggregate: because VAT liability by sector is excluded from the calibration optimiser (Section 3), the framework does not yet report the sectoral incidence of a taper or reduced rate, a cut of obvious policy interest, and the split of band-total liability across sub-band turnover—which the anchor and relocation costings use—rests on the generator’s common input-share distribution rather than on sector-composition data. The analysis is partial-equilibrium, abstracting from pass-through into prices (Benedek et al., 2015; Copestake et al., 2026). The chief outstanding empirical task is to *identify* the turnover elasticity from administrative micro-data, which would turn the *e*-sensitivity range into a point estimate; a welfare/MVPF accounting would net compliance costs against the misallocation distortion and support revenue-neutral schedule designs. Above all, validation against administrative firm-level data with genuine location choice would replace the present internal-consistency checks with out-of-sample tests.

References

- Asatryan, Zareh and Andreas Peichl (2017) “Responses of Firms to Tax, Administrative and Accounting Rules: Evidence from Armenia,” CESifo Working Paper 6754, CESifo, https://ideas.repec.org/p/ces/ceswps/_6754.html.
- Bashir, Muhammad, Zehra Farooq, Usama Jamal, and Mazhar Waseem (2024) “Size-Based Policies and Firm Growth: Evidence from Pakistan,” working paper, University of Manchester, https://mazharwaseem.com/static/uploads/PAK_VAT_Threshold_October_2024.pdf.
- Benedek, Dora, Ruud A. de Mooij, Michael Keen, and Philippe Wingender (2015) “Estimating VAT Pass Through,” IMF Working Paper WP/15/214, International Monetary Fund, <https://www.elibrary.imf.org/view/journals/001/2015/214/001.2015.issue-214-en.xml>.
- Bertanha, Marinho, Andrew H. McCallum, and Nathan Seegert (2023) “Better Bunching, Nicer Notching,” *Journal of Econometrics*, 237 (2), 105512, <https://doi.org/10.1016/j.jeconom.2023.105512>.
- Best, Michael Carlos, Anne Brockmeyer, Henrik J. Kleven, Johannes Spinnewijn, and Mazhar Waseem (2015) “Production versus Revenue Efficiency with Limited Tax Capacity: Theory and Evidence from Pakistan,” *Journal of Political Economy*, 123 (6), 1311–1355, <https://doi.org/10.1086/683849>.
- Blomquist, Sören, Whitney K. Newey, Anil Kumar, and Che-Yuan Liang (2021) “On Bunching and Identification of the Taxable Income Elasticity,” *Journal of Political Economy*, 129 (8), 2320–2343, <https://doi.org/10.1086/714446>.
- Chetty, Raj, John N. Friedman, Tore Olsen, and Luigi Pistaferri (2011) “Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records,” *Quarterly Journal of Economics*, 126 (2), 749–804, <https://doi.org/10.1093/qje/qjr013>.
- Copestake, Alexander, Matthieu Bellon, and Wenzhang Zhang (2026) “Supply and Demand Determinants of Heterogeneous VAT Pass-Through,” *National Tax Journal*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4207882, Forthcoming.
- Council of the European Union (2020) “Council Directive (EU) 2020/285 on the Special Scheme for Small Enterprises,” Official Journal of the European Union, <https://eur-lex.europa.eu/eli/dir/2020/285/oj>.
- Garicano, Luis, Claire Lelarge, and John Van Reenen (2016) “Firm Size Distortions and the Productivity Distribution: Evidence from France,” *American Economic Review*, 106 (11), 3439–3479, <https://doi.org/10.1257/aer.20130232>.
- Gourio, François and Nicolas Roys (2014) “Size-Dependent Regulations, Firm Size Distribution, and Reallocation,” *Quantitative Economics*, 5 (2), 377–416, <https://doi.org/10.3982/QE338>.

- Harju, Jarkko, Tuomas Matikka, and Timo Rauhanen (2019) “Compliance Costs vs. Tax Incentives: Why Do Entrepreneurs Respond to Size-Based Regulations?” *Journal of Public Economics*, 173, 139–164, <https://doi.org/10.1016/j.jpubeco.2019.02.003>.
- HM Revenue and Customs (2024) “Increasing the VAT Registration Threshold,” Tax information and impact note, GOV.UK, <https://www.gov.uk/government/publications/vat-increasing-the-registration-and-deregistration-thresholds/increasing-the-vat-registration-threshold>.
- HM Treasury (2018) “VAT Registration Threshold: Call for Evidence,” HM Treasury, March 2018, <https://www.gov.uk/government/consultations/vat-registration-threshold-call-for-evidence>.
- (2024) “Spring Budget 2024: Policy Costings,” HM Treasury, March 2024, https://assets.publishing.service.gov.uk/media/65e7920c08eef600155a5617/Published_Costing_Document_Spring_Budget_2024_Final.pdf.
- Keen, Michael and Jack Mintz (2004) “The Optimal Threshold for a Value-Added Tax,” *Journal of Public Economics*, 88 (3–4), 559–576, [https://doi.org/10.1016/S0047-2727\(02\)00165-2](https://doi.org/10.1016/S0047-2727(02)00165-2).
- Kleven, Henrik J. (2016) “Bunching,” *Annual Review of Economics*, 8, 435–464, <https://doi.org/10.1146/annurev-economics-080315-015234>.
- Kleven, Henrik J. and Mazhar Waseem (2013) “Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan,” *Quarterly Journal of Economics*, 128 (2), 669–723, <https://doi.org/10.1093/qje/qjt004>.
- Liu, Li, Ben Lockwood, Miguel Almunia, and Eddy H. F. Tam (2021) “VAT Notches, Voluntary Registration, and Bunching: Theory and UK Evidence,” *Review of Economics and Statistics*, 103 (1), 151–164, <https://direct.mit.edu/rest/article/103/1/151/97767>.
- Liu, Li, Ben Lockwood, and Eddy H. F. Tam (2024) “Small Firm Growth and the VAT Threshold: Evidence for the UK,” IMF Working Paper WP/24/33, International Monetary Fund, <https://www.imf.org/en/publications/wp/issues/2024/02/16/small-firm-growth-and-the-vat-threshold-evidence-for-the-uk-544462>.
- Nandi, Tushar and Ross Warwick (2020) “How Businesses Respond to VAT: Evidence from West Bengal, India,” TaxDev / Institute for Fiscal Studies, <https://www.taxdev.org/research-projects/how-businesses-respond-vat>.
- Office for Budget Responsibility (2023) “The Impact of the Frozen VAT Registration Threshold,” Economic and Fiscal Outlook, March 2023, <https://obr.uk/box/the-impact-of-the-frozen-vat-registration-threshold/>.
- Office of Tax Simplification (2017) “Value Added Tax: Routes to Simplification,” OTS report, November 2017, https://assets.publishing.service.gov.uk/media/5a822577e5274a2e87dc1620/Value_added_tax_routes_to_simplification_web.pdf.

- Onji, Kazuki (2009) “The Response of Firms to Eligibility Thresholds: Evidence from the Japanese Value-Added Tax,” *Journal of Public Economics*, 93 (5–6), 766–775, <https://doi.org/10.1016/j.jpubeco.2008.10.003>.
- PolicyEngine (2025) “PolicyEngine UK: Open-Source Tax-Benefit Microsimulation,” <https://policyengine.org>.
- Saez, Emmanuel (2010) “Do Taxpayers Bunch at Kink Points?” *American Economic Journal: Economic Policy*, 2 (3), 180–212, <https://doi.org/10.1257/pol.2.3.180>.
- Seely, Antony (2024) “VAT: The Registration Threshold,” House of Commons Library Briefing Paper SN00963, <https://commonslibrary.parliament.uk/research-briefings/sn00963/>.

A Appendix

This appendix collects supporting material in two parts: data construction and calibration robustness (Appendix A.1); and statistical inference for the reduced-form bunching results (Appendix A.2), including the placebo and power tests establishing that band-calibrated synthetic data cannot support bunching inference in either direction.

A.1 Data construction detail

Section 3 describes the two-stage synthetic-population pipeline, parameterised by the registration threshold, and its calibration to HMRC and ONS aggregates. This appendix records the implementation detail deferred from that account.

Each of the approximately 2.94 million firm rows carries a turnover, an input expenditure, a net VAT liability ($v_i = 0.20 (y_i - x_i)$), a calibration weight, and a registration status; the weights are chosen so the weighted totals reproduce the official targets while the population sums to about 2.5 million firms. Because the population is calibrated *to* the HMRC aggregates, agreement between my weighted outputs and those aggregates is a check of *internal consistency*, not external validation: any target used in calibration is reproduced by construction. The headline calibration accuracies (89.2% and 92.8% for the two vintages) and the anchor comparison of Section 4 are therefore consistency checks on the pipeline, not out-of-sample tests of firm behaviour.

The reported calibration score is a bounded summary statistic. For each dimension I compute $\max\{0, 1 - |\hat{T} - T|/|T|\}$, so a dimension missed by more than 100 percent receives zero rather than a negative score. The headline “overall” number is the unweighted mean of the five dimensions used as calibration targets—HMRC turnover bands, ONS population, employment bands, HMRC sector counts, and VAT liability by turnover band. VAT liability by sector is excluded from that mean and reported separately as an informational diagnostic, because the current generator does not calibrate sector-specific input/output VAT structure.

Ledger and Populace migration check. The official target tables are being moved into PolicyEngine Ledger, with Populace providing the shared synthetic-population generator. I therefore keep the paper’s archived processed CSVs as the reproduction source for the reported results, and treat the pinned Populace/Ledger path as an auditable migration check rather than a silent replacement. The snapshot used here is PolicyEngine/ledger pull request 67 at commit cd98b5c and PolicyEngine/populace pull request 223 at commit fa20daf; both upstream pull requests merged on June 30, 2026. For the 2024–25 vintage, the Ledger-backed targets match the paper’s processed numeric inputs exactly after dropping presentation-only labels, totals, and the HMRC “Unknown” column that the generator does not calibrate: six normalized source tables checked, zero mismatches, and maximum numeric difference zero. (The “2024–25 vintage” here denotes the HMRC/ONS statistical release labelled 2024–25, the same target tables the paper’s 2024–25 population calibrates to.) A 1,000-iteration Populace run from those

targets generated 2,946,015 firm rows; its validator reported a 93.8 percent overall score against the paper validator’s then-90.5 percent, a comparison that is not like-for-like (different band sets and sector-target definitions). Two caveats now attach to that pinned snapshot. First, it predates this paper’s net-liability correction (Section 3, footnote): both generators at the pin date shared the mis-scaled $v_i = y_i - x_i$ convention, so the experimental Populace firm generator inherits the same defect, which I have reported upstream; the target-parity result (zero mismatches) is unaffected, since it concerns the input tables, not the generated population. Second, the paper validator’s scores on the corrected build are 89.2 percent (2023–24) and 92.8 percent (2024–25). The comparison table and structured provenance are reproduced by `firm-microsim-populace-ledger` and checked into `results/populace_ledger_comparison.txt` and `results/populace_ledger_provenance.txt`.

Counterfactual exclusion window and polynomial degree. The no-VAT counterfactual density of Section 5 is fitted by polynomial regression on the observed density outside a manipulation window around the threshold. The baseline excludes $[T^* - 15k, T^* + 15k]$ and fits a degree-7 polynomial. Widening the window guards against contamination of the counterfactual by the behavioural response but reduces the data available for the fit; raising the degree improves in-sample fit but risks over-fitting the tails. The sensitivity of the excess-mass and bunching-ratio estimates to window width and degree, with associated uncertainty, is reported in Appendix A.2.

A.2 Inference for the bunching estimator

This appendix documents the inference procedure for the reduced-form bunching results of Section 5: the mass-conservation constraint that disciplines the counterfactual density, why no standard errors are reported, sensitivity to the polynomial degree, the exclusion window, and the generator seed, the placebo tests, and the power (recovery) test. All numbers are produced by the released estimator run on the regenerated weighted synthetic populations for both vintages; the machine-readable artifacts are `results/bunching_inference.txt`, `results/placebo_bunching.txt`, `results/recovery_bunching.txt`, and `results/seed_sensitivity.txt`. The estimator reports the geometry only—the bunching ratio b , its Liu et al. (2021)-normalised counterpart b_{LLAT} , the excess mass E , and the mass-conservation objects (Δ_R, y_R) —and computes no elasticities: the mapping from excess mass to an elasticity is not identified without functional-form assumptions (Blomquist et al., 2021; Bertanha et al., 2023), and on these data there is no behavioural mass to convert in the first place.⁷

⁷Earlier versions of the estimator also printed a CES substitution elasticity and two wedge-normalised turnover elasticities, the latter inheriting an ad hoc $\tau/2$ normalisation from a deleted smoothed-schedule model. These outputs were labelled non-identified but were still reported; they have been removed from the code and the paper.

Mass-conservation constraint

The canonical bunching estimator of [Chetty et al. \(2011\)](#) and [Kleven and Waseem \(2013\)](#) imposes that the counterfactual density conserves mass: every firm removed from above the threshold reappears as excess mass below it. An unconstrained polynomial fit on bins outside a hand-set window computes the excess mass E (below T^*) and the missing mass Δ_R (above T^*) *independently*, with no guarantee that $E = \Delta_R$ and no endogenous location of the marginal buncher. I impose the constraint in two ways. First, the counterfactual polynomial in $(y - T^*)$ is fit by an iterated procedure in which the density of the bins above the excluded region is rescaled until the fitted counterfactual integrates to the observed total mass over the estimation range $[\text{£}20\text{k}, \text{£}140\text{k}]$ (the regressor is normalised to $[-1, 1]$ for numerical stability of the high-degree basis). Second, the upper edge of the manipulation region—the marginal buncher y_R —is located endogenously by integrating missing mass upward from T^* until it equals the excess mass below,

$$\int_{T^*}^{y_R} [f^{\text{cf}}(y) - f^{\text{obs}}(y)]_+ dy = \int_{T^*-W}^{T^*} [f^{\text{obs}}(y) - f^{\text{cf}}(y)]_+ dy = E.$$

The constraint’s behaviour differs by vintage, diagnostically so. On the 2024–25 population it is satisfied in the interior ($y_R = \text{£}101\text{k}$, $\Delta_R = 11,452 \geq E$): the band-edge artifact is not a local spike. On the 2023–24 population it *fails to bind*: the walk is censored at the search cap ($y_R = \text{£}100\text{k} = T^* + W$) with $\Delta_R = 5,761 < E = 7,933$, and decomposing the missing mass locates all of it in $(\text{£}90\text{k}, \text{£}100\text{k}]$ and none in the OBR-pinned $(\text{£}85\text{k}, \text{£}90\text{k}]$ segment. The counterfactual exceeds the observed density only beyond the $\text{£}90,000$ fine-target boundary—the same calibration-band-edge mechanism as the 2024–25 exhibit, operating inside the headline estimator’s own window—so the reported y_R is a property of the target boundary, not evidence of administrative-style missing mass.

Why no standard errors are reported

Earlier drafts attached bootstrap standard errors to the bunching statistics (resampling firm rows with replacement and re-running the estimator). Those numbers are not reported anywhere in this paper, because the exercise has no estimand: the synthetic file is a deterministic construction from published targets, not a sample from a population, and row-resampling dispersion scales as $1/\sqrt{n}$ in a row count the analyst chooses. A standard error that can be made arbitrarily small by generating more synthetic rows conveys precision that does not exist. Two reportable quantities replace them. First, the specification grids below, which show how the point estimates move with the estimator’s own knobs. Second, generator-seed sensitivity (`results/seed_sensitivity.txt`): regenerating the full 2023–24 population under three seeds moves E by ± 161 firms ($\pm 2\%$), b_{LLAT} by ± 0.02 , the raise-to- $\text{£}100\text{k}$ cost by $\pm \text{£}2.1\text{m}$, and the registered base by $\pm \text{£}0.08\text{bn}$ —the pipeline’s actual Monte Carlo error. Table [A.1](#) reports the point estimates. Two conventions matter for reading them. Bins are centred on integer £k values (edges at half-integers), so the bin containing the threshold spans $\text{£}84.5\text{k}$ – $\text{£}85.5\text{k}$ and

Table A.1: Headline bunching statistics (point estimates; degree-7 polynomial, $\pm£15k$ window). Uncertainty is characterised by the specification grids and the generator-seed sensitivity, not by standard errors—see the text.

Vintage	E	Δ_R	b_{LLAT}	y_R (£k)
2023–24 (£85k, OBR fine targets)	7,933	5,761	1.91	100.0
2024–25 (£90k, coarse bands only)	11,072	11,452	1.23	101.0

its mass is attributed to the above side, attenuating E and Δ_R relative to edge-aligned binning. And E sums the *positive part* of the observed-minus-counterfactual gap in the excluded window below the threshold; the net gap there is 1,570, so the statistic reads gross rather than net excess.

Sensitivity to polynomial degree and exclusion window

Table entries for a grid of polynomial degree $\in \{5, 6, 7, 8\}$ and symmetric exclusion window $\in \{£10k, \dots, £25k\}$ are in `results/bunching_inference.txt`; the pattern matters more than the cells. On the 2023–24 population the target-inherited excess ranges over roughly 6,700–9,800 in the interior cells (degrees 6–7, windows £10–15k; 4 of the 16 grid cells), inflating in the corner cells where the polynomial confuses the profile’s shoulders with its counterfactual. On the 2024–25 population the band-edge artifact swings four-fold across the grid (E between roughly 8,600 and 36,700; b_{LLAT} between 0.94 and 4.46) while its headline cell lands within 10% of the administrative estimate: neither specification stability nor literature-proximity certifies a synthetic bunching statistic—a caution for any bunching exercise on calibrated synthetic data.

Placebo tests

The placebo makes the target-inheritance mechanism precise by manipulating the calibration targets and re-running the estimator (Table A.2; Section 5.3 describes the constructions). The estimates track the targets exactly: the control, calibrated to the OBR fine-band profile, returns $E = 7,933$; Placebo A (smooth reweighting) returns $E = 0$; and Placebo B (no fine targets, un-stepped coarse bands, full regeneration and recalibration) returns $E = 0$. Removing structure from the targets removes structure from the estimates because the targets are the only place sub-band structure can come from: the generator has no firm location-choice.

Power (recovery) test

The placebo establishes *specificity*: no false positive on step-free data. The complementary property is *power*. Starting from the step-free Placebo A population, I inject a Kleven–Waseem-consistent behavioural signal of known magnitude: firms in the

Table A.2: Placebo tests, 2023–24 population (£85,000 threshold).

Population	b_{LLAT}	E
Actual (control, OBR fine-band targets)	1.911	7,933
Placebo A (reweight density smooth)	0.000	0
Placebo B (no fine targets, un-stepped)	0.395	0

Table A.3: Power (recovery) test on the step-free population: Kleven–Waseem-consistent injections with true marginal buncher at £112,000. Baseline (no injection): $E = 0$.

E_{true} (injected)	E_{hat}	Recovery	$\hat{y}_R$ (true 112)
2,000	298	14.9%	85.8
5,000	2,445	48.9%	91.9
8,000	4,918	61.5%	100.0

donor region (T^* , £112k] relocate to $[\text{£}80\text{k}, T^*)$ with probability declining linearly from the threshold to zero at the true marginal buncher $y_R^{\text{true}} = \text{£}112\text{k}$, so the missing-mass region extends £12k *beyond* the exclusion window—as with genuine bunching—and the counterfactual fit confronts genuinely depressed above-window bins. Recovery is scored with the headline estimator alone (Table A.3).⁸

Recovery rises with signal size—half to two-thirds of moderate (5,000–8,000-firm) injections, while a small 2,000-firm injection sits near the detection floor—and the estimator *under*-states the truth rather than inflating it. The attenuation has a measured mechanical source: the depressed bins between the window top (£100k) and the true marginal buncher (£112k) enter the counterfactual fit and pull it down, truncating measured missing mass and drawing $\hat{y}_R$ toward the threshold. The implication is the one Section 5 turns on: the target-inherited excess on the 2023–24 data is read with the same attenuation a genuine signal would face, so the estimator, if anything, under-states the targeted profile; and no attenuation argument can turn the 2024–25 band-edge artifact into behaviour. Both tests are part of the released code and are fully reproducible.

⁸An earlier version of this exercise relocated mass entirely *within* the exclusion window. The fitted counterfactual is identical across such injections by construction (it is fit outside the window and rescaled by total mass, which relocation preserves), so that version measured deposit bookkeeping, not power, and its stated attenuation mechanism was wrong. The redesigned test supersedes it.