Tax Evasion under Progressive Fines: Revenue, Deterrence, and Incidence*

Duccio Gamannossi degl'Innocenti[†] [‡] Antoine Malézieux[§]
Julie Rosaz[¶] Annalisa Tirozzi[‖]

September 2026

Abstract

Progressive penalties are common in real-world deterrence systems, yet the evidence on fines mainly concerns raising their level rather than steepening their gradient. We compare progressive and linear fine schedules in a laboratory tax-evasion game with complementary between- and within-subject designs. The implemented progressive schedule increases total revenue by about 9% between subjects and, under additivity, about 7% per switch within. Between subjects, average evasion does not change detectably. Even the largest evasion response our equivalence analysis permits would move expected revenue by less than 40% of the unadjusted between-subject difference. Because most evaded income falls in the highest-penalty brackets, the schedule also raises the evasion-weighted fine rate, which accounts for much of the aggregate gain. Once the level difference is rescaled away, curvature alone shifts revenue from smaller to larger offenses at fixed behavior. Revenue also rises among subjects predicted to evade most, the only ones for which our model signs the effect on total revenue. Overall, the schedule we implement raises and reallocates revenue more clearly than it changes evasion. Policymakers seeking to reduce evasion should pair such reforms with complementary measures that the additional revenue can help finance.

Keywords: progressive fines, tax evasion, sanction design, incidence, notches.
JEL Classification: C91; H22; H26; K42.

*We thank Roberto Bonfatti, Andrea Guido, Enrico Rettore and Quentin Thevenet. This research was supported by internal departmental funds from the Department of Economics and Statistics (DISES), Università degli Studi di Napoli Federico II. The authors declare no competing financial or personal interests.

[†]Corresponding author. Department of Economics and Management, University of Padua, Via del Santo 33, 35123 Padua, Italy (duccio.gamannossi@unipd.it).

[‡]Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA (dgamannossi@gov.harvard.edu).

[§]Université Bourgogne Europe, Burgundy School of Business, CEREN EA 7477, F-21000 Dijon, France (antoine.malezieux@bsb-education.com).

[¶]Université Bourgogne Europe, Burgundy School of Business, CEREN EA 7477, F-21000 Dijon, France (julie.rosaz@bsb-education.com).

[‖]Department of Economic Policy, Università Cattolica del Sacro Cuore, Via Necchi 5, 20123, Milan, Italy (annalisa.tirozzi@unicatt.it).

1 Introduction

Progressive penalties are a common feature of deterrence systems. Tax-evasion penalties range from 20% to 140% of the underpaid tax in Italy and from 10% to 80% in France. In the United States, simple mistakes face a 20% penalty while proven evasion can trigger 75%. The amounts at stake are large: in 2024, tax enforcement identified €14.8 billion in additional tax in Italy, €16.7 billion in taxes and penalties in France, and $29.0 billion in recommended additional tax in the United States (Ministero dell'Economia e delle Finanze, Dipartimento delle Finanze, 2025; Direction Générale des Finances Publiques, 2025; Internal Revenue Service, 2025). Despite this institutional relevance, little is known about what the shape of the fine schedule does: whether it deters evasion and, if not, where along the offense distribution it moves enforcement revenue and behavior. Since Friedland et al. (1978), the experimental literature has mainly varied audit probabilities and fine levels rather than the penalty gradient. This paper studies the incidence margin: when sanction intensity is reallocated from small to large offenses, where is expected enforcement revenue collected, who responds, and how does aggregate evasion change?

In the laboratory tax-evasion game detailed in Section 2, we compare two fine schedules: a progressive piecewise-linear one and a linear benchmark. The progressive multiplier ranges from 1 to 5 times the unpaid tax, so low evasion is penalized lightly relative to high evasion. The linear benchmark uses a multiplier of 3.25, a flat counterpart close to standard values in the experimental literature. Given how much subjects evade and which brackets those amounts fall in, the progressive schedule applies a higher fine rate per unit evaded, so the implemented change moves shape and level together. We treat that bundle as the policy-relevant treatment, because no schedule can be calibrated in advance to equalize the fine burden.

In the between-subject design (T1–T2) we estimate the effect of facing the progressive (T1) rather than linear (T2) schedule throughout. In the within-subject design, T3 and T4 expose the same subjects to both schedules in opposite randomized orders. The pooled T3/T4 estimate compares the pre–post changes across those orders. Each design is exposed to a different source of bias: individual differences between subjects, timing within them. We therefore treat the two as complementary. Our main claims are the ones both designs support, and we mark the places where they diverge.

The model we provide in Section 3 makes explicit why progressivity need not have uniform effects. High-evasion types face higher marginal sanctions than under the linear benchmark, so progressivity should weakly raise compliance, declared-tax revenue, and total expected revenue for them, while leaving fine revenue ambiguous. For low- and mid-evasion types, the model delivers ambiguous individual predictions, since two

forces compete: locally weaker fines push evasion up, while attraction to bracket thresholds pulls declarations toward the cutoffs. The aggregate effect then depends on the population's type composition and the responses of each type. Heterogeneity analysis is empirically delicate because realized position on the progressive schedule is chosen by the subject: grouping by realized progressive-round evasion would condition on a treatment-affected outcome, the standard endogenous-stratification problem in randomized experiments (Rosenbaum, 1984; Frangakis and Rubin, 2002; Abadie et al., 2018). We therefore fix predicted evasion groups (Low, Mid, and High) using characteristics and elicited measures external to realized progressive-round evasion. We then examine whether the results persist under a partition based only on pre-treatment demographics, following work on endogenous stratification and post-ML heterogeneity inference (Athey and Imbens, 2016; Chernozhukov et al., 2025; Imai and Li, 2025). Section 4 provides the details on the estimands, the first-stage machine learning, and the diagnostics.

Total revenue rises in both designs, as Section 5 shows. In the between-subject design, facing the progressive schedule throughout raises total revenue by 2.055 ECU, or 9.2%. The pooled randomized difference in pre–post changes across the two schedule orders is 2.874 ECU. Under common time effects and no carryover, half of the pooled estimate is the average progressive-minus-linear effect of one switch, 1.437 ECU or about 7%. In T1–T2, exact randomization inference rejects equality of the expected-total-revenue distributions but not equality of the evasion distributions. The equivalence analysis rules out a uniform evasion shift larger than about 20% of mean evasion in the linear arm. Even at this bound, the implied change in expected revenue is less than 40% of the unadjusted T1–T2 expected-revenue difference, so the bulk of the difference is not a deterrence response.

To locate the revenue difference along the schedule, we decompose it across the brackets subjects actually choose. The upper-tail gain occurs where higher fine rates apply with little estimated change in evasion. In the bottom bracket, weaker fines coincide with more evasion and lower within-bracket revenue. Declarations also move toward the bracket thresholds: within subjects the pooled estimates show a shift to interior cutoffs, whereas the between-subject evidence is weaker.

Following mechanical and behavioral revenue decompositions (Advani et al., 2023), we hold behavior fixed and find that curvature alone does not account for the aggregate gain. At observed T2 behavior, the progressive schedule applies an evasion-weighted fine rate about 30% above the linear one. That higher level mechanically raises expected revenue by 1.593 ECU, about four-fifths of the adjusted aggregate gain. Under a burden-equivalent rescaling that sets the aggregate mechanical effect to zero by construction, the remaining fixed-behavior shape component reallocates expected revenue from the lower

brackets to the top bracket.

Total revenue rises within the predicted High group in both designs, consistent with the model's signed upper-tail prediction. The Low estimates are also positive, although the model leaves the Low group's revenue effect unsigned. Within subjects, the Low result is concentrated in T4, while the Mid estimates remain near zero. The separate T3 and T4 estimates are asymmetric, so we interpret the pooled per-switch average cautiously. Section 6 shows that these results survive alternative partitions and classifiers, and individual and round fixed effects in the within-subject design.

The role of the fine function has long been central in tax-compliance theory. After Allingham and Sandmo (1972), Srinivasan (1973) studied increasing and convex fine functions, and Yitzhaki (1974) showed that basing penalties on evaded taxes rather than undeclared income changes the model's predictions. Later theoretical work studied nonlinear tax schedules, retroactive penalties, state-dependent penalties, penalty-function progressivity, maximal penalties, and reward-penalty combinations (Pencavel, 1979; Landsberger and Meilijson, 1982; Rickard et al., 1982; Koskela, 1983; Cowell, 1985; Falkinger and Walther, 1991). Our estimates give that tradition a direct test: steepening the gradient moves revenue across the offense distribution, but it does not detectably move aggregate evasion.

Our primary contribution is controlled evidence on sanction-schedule design in tax compliance. The incidence analysis links this evidence to nonlinear-fine and optimal-enforcement theory: graduated sanctions preserve marginal deterrence while treating small offenses more leniently (Stigler, 1970; Shavell, 1992; Mookherjee and Png, 1994; Polinsky and Shavell, 2000), and observed sentencing practice tracks this gradient (Crinò et al., 2019). Our evidence is consistent with that logic but limits what it delivers in practice. At the steepness we implement, evasion is not detectably lower where the schedule fines more, while it is higher where it fines less, so the revenue gain comes from the rate rather than from behavior. Because the implemented rates exceed the penalties applied in practice, this null is unlikely to reflect fines too mild to matter. The bunching literature (Saez, 2010; Kleven and Waseem, 2013; Kleven, 2016) provides the framework for interpreting responses to the schedule's liability notches, alongside field evidence on notched speeding penalties, income-based fine discontinuities, and fine-level effects on payment, revenue, and financial burden (Traxler et al., 2018; Kaila, 2026; Traxler and Dušek, 2025; Mello, 2025; Norris and Rose, 2024). Related experimental work shows that institutional design can matter beyond average incentives: Caffera et al. (2021) compare linear and convex fines in environmental compliance, and Heinemann and Kocher (2013) show that tax compliance depends on both tax progressivity and reform direction. Our results extend that design perspective to the sanction schedule itself, where the margin

that moves is the incidence of enforcement rather than its aggregate deterrent effect.

2 Experimental design

The experiment has three parts: a 20-round Tax Evasion Game, an incentivized risk-elicitation task, and a socio-demographic questionnaire. The fine schedule is the only treatment variation. The game follows the standard design (Malézieux, 2018; Alm and Malézieux, 2021), with the same decision environment repeated in every round.[1] Instructions are framed in tax-related terms (see Appendix A.2).

In each round, participants receive an initial income of 125 ECU.[2] The income level is fixed to abstract from wealth effects, and endowing participants, rather than having them earn income, has been shown not to materially alter evasion decisions (Alm and Malézieux, 2021). Participants can declare any amount between 0 and 125 ECU, and their declared income is taxed at a linear rate $\tau = 20\%$. Collected taxes are invested in a research fund, and participants do not benefit from them. Audits occur with a constant probability of 20%, so the number of realized audits varies across rounds. If a participant is audited, the fine equals the fine rate q times the unpaid tax, where q varies with the treatment. Two rounds are randomly drawn for final payment.

2.1 Fine schedules and treatment arms

In the progressive regime, evasion is fined in five brackets of 25 ECU each, with the multiplier q rising from 1 times the unpaid tax in the bottom bracket to 5 in the top. Given the fiscal parameters, the schedule is devised to maximize the spread of multipliers while keeping fines within the round's income, so earnings cannot turn negative. This range covers the penalties applied in real-world tax systems, with the top bracket well above them. In the linear regime, q is fixed at 3.25, close to the average fine implemented in the literature, i.e., 3.01 in Alm and Malézieux (2021).

Across all linear-schedule rounds, applying the progressive schedule to the observed choices would yield an average multiplier of 3.24 across positive-evasion occurrences, essentially the linear 3.25. However, because the highest multipliers apply at high evasion, weighting by evaded ECUs raises this average to 4.22.

Thus, the progressive schedule is roughly neutral for the typical evasion choice and about 30% harsher on the typical evaded ECU, so it shifts sanction intensity toward high

[1] In this respect, we closely follow Heinemann and Kocher (2013), who study tax-rate progressivity in a 20-round experiment.

[2] The constant exchange rate is 1.00€ per 25 ECUs.

evasion rather than equalizing the fine burden point-by-point.[3] An arm that equalizes burden cannot be designed in advance, because the equalizing factor is only known once behavior has responded to it (Subsection 4.1).

The four treatment arms combine these two regimes in between- and within-subject designs. In T1, subjects face the progressive schedule in all 20 rounds; in T2, they face the linear schedule in all 20 rounds. In T3, subjects face the linear schedule for the first 10 rounds and the progressive schedule for the last 10 rounds. T4 reverses this order, with the progressive schedule first and the linear schedule second. In the between-subject arms (T1–T2), treatment was assigned at the session-day level.[4]

2.2 Experimental protocol

After the tax task, participants complete an incentivized risk-elicitation task (Gneezy and Potters, 1997) and a questionnaire covering socio-demographics, prior experimental experience, self-reported risk willingness, spending, and savings. Appendix A.2 reports the full wording.

Overall, 490 subjects participated in the experiment (131 in T1, 127 in T2, 109 in T3, 123 in T4). The four treatment arms were conducted over 6, 7, 11, and 8 distinct session days, respectively. Treatment compliance was complete: all subjects faced their assigned schedule, completed all 20 rounds, and entered the analysis. No subjects were excluded, and no revenue outcomes are missing.

The experiment took place between January and December 2022 and was run exclusively online, because France was still under varying social-distancing rules when it was designed in late 2021. Participants were students from the Burgundy School of Business (Dijon, France), recruited through ORSEE (Greiner, 2015). They earned on average 9.33€ (SD: 5€) and were paid through Lydia, an electronic payment service that transfers funds using a cellular phone number, with a delay of about a week after participation.

3 Theoretical framework

This section develops a one-period model[5] of evasion under linear and progressive fines. The model is tailored to the experimental environment and isolates the effect of changing

[3] As a descriptive benchmark computed at observed behavior, the top bracket accounts for about 23% of observations and roughly two-thirds of evaded ECUs. Moving from the linear to the progressive schedule should therefore be read as changing the shape and incidence of sanctions, not as equalizing the average fine burden. Both figures pool all linear-schedule rounds. The T2 arm alone gives 3.18 and 4.24, reported in Panel A of Table 3.

[4] Subsection 4.4 uses this assignment structure for inference.

[5] The absence of an economically meaningful learning trend over rounds (see Figure A1) makes a one-period model a natural benchmark.

the fine schedule while holding fixed income, the tax rate, and the audit probability. Because the fine applies to evaded tax, a change in the tax rate would move wealth alone, while the fine schedule moves the relative price of evasion (Yitzhaki, 1974). It organizes the treatment comparisons and delivers the incidence, heterogeneity, and aggregate revenue predictions taken to the data.

3.1 Setup

We consider an individual with income Y, proportional tax rate τ, and audit probability p. The individual declares $X \in [0, Y]$ and evades

$$E = Y - X \in [0, Y].$$

Conditional disposable income is

$$C^{na}(E) = Y - \tau(Y - E), \qquad C^{a}(E; f) = Y - \tau(Y - E) - \tau f(E)E, \tag{1}$$

where C^{na} and C^{a} are the non-audit and audit values, and $f(E)$ is the fine schedule.

Expected utility from evading E is

$$\mathcal{U}(E; \rho, f) := (1 - p)\, U\big(C^{na}(E); \rho\big) + p\, U\big(C^{a}(E; f); \rho\big), \tag{2}$$

and the individual chooses E to maximize it over $[0, Y]$.

Throughout, function subscripts denote partial derivatives with respect to the subscripted argument, so U_c and U_{cc} are the first and second derivatives of U with respect to consumption. Utility is increasing in consumption ($U_c > 0$), with curvature allowed to vary by type (U_{cc} can be negative or positive).

Evasion raises consumption when no audit occurs but increases the sanction base under audit. The latent type ρ governs how strongly a subject trades off the non-audit gain against the audit-state loss. Our preferred interpretation of ρ is risk attitude, but more generally, ρ can stand for any stable preference trait that operates through the curvature of U.

We focus on the two fine schedules of the experimental design, writing $f(E)$ for the multiplier called q in Section 2. The linear schedule f_{lin} and the progressive schedule f_{prog} are given by

$$f_{lin}(E) = f_{lin} = 3.25,$$

$$f_{prog}(E) = \begin{cases} f_1 = 1, & 0 = B_0 \le E \le B_1 = 25, \\ f_2 = 2, & B_1 < E \le B_2 = 50, \\ f_3 = 3, & B_2 < E \le B_3 = 75, \\ f_4 = 4, & B_3 < E \le B_4 = 100, \\ f_5 = 5, & B_4 < E \le B_5 = Y = 125. \end{cases} \tag{3}$$

3.2 Characterization

On any flat segment with multiplier $\bar{f}$, audited consumption can be written as

$$C^a(E; \bar{f}) = Y - \tau(Y - E) - \tau \bar{f} E,$$

and the marginal return to evasion is

$$M(E; \rho, \bar{f}) := \tau \Big[(1-p)\, U_c\big(C^{na}(E); \rho\big) + p\,(1-\bar{f})\, U_c\big(C^a(E; \bar{f}); \rho\big) \Big],$$

so that an interior optimum satisfies

$$M(E; \rho, \bar{f}) = 0. \tag{4}$$

On each flat segment we impose strict concavity of the objective in E, that is, $M_E < 0$, so within-bracket optima are unique. We maintain this concavity condition together with the single-crossing and deterrence restrictions

$$M_\rho(E; \rho, \bar{f}) < 0, \qquad M_f(E; \rho, \bar{f}) \le 0.$$

These restrictions imply that higher ρ lowers the local incentive to evade and that stronger fines weakly reduce it. On each flat segment they give decreasing differences in (E, ρ), so the within-bracket choice is monotone in the latent type (Milgrom and Shannon, 1994). A comparison across brackets is separated by a notch and turns on utility levels rather than marginal returns, which these restrictions do not sign, so Proposition 1 states the bracket mapping as a bound rather than an exact assignment. The signs of the three M derivatives hold under standard risk-averse expected utility ($U_{cc} < 0$) and sanctions that at least recover the evaded tax ($\bar{f} \ge 1$), but we specify them as shape restrictions because we also allow risk-loving types, a relaxation the top bracket turns out to need.[6]

[6]The relative price of evasion is the audit-state loss per unit of non-audit gain, $\bar{f} - 1$. Differentiating M in $\bar{f}$ gives a direct price term in U_c and a curvature term in U_{cc}, the latter from lower audited consumption. The two reinforce under risk aversion and conflict when $U_{cc} > 0$.

Proposition 1 *Assume $U_c > 0$ and, on each flat segment, maintain $M_E < 0$, $M_\rho < 0$, and $M_f \leq 0$. Assume also that the sign-bracketing conditions of Appendix A.3.1 hold, so that the type cutoffs used below exist. Under these conditions:*

1. *Under the linear schedule $f(E) \equiv f_{lin}$, the problem has a unique optimizer $E^{lin}(\rho) \in [0, Y]$, with sufficiently high-ρ types selecting full compliance and sufficiently low-ρ types selecting full evasion. On the interior, $E^{lin}_\rho < 0$ and, for flat-rate comparisons in the fine level, $E^{lin}_f \leq 0$.*

2. *Under the progressive schedule, the optimizer is selected from the finite set of within-bracket interior candidates and bracket thresholds. Thresholds at which the return to additional evasion is weakly positive just below and weakly negative just above are locally optimal and can generate bunching when the type distribution has positive density nearby.*

3. *Let $\mathcal{G}_k$ collect the types whose progressive choice falls in bracket k. These sets partition the type space. The type cutoffs bound that choice. A type above $\rho^{B_k}_{prog}$ does not evade above B_k, and because the cutoffs are ordered the bound falls from $k = 5$ to $k = 1$ as ρ rises. Within the bound a type selects an interior candidate or a bracket threshold, and types above ρ^0_{prog} select full compliance ($E = 0$).*

Proof. See Appendix A.3.1.

The bracket mapping shows that the regime switch moves incentives in three directions: relative to the flat $f_{lin} = 3.25$, the progressive multiplier is lower in the two bottom brackets ($f_1 = 1$, $f_2 = 2$), similar in the middle one ($f_3 = 3$), and higher in the top two ($f_4 = 4$, $f_5 = 5$). We group types by the direction of the incentive change they face, which drives the predictions below:

$$\begin{aligned} \mathcal{G}^L \text{ (Low evasion)} &:= \mathcal{G}_1 \cup \mathcal{G}_2, \\ \mathcal{G}^M \text{ (Mid evasion)} &:= \mathcal{G}_3, \\ \mathcal{G}^H \text{ (High evasion)} &:= \mathcal{G}_4 \cup \mathcal{G}_5. \end{aligned}$$

Throughout, we refer to the two top brackets ($E > B_3$), those selected by types in $\mathcal{G}^H$, as the *upper tail* of the schedule.

The switch thins the extremes. Because the top rate rises and the bottom rate falls, fewer types fully evade and fewer fully comply, so mass moves from the corners into the interior brackets, where the liability notches at the bracket thresholds concentrate it. At the top the thinning is sharp. Since $pf_5 = 1$, full evasion is an actuarially fair gamble against full compliance, so only types with $U_{cc} \geq 0$, permitted by the relaxation above, remain at $E = Y$.

3.3 Revenue predictions

Expected revenue is the sum of declared-tax revenue $R^d(E) := \tau(Y - E)$ and expected penalty revenue $R^f(E; f) := p\tau f(E)E$:

$$R^{tot}(E; f) = R^d(E) + R^f(E; f) = \tau(Y - E) + p\tau f(E)E. \tag{5}$$

Fix a type $\rho \in \mathcal{G}_k$ and define

$$\Delta^{lin \to prog} E_k(\rho) := E^{prog}(\rho) - E^{lin}(\rho).$$

The revenue change under $lin \to prog$ is

$$\begin{aligned} \Delta^{lin \to prog} R_k^{tot} &:= R^{tot}\big(E^{prog}; f_k\big) - R^{tot}\big(E^{lin}; f_{lin}\big) \\ &= \tau\Big[(pf_k - 1)\Delta^{lin \to prog} E_k + p(f_k - f_{lin})E^{lin}\Big]. \end{aligned} \tag{6}$$

Equation (6) separates the behavioral response in evasion from the mechanical change in the fine multiplier. The mechanical term depends only on how the multipliers compare across regimes, so it delivers a prediction that holds whatever behavior does.

Prediction 1 *Holding behavior fixed, a switch from linear to progressive fines reallocates expected penalty revenue from the lower brackets to the upper brackets. The direction of this reallocation is independent of behavioral responses, while its magnitude depends on where evaded amounts lie.*

The behavioral term operates through $\Delta^{lin \to prog} E_k$, which is shaped by two forces. First, the switch changes the local fine multiplier, and evasion moves in the opposite direction to the fine ($M_f \leq 0$). Second, the notches create threshold attraction, which pulls choices toward the bracket cutoffs.

For high-evasion types $\mathcal{G}^H$ ($k = 4, 5$), the mechanical term and both behavioral forces point toward higher revenue. The fine multiplier exceeds the linear benchmark ($f_k > f_{lin} = 3.25$), so the mechanical term is positive, the higher marginal sanction pushes evasion down, and threshold attraction pulls choices down to a bracket cutoff. Formally, each type in $\mathcal{G}^H$ has its linear optimum above B_3 and satisfies $\Delta^{lin \to prog} E_k \leq 0$, whether the progressive choice is interior, a threshold, or the full-evasion corner (Lemma 1 in Appendix A.3.2).

The signed evasion response, together with the higher multiplier, fixes the direction of two of the three revenue outcomes. Declared-tax revenue, which is proportional to reported income, weakly increases. Total expected revenue also weakly increases because,

under the experimental range ($p = 0.2$, $f_k \leq 5$), both terms in (6) are weakly non-negative. Expected penalty revenue remains ambiguous because a higher fine multiplier raises revenue conditional on evasion, while lower evasion shrinks the penalty base. The fine-revenue response can therefore offset the signed declared-tax response even though total revenue remains weakly positive.

For low- and mid-evasion types ($k = 1, 2, 3$), the mechanical term is negative because $f_k < f_{lin}$, and the behavioral response is ambiguous because the two forces compete: lower local fines tend to raise evasion, whereas threshold attraction can pull choices toward bracket cutoffs. As a result, $\Delta^{lin \to prog} E_k$ is not signed, and the ambiguity carries over to declared-tax and total expected revenue. Expected penalty revenue is ambiguous for an additional reason: even at signed behavior, a larger evaded base and a lower multiplier push in opposite directions. The revenue-sign derivations for both regions are in Appendix A.3.2.

Prediction 2 *Under a switch from linear to progressive fines, high-evasion types weakly increase compliance, declared-tax revenue, and total expected revenue, while expected penalty revenue is ambiguous. For low- and mid-evasion types, all revenue effects are ambiguous. Under the reverse switch, the signed predictions for compliance, declared-tax revenue, and total expected revenue reverse, while expected penalty revenue remains ambiguous.*

3.3.1 Aggregate predictions

Let F_ρ be the distribution of ρ. Aggregate revenue change under the switch is

$$\Delta^{lin \to prog} R^{tot} = \sum_{k=1}^{5} \int_{\mathcal{G}_k} \Delta^{lin \to prog} R_k^{tot}(\rho)\, dF_\rho(\rho), \tag{7}$$

with $\Delta^{lin \to prog} R_k^{tot}$ given by (6). Aggregate gains from progressivity are therefore larger when the type distribution concentrates in the upper tail, where the bracket-level effect is signed, and ambiguous when it concentrates in the three lower brackets, where that effect is not. Under the reverse switch, the bracket-level signs flip and the same composition logic applies. Equation (7) also clarifies that aggregate treatment effects may differ across samples, even with exogenous assignment, because they are composition-dependent through F_ρ. The expanded aggregate decomposition is reported in Appendix A.3.2.

Prediction 3 *The aggregate effect of switching from linear to progressive fines on total revenue is ambiguous ex ante and depends on the distribution of types across brackets.*

We refer to the vector of bracket-level revenue changes $\{\Delta^{lin\to prog} R_k^{tot}\}_{k=1}^5$ as the *incidence* of the schedule change. It is the model's sharpest object: its mechanical component is signed bracket by bracket regardless of behavior, whereas the behavioral component is signed only in the upper tail. Section 5 mirrors this structure, testing the incidence first (Prediction 1) and then asking who responds (Predictions 2 and 3).

4 Empirical strategy and identification

This section takes the model to the data in two layers, following its predictions: an incidence layer testing where revenue moves (Prediction 1), and a heterogeneity layer testing who responds (Predictions 2 and 3). Subsections 4.1 and 4.2 define what each layer estimates, Subsection 4.3 sets out the experimental variation that identifies it, and Subsection 4.4 gives the hierarchy and inference used to test it.

4.1 Where revenue moves: incidence objects

The incidence layer measures where the schedule change moves revenue, using four objects defined here and reported in Section 5: distributional comparisons across arms, a bracket decomposition, an excess-mass statistic at the schedule cutoffs, and a burden-equivalent (level/shape) benchmark. The bracket decomposition and the burden-equivalent benchmark are accounting decompositions at observed behavior, so their direction holds by construction and they describe the mechanical reallocation of Prediction 1. The distributional comparisons and the excess-mass statistic are behavioral, carry their own uncertainty, and rely only on the randomized regime variation.

The distributional comparisons cover declared-tax revenue, fine revenue, total revenue, and evasion across arms. For each subject we average these variables over the 20 rounds. For expected total revenue and evasion, we test equality of the T1 and T2 distributions with an exact session-day randomization-inference Kolmogorov–Smirnov (KS) test based on the largest absolute ECDF gap. Pointwise subject-bootstrap bands on $\Delta(y) = F_{\text{Prog}}(y) - F_{\text{Lin}}(y)$ then localize differences in the displayed distributions. Companion quantile treatment effects are reported in the appendix, though the ECDF evidence is the more reliable because the variables have mass points, which flatten the empirical quantile functions. For evasion we also report equivalence bounds and minimal detectable effects, which state how large an aggregate response the data can rule out.

The bracket decomposition asks whether a bracket's revenue difference comes from a change in its mean R^{tot} (m_r) or in the share of observations it holds (s_r), for regime $r \in \{prog, lin\}$. The two regime changes are $\Delta m = m_{prog} - m_{lin}$ and $\Delta s = s_{prog} - s_{lin}$.

The bracket's total-revenue difference, $s_{prog}m_{prog} - s_{lin}m_{lin}$, can be split into a *Within bracket* term $s_{lin}\Delta m$, the change in the bracket mean, and a *Cell weight* term $m_{lin}\Delta s$, the change in the share of observations the bracket attracts, with residual $\Delta s\Delta m$. The decomposition is the empirical counterpart of the incidence vector $\{\Delta^{lin\to prog}R_k^{tot}\}_{k=1}^5$ and locates the revenue differential along the schedule. The bracket-level evasion change ΔE is reported alongside to track the behavioral margin in (6), separating changes in evasion from changes in fine rates. Because it conditions on realized brackets, we read it as an accounting decomposition of the randomized difference rather than a causal effect.

The excess-mass statistic isolates bunching exactly at a declaration cutoff relative to its neighborhood. Let $P_r(X = x)$ be the share of observations at declaration x, and fix a window width w. At each cutoff $b \in \{0, 25, 50, 75, 100, 125\}$,

$$em_b^r := P_r(X = b) - \frac{1}{2w}\sum_{0<|j|\leq w} P_r(X = b + j)$$

compares the mass at the cutoff against the average mass at the $2w$ surrounding declarations. The regime difference $em_b^{prog} - em_b^{lin}$ nets out focal-value bunching common to both schedules and tests the notch mechanism of Proposition 1(ii), while the bracket ordering of Proposition 1(iii) predicts where displaced mass should reappear.

The burden-equivalent benchmark separates the level of the schedule from its shape. A scale factor c^* rescales the progressive multipliers so that, at observed linear-regime choices, the evasion-weighted average multiplier equals the flat rate. Rescaling removes the mechanical level premium, so the aggregate mechanical effect of the rescaled schedule is zero by construction and its remaining bracket pattern isolates the fixed-behavior reallocation produced by curvature alone. The benchmark is an accounting exercise, not a behavioral counterfactual: no subject faced the rescaled schedule, and c^* is computed from observed behavior.

A burden-equivalent arm cannot be guaranteed by design: the equalizing factor depends on the behavior that arm itself would induce, so any ex-ante calibration would identify the effect of yet another shape-and-level bundle rather than a pure shape effect at equal realized burden, a constraint that actual reforms share. We treat the implemented schedule change, which combines a change in shape with a higher evasion-weighted fine level, as the policy-relevant treatment.

4.2 Who responds: estimands and predicted groups

The heterogeneity layer asks who responds. For grouped estimands, the target is the regime effect within a fixed predicted group, used as an empirical proxy for the model's

low-, mid-, and high-evasion types $\mathcal{G}^L, \mathcal{G}^M, \mathcal{G}^H$ of Subsection 3.2 (Prediction 2). Its causal interpretation requires the group label to be unaffected by assignment. Aggregate estimands compare revenues under the progressive and linear regimes without conditioning on a predicted label and are the direct policy objects. Both designs inform grouped and aggregate versions of these estimands.

Across grouped analyses, each subject i is assigned to a predicted evasion group $\widehat{G}_i \in \{L, M, H\}$, held fixed across rounds to represent the model's stable heterogeneity, with the hat marking a prediction rather than an observed type.[7] The baseline risk set uses incentivized risk investment and self-reported risk willingness, the closest available proxies for the model's heterogeneity channel. A demographic set based on age and a female indicator trades informativeness for greater exogeneity.

We apply the same supervised multiclass procedure to both sets and compare four model families over fixed hyperparameter grids: penalized multinomial logit, feed-forward neural networks, gradient-boosted trees, and probability forests. We rank the resulting specifications lexicographically by repeated out-of-fold macro-F1, then multiclass log loss, then classification error. Each specification returns a subject-level posterior $\widehat{\pi}_i(g)$ over the three groups, and the label is its mode, $\widehat{G}_i := \arg\max_{g \in \{L,M,H\}} \widehat{\pi}_i(g)$. The top-ranked risk-set specification defines the groups used in the main analysis (Section 5). Section 6 examines whether the results carry over to the demographic set, while Appendix A.6 probes sensitivity to hyperparameter selection and uncertainty in individual group classification.

Several qualifications attach to these labels. The baseline risk measures are elicited after the tax task, so a causal interpretation within them rests on the identifying assumption that treatment does not affect those measures or the labels they produce. Being pretreatment, the demographic partition satisfies the condition by construction, at the cost of informativeness (Subsection 6.1). A grouped estimand needs its treatment-by-group cell to be populated, and all are apart from a thin Mid group (Table A1), which is why Mid is estimated least precisely. Assignment is random, so the adjusted specifications address realized covariate differences rather than identification (Tables A2 and A1). With a baseline macro-F1 of 0.371, we read the groups as imperfect proxies rather than observed model types.

Grouping on observed progressive-session evasion is generally endogenous, because it conditions on a treatment-affected outcome. Under the model, however, this concern is lifted: the single-crossing restriction $M_\rho < 0$, implying that higher types evade less, makes optimal evasion monotone in the one-dimensional latent type, so evasion rank tracks type

[7] To prevent leakage from realized outcomes into the labels, subjects whose progressive-round behavior enters the training target receive out-of-fold labels. Appendix A.4.1 gives the cross-fitting and full-sample assignment rules.

within each arm. Combined with session-day randomization, that monotonicity makes ranks comparable across arms, which is the rank-similarity condition (Chernozhukov and Hansen, 2005) rather than the stronger individual rank invariance (Heckman et al., 1997).[8] Aligned with the model and simple to interpret, we report this grouping as a complement to our preferred ML design. When traits outside the single index drive evasion, the same estimates need not recover type-specific effects, but they remain informative descriptive comparisons of heterogeneity.

4.3 Regime variation: T1–T2 and T3/T4

The two designs supply different regime variation: T1–T2 varies the schedule between subjects, T3/T4 within them. They also face different identification threats. In T1–T2, randomization makes the groups comparable in expectation, but the between-subject design does not difference out realized individual characteristics and assignment occurs across only 13 session-day clusters. T3/T4 removes stable individual differences by construction but ties regime exposure to timing, where learning, fatigue, carryover, and sequence-specific shocks might occur. Neither dominates, so we treat them as complementary and use their agreement to assess the evidence.

Both designs use the groups of Subsection 4.2 unchanged, together with common outcome and covariate definitions. Let t index rounds. The round-level outcome is $y_{it} \in \{R^d_{it}, R^f_{it}, R^{tot}_{it}\}$. These outcomes correspond to R^d_k, R^f_k, and R_k in the model. The full-sample mean of R^{tot}_{it} is the sample analogue of the model's expected aggregate revenue.

4.3.1 Between-subject comparison: T1–T2

In the between-subject design, treatment is assigned at the session-day level: subjects in T1 face the progressive schedule in all 20 rounds, while subjects in T2 face the linear schedule. Let $\overline{y}_i$ denote subject i's observed mean outcome over 20 rounds, and let $\overline{y}_i(T1)$ and $\overline{y}_i(T2)$ denote the corresponding mean potential outcomes. For outcome y, session-day randomization identifies the aggregate ITT, $\Delta_y^{T1-T2} := \mathbb{E}[\overline{y}_i(T1) - \overline{y}_i(T2)]$. With perfect compliance, this is also the average effect of the progressive rather than linear schedule in the experimental sample.

We estimate the effect using the subject-level ANCOVA[9]

$$\overline{y}_i = \alpha + \beta\, \mathbf{1}\{i \in T1\} + \boldsymbol{\delta}'\mathbf{Z}_i + \varepsilon_i. \tag{8}$$

[8] Appendix Section A.4.2 reports the details.

[9] A Tobit specification would accommodate the outcome bounds explicitly, but it would add latent-index and distributional assumptions without strengthening identification.

The coefficient $\widehat{\beta}_y$ is the covariate-adjusted T1–T2 difference in the subject-level mean of outcome y, with positive values indicating higher revenue under the progressive schedule. The controls $\mathbf{Z}_i$ are age, study year, and indicators for female, French nationality, economics enrollment, and prior experimental experience. Because the dependent variable is a subject-level average, the specification includes no round controls.

Within predicted group g, the target is $\Delta_{g,y}^{T1-T2} := \mathbb{E}[\overline{y}_i(T1) - \overline{y}_i(T2) \mid \widehat{G}_i = g]$. Session-day randomization identifies this grouped effect under the label condition of Subsection 4.2. Re-estimating the ANCOVA within g gives the adjusted estimator $\widehat{\beta}_{g,y}$ used in the results. The notation $\widehat{\Delta}_{g,y}^{T1-T2}$ denotes the corresponding unadjusted difference in group means. We omit the outcome subscript when it is clear from context.

4.3.2 Within-subject switch comparison: T3/T4

In T3 and T4, each subject faces both regimes in two ten-round blocks with opposite orders: T3 switches from linear to progressive after round 10, while T4 switches from progressive to linear. Let $Post_{it} = \mathbf{1}\{t \geq 11\}$ and let y_{it}^* denote the latent outcome. For each outcome, the pooled random-intercept Tobit is

$$y_{it}^* = \alpha + \beta_1 \mathbf{1}\{i \in T3\} + \beta_2\big(\mathbf{1}\{i \in T3\} \times Post_{it}\big) + \sum_r \delta_r \mathbf{1}\{t = r\} + \boldsymbol{\gamma}'\mathbf{Z}_i + u_i + \varepsilon_{it}. \quad (9)$$

We use the same controls as in the T1–T2 ANCOVA. Full round fixed effects absorb the post indicator and changes common to both sequences, while the random intercept u_i captures persistent subject heterogeneity. We estimate the model separately by predicted group and for the full sample. Because the policy objects are changes in expected revenue, we report fitted differences on the observed revenue scale rather than latent-index coefficients.[10]

Let $m_{T,g}^{\text{pre}}$ and $m_{T,g}^{\text{post}}$ denote the observed-scale means before and after the switch in sequence $T \in \{T3, T4\}$ and group g. The grouped pooled target is

$$\Delta_g^{\text{T3/T4}} := \big(m_{T3,g}^{\text{post}} - m_{T3,g}^{\text{pre}}\big) - \big(m_{T4,g}^{\text{post}} - m_{T4,g}^{\text{pre}}\big),$$

with plug-in estimator $\widehat{\Delta}_g^{\text{T3/T4}}$.[11] Random assignment identifies this difference as the randomized-sequence ITT on the pre–post change and, with perfect compliance, as the effect of assignment to the observed sequences. Because T3 and T4 switch in opposite

[10] Appendix Section A.4.3 details the Tobit implementation and the mapping from latent indices to observed-scale means.

[11] The random-effects Tobit accommodates the censoring of declarations at 0 and 125 and the repeated-measures structure. We also report the model-free difference in subject-level pre-post changes as a specification check.

directions, the pooled estimate combines the two schedule changes.

If round dynamics are common across sequences and the first schedule does not affect behavior under the second, the T3 change measures the move from linear to progressive fines and the sign-reversed T4 change measures the same direction. Halving the pooled estimate then gives the average progressive-minus-linear effect of one switch, $\widehat{\Delta}_g^{\mathrm{T3/T4}}/2$, comparable in scale to T1–T2.

To describe the asymmetry between the two schedule orders, we estimate progressive-minus-linear differences separately within T3 and T4 using RE-Tobit models with a linear round trend. Within each sequence, however, schedule and time move together. The separate estimates therefore cannot distinguish an effect of switch direction from timing, learning, fatigue, or carryover, and we use them only as diagnostics. We denote the fitted observed-scale difference for sequence $T \in \{T3, T4\}$ and group g by $\widehat{\Delta}_g^T$, with outcome subscripts added as needed. Appendix Section A.4.3 gives their formal construction.

4.4 Testing hierarchy and cross-design inference

We fix the testing hierarchy by the model's predictions and set it out explicitly, since the study was not preregistered.

The grouped analysis spans three predicted groups, three revenue outcomes, and two designs. We give primary weight to aggregate total revenue because it is the direct policy object of Prediction 3. We also expect higher precision in the total-revenue estimates than in their components, as evasion moves declared-tax and fine revenue in opposite directions. The High-group total-revenue difference is the theory-prioritized grouped result because Prediction 2 signs total revenue for high-evasion types while leaving fine revenue ambiguous. High declared-tax revenue is supporting evidence for the predicted compliance channel, but it is estimated imprecisely.

We assess aggregate and High total revenue through agreement across the two designs rather than through a grid-wide adjustment that would combine these cells with unsigned fine-revenue outcomes and the Low and Mid groups. The incidence objects of Subsection 4.1 lie outside this grouped testing family: the accounting decompositions are not significance tests, and the behavioral comparisons are not primary tests in the heterogeneity family.

Inference follows the assignment structure of each design. Formal T1–T2 tests use randomization inference at the session-day level, matching the experimental assignment. Under the sharp null this procedure is exact and avoids cluster-asymptotic approximations with few assignment clusters. These tests cover the main grouped and aggregate estimates, the evasion-equivalence exercise, the global KS tests for expected total revenue and evasion, the four T1–T2 interior-threshold excess-mass differences and their joint test,

and the T1–T2 panel of the alternative-partition frontier. We enumerate all $\binom{13}{6} = 1{,}716$ admissible assignments to obtain two-sided randomization p-values and, where reported, confidence intervals by test inversion. The interior-notch joint test is the single directional exception and uses a one-sided maximum-signed statistic.

The subject bootstrap instead provides inference for the pooled and sequence-specific T3/T4 estimates, the hyperparameter-selection and assignment-uncertainty re-estimation checks, the excess-mass endpoints at 0 and 125, all T3/T4 excess-mass results, and all excess-mass confidence intervals. For grouped bootstrap inference under either grouping, we condition on the realized partition: the group labels are held fixed rather than re-estimated in each replication. We report bootstrap standard errors, two-sided plus-one p-values, and percentile confidence intervals. The latent RE-Tobit coefficient tables report model-based standard errors and two-sided normal-approximation p-values.

In Online Appendix Section B.5 we compare the primary T1–T2 randomization-inference intervals with a subject bootstrap that resamples subjects within treatment arm. That bootstrap is a diagnostic rather than the reported T1–T2 inference, because it resamples subjects rather than the session-day assignment. Cross-design agreement therefore rests on the estimates themselves, not on a shared resampling method.

5 Results

5.1 Descriptive diagnostics

Figure 1 reports the raw distribution of evasion E in each regime (equivalently, declarations $X = 125 - E$). This comparison pools all progressive-schedule rounds and all linear-schedule rounds across the four arms, so it is a regime-level description rather than the T1–T2 contrast used for the estimates below.

Choices cluster at the boundaries under both schedules. The progressive regime places slightly less mass at the extremes than the linear regime (49.8% vs. 53.5%), largely because full compliance is less frequent under progressive fines (38.4% vs. 42.0%). Progressivity instead places more mass at the schedule thresholds $X \in \{25, 50, 75, 100\}$. These raw concentrations may combine focal choices, which occur under both schedules, with incentive-driven bunching at the progressive notches.

Treatment-arm support is comparable and aggregate covariate balance is generally strong. The modest T1–T2 differences in age and prior experimental experience motivate covariate adjustment. Declarations are broadly stable over rounds, although a mild downward drift warrants round controls in T3/T4 (Appendix Figure A2). Appendix Subsection A.5.1 reports the full treatment-arm summaries, balance tables, and round-level

plots.

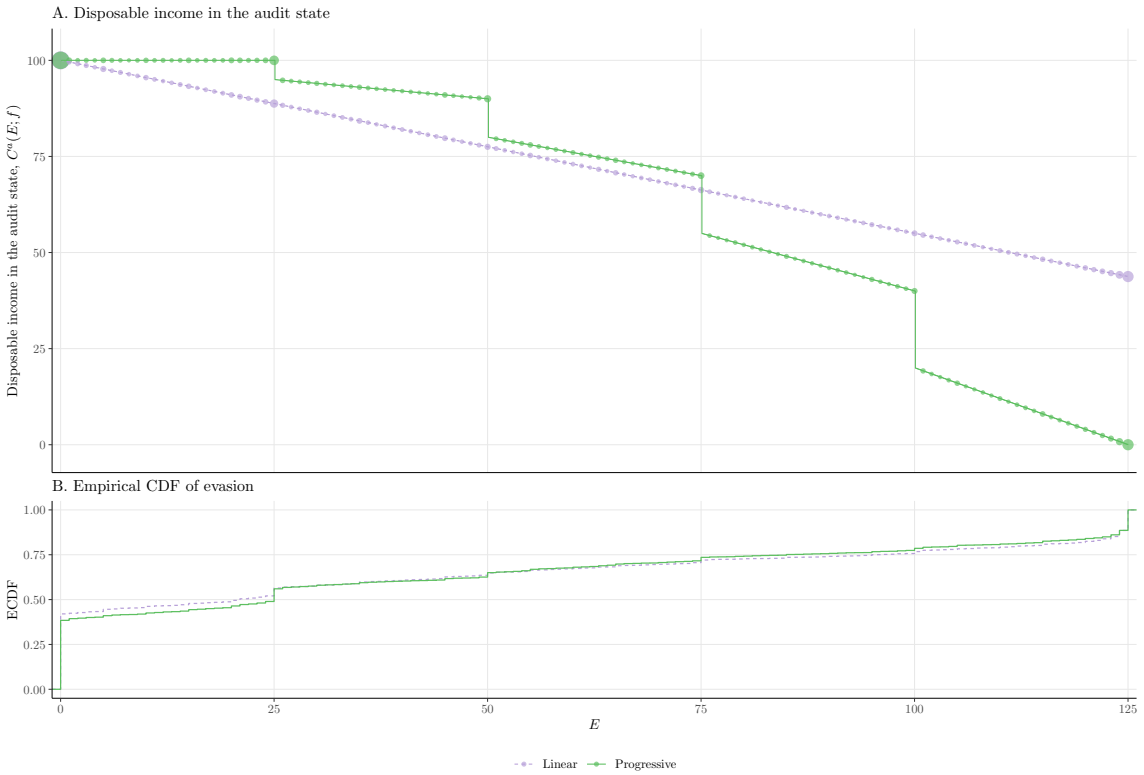

Figure 1: Schedule and evasion distribution by fine regime. Top panel: disposable income in the audit state as a function of evasion E under the linear and progressive fine schedules; point sizes show observed mass at each evasion choice. Bottom panel: empirical CDF of evasion E by regime. Vertical jumps in the top panel mark progressive-bracket thresholds.

5.2 Distributional evidence

We begin at the outcome level, comparing the T1 and T2 distributions of subject-level means directly from random assignment. The ECDF differences described in Subsection 4.1 trace where along the outcome distribution the schedule change has an impact. Each revenue outcome comes in two forms. Expected revenue averages the audited and unaudited outcomes for each choice using the 20% audit probability and is the smooth object predicted by the model. Realized revenue is the amount actually collected after the audit draw and is therefore noisier.

The exact session-day RI KS test rejects equality of the expected-total-revenue distributions, with a largest gap of 0.353 at 22.5 ECU ($p = 0.006$), while it does not reject equality of the evasion distributions (largest gap 0.089, $p = 0.69$). The pointwise subject-bootstrap bands in Figure 2 help locate where the distributions differ. The sample expected-total-revenue difference is nonpositive throughout, so the progressive arm's distribution lies weakly to the right of the linear one at every threshold. Its 95% band

excludes zero from 17.73 to 24.98 ECU. The corresponding largest gap for realized total revenue is 0.171.

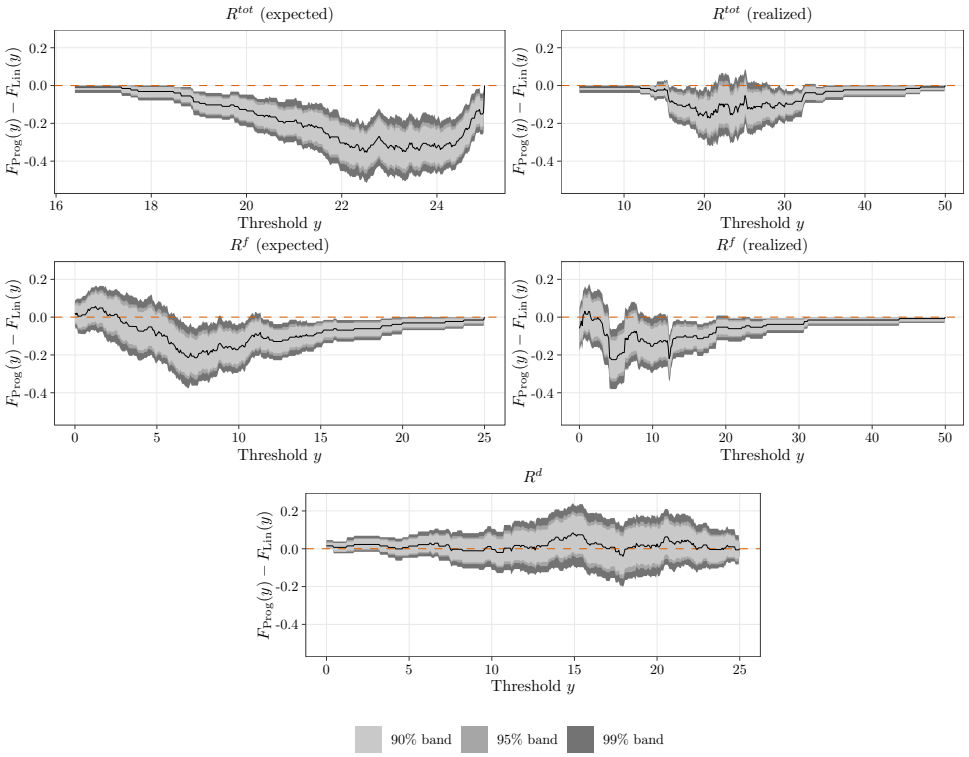

Figure 2: Distribution-function differences in T1–T2. Panels plot $F_{\text{Prog}}(y) - F_{\text{Lin}}(y)$ for subject-level 20-round means, with nested pointwise 90%, 95%, and 99% subject-bootstrap localization bands. Negative values indicate more progressive-arm mass above y, so higher revenue under progressivity appears here as a negative difference. The bands locate that difference along the outcome distribution. Rows report total revenue, fine revenue, and declared-tax revenue. Expected and realized outcomes are shown where relevant.

The difference is concentrated in fine revenue, where the largest gaps are 0.214 for expected and 0.225 for realized fine revenue. In both, the progressive arm holds more mass at higher values. The declared-tax component shows no separated distributional difference. Quantile effects are positive at every reported quantile but smaller near the top of the expected-revenue distribution (Appendix Table A3).

A non-rejection does not establish the absence of an effect, so we bound the evasion response. Table 1 reports equivalence bounds for subject-mean evasion and the full-compliance share, its extensive-margin counterpart.[12] For adjusted subject-mean evasion, the estimated T1–T2 difference is −1.78 ECU, a 4.4% decline relative to the linear-arm mean, with an exact 90% randomization-inference interval of $[-8.22, 4.87]$. The smallest symmetric equivalence margin rejected at the 5% level is therefore 8.22 ECU, or 20.4% of benchmark evasion. The corresponding 80%-power minimum detectable effect is slightly

[12]The full-evasion corner is examined through the excess-mass endpoints in Subsection 5.3.

larger, at 9.12 ECU, or 22.7%. For adjusted full compliance, the analogous equivalence bound is 8.9 percentage points, or 19.9% of the linear-arm mean.

The revenue implication of this bound is tighter than its percentage of evasion suggests. Under the linear schedule, each additional ECU of evasion lowers declared-tax revenue by 0.20 ECU but raises expected fine revenue by 0.13 ECU. Expected total revenue therefore falls by only 0.07 ECU. Using the adjusted 8.22-ECU equivalence bound as a conservative magnitude benchmark gives about 0.58 ECU of expected revenue. This is 2.6% of linear-arm expected total revenue and 39% of the unadjusted T1–T2 expected-revenue difference of 1.467 ECU. Thus, at the linear fine rate, even the largest common evasion response not ruled out corresponds to less than 40% of the sample expected-revenue difference. This is a scale comparison rather than a decomposition share because the revenue effect of evasion varies across the progressive brackets. The fixed-behavior analysis below directly isolates the mechanical fine-level premium.

Table 1: Equivalence bounds and detectable effects for evasion and full compliance in T1–T2.

	Subject-mean evasion		**Full-compliance share**	
	Unadjusted	**Adjusted**	**Unadjusted**	**Adjusted**
Treatment difference: T1–T2				
Estimate	1.956	-1.775	-3.6 pp	-4.2 pp
CI_{99}	[-13.671, 14.154]	[-14.533, 10.064]	[-17.9, 7.9] pp	[-15.8, 6.0] pp
CI_{95}	[-8.458, 10.173]	[-9.882, 6.254]	[-11.2, 4.9] pp	[-10.4, 2.7] pp
CI_{90}	[-5.916, 8.258]	[-8.224, 4.870]	[-9.3, 3.7] pp	[-8.9, 1.5] pp
p-value	0.601	0.622	0.402	0.189
Equivalence bound and design MDE				
T2 mean	40.219	40.219	44.9%	44.9%
δ^*	8.258 (20.5%)	8.224 (20.4%)	9.3 pp (20.6%)	8.9 pp (19.9%)
MDE80	9.823 (24.4%)	9.123 (22.7%)	10.7 pp (23.8%)	8.7 pp (19.3%)
MDE90	11.367 (28.3%)	10.557 (26.2%)	12.4 pp (27.5%)	10.0 pp (22.4%)

Notes. The unit of analysis is the subject-level mean over 20 rounds. T1 has 131 subjects and T2 has 127 subjects. Adjusted columns estimate the T1 coefficient in an ANCOVA on subject means using experience, age, female, study year, the economics-student indicator, and the French-nationality indicator. Confidence intervals, two-sided p-values, and $\delta^* = \max\{|CI_{90,L}|, |CI_{90,U}|\}$ use exact session-day randomization inference by test inversion over all $\binom{13}{6} = 1{,}716$ reassignments. MDE80 and MDE90 are computed from the standard deviation of the effect statistic over all 1,716 session-day assignments, without bootstrap resampling. They equal $2.80 \times SD_{exact}$ and $3.24 \times SD_{exact}$. Parentheses report each bound as a percent of the T2 mean.

5.3 The incidence of the schedule change: edge responses and threshold attraction

Having found a revenue difference without a detectable aggregate evasion difference, we ask where the point estimates place it along the schedule. Table 2 allocates the between-subject difference and the pooled within-subject difference across realized brackets. Because subjects choose their position on the schedule, these components are accounting

quantities rather than causal effects within brackets. The point estimates place the largest changes near the two edges. The middle brackets' estimates are smaller and vary in sign.

Table 2: Revenue-incidence decomposition and threshold attraction in T1–T2 and T3/T4 (Prog. − Lin.).

Panel A. Bracket decomposition for R^{tot}

		T1–T2	T3/T4	T3	T4
Bracket	**Component**				
$B_4 < E \leq B_5$	ΔE	0.82	0.09	1.05	-0.77
	Within-cell	2.70	2.03	2.04	2.02
	Cell weight	0.04	-0.73	0.12	-1.49
	Total effect	2.75	1.01	2.26	-0.10
$B_3 < E \leq B_4$	ΔE	-0.42	0.08	-1.36	1.36
	Within-cell	0.10	0.00	0.43	-0.37
	Cell weight	0.18	-0.07	-0.06	-0.08
	Total effect	0.31	-0.08	0.32	-0.43
$B_2 < E \leq B_3$	ΔE	-0.44	-0.79	-2.18	0.44
	Within-cell	-0.16	0.09	-0.22	0.36
	Cell weight	0.17	0.31	-0.24	0.80
	Total effect	-0.02	0.51	-0.43	1.35
$B_1 < E \leq B_2$	ΔE	-0.93	-0.12	-0.76	0.45
	Within-cell	-0.03	-0.16	-0.14	-0.17
	Cell weight	-0.09	0.43	0.32	0.52
	Total effect	-0.12	0.24	0.16	0.31
$B_0 \leq E \leq B_1$	ΔE	1.16	2.45	1.23	3.53
	Within-cell	-0.29	-0.51	-0.62	-0.41
	Cell weight	-0.39	0.17	-0.21	0.50
	Total effect	-0.67	-0.35	-0.82	0.07

Panel B. Excess-mass differences at thresholds

Threshold	Statistic	T1–T2	T3/T4	T3	T4
0	Excess-mass	2.65	-5.80**	-0.28	-5.53***
	p-value	0.295	0.027	0.899	<0.001
	CI_{95}	[-2.34, 7.82]	[-11.24, -0.71]	[-4.22, 3.76]	[-9.19, -2.36]
25	Excess-mass	0.13	0.07	-0.25	0.33
	p-value	0.783	0.898	0.688	0.387
	CI_{95}	[-0.84, 1.15]	[-1.34, 1.41]	[-1.44, 0.87]	[-0.45, 1.08]
50	Excess-mass	0.31	1.21	-1.31*	2.52***
	p-value	0.439	0.341	0.094	<0.001
	CI_{95}	[-0.53, 1.27]	[-1.24, 3.76]	[-3.12, 0.16]	[0.89, 4.59]
75	Excess-mass	0.97	2.94**	1.86*	1.08*
	p-value	0.368	0.013	0.078	0.067
	CI_{95}	[-0.32, 2.57]	[0.51, 5.58]	[-0.18, 4.27]	[-0.04, 2.24]
100	Excess-mass	1.52	7.22***	2.73*	4.49***
	p-value	0.256	<0.001	0.080	<0.001
	CI_{95}	[-1.88, 4.91]	[3.51, 11.51]	[-0.23, 6.19]	[2.24, 6.93]
125	Excess-mass	-3.62	-7.34**	-3.03	-4.31*
	p-value	0.342	0.043	0.267	0.056
	CI_{95}	[-11.03, 3.51]	[-14.27, -0.29]	[-8.72, 2.48]	[-8.86, 0.08]

Notes. Panel A reports bracket-level accounting decompositions of the Prog. − Lin. R^{tot} differences for T1–T2 and T3/T4 into Within-cell, Cell weight, and Total effect components. These entries hold by construction and are not significance tests, so Panel A reports no inferential stars. The pooled T3/T4 components are observation-count-weighted averages of the T3 and T4 sequence-specific components. The T3 and T4 columns report those components separately. Panel B reports Prog. − Lin. excess-mass differences at declaration thresholds $X \in \{0, 25, 50, 75, 100, 125\}$ using window $w = 2$. At the endpoints, the entries are the corresponding full-evasion and full-compliance mass differences. Its T3/T4 column adds the sign-aligned T3 and T4 pre-post changes. For the T1–T2 interior thresholds $X \in \{25, 50, 75, 100\}$, two-sided p-values use exact session-day randomization inference over all $\binom{13}{6} = 1{,}716$ assignments. The remaining p-values and all displayed confidence intervals use the subject bootstrap. Stars follow the indicated p-values at $p < 0.10$, $p < 0.05$, and $p < 0.01$.

In the top bracket ($B_4 < E \leq B_5$), Panel A of Table 2 allocates a total-revenue difference of 2.75 ECU between subjects and 1.01 ECU within subjects. Between subjects the difference is almost entirely within-bracket, at 2.70 ECU against a cell-weight term of 0.04. Within subjects the within-bracket component is larger still, at 2.03 ECU, but the top bracket attracts a smaller share of observations, with a cell-weight term of −0.73 and a small negative residual. That share decline is the thinning of the full-evasion corner the model anticipates, and it matches the fall in excess mass at $X = 0$ in Panel B, which the between-subject design does not detect. In both designs the estimated top-bracket evasion difference is small and imprecise. This pattern is consistent with a higher fine rate applied to similar evasion. At the bottom bracket ($B_0 \leq E \leq B_1$), the within-bracket

component is negative, at −0.29 ECU between subjects and −0.51 ECU within subjects, while ΔE is positive and sizable, at 1.16 and 2.45 ECU. More evasion together with lower within-bracket revenue is consistent with weaker local fines.

Evidence of threshold attraction differs across the two designs. In pooled T3/T4, excess mass at full evasion and full compliance ($X = 0, 125$) falls and reappears at interior cutoffs, concentrated at $X = 100$ and more modestly at $X = 75$, a pattern consistent with attraction to the progressive schedule's notches.[13] T1–T2 shows this reallocation only weakly, and its exact session-day test does not reject ($p = 0.138$). Much of the pooled pattern comes from T4, where subjects face the progressive schedule first. This may reflect greater attention to the cutoffs early in the experiment or persistence of choices formed under the linear schedule in T3. The within-subject comparison may also detect the movement more precisely because it removes stable differences in focal-choice behavior. Since schedule, timing, and prior exposure move together within each sequence, the experiment cannot distinguish these explanations. We therefore treat threshold attraction as mechanism-consistent evidence rather than an identified response mechanism.

5.4 Separating level from shape: a burden-equivalent benchmark

The upper-tail gain is a fine-rate effect, and the aggregate gain is partly mechanical as well. At observed behavior, the progressive schedule is not merely shaped differently from the flat benchmark but harsher on average, with a higher evasion-weighted fine level, so part of the aggregate T1–T2 revenue gain arises from this level even with behavior held fixed. Table 3 separates this level from the schedule's shape on the expected-revenue scale and compares the resulting mechanical premium with the aggregate observed-revenue ANCOVA estimate. Panel A shows that the progressive schedule raises the evasion-weighted multiplier from 3.25 to 4.24, or about 30%, at observed T2 behavior.

At the point estimate, the fixed-behavior level premium is large relative to the aggregate gain. Holding T2 choices fixed, applying the progressive schedule instead of the linear schedule raises expected revenue by 1.593 ECU. The unadjusted T1–T2 expected-revenue difference is 1.467 ECU, about 8% smaller, so the premium slightly exceeds it: Table 3 puts the ratio at 108.5%. Thus, differences in actual choices and in the composition of the two arms partly offset the fixed-behavior gain, although the between-subject design does not allow us to attribute this residual entirely to behavioral responses. Relative to the covariate-adjusted aggregate T1–T2 R^{tot} difference of 2.055 ECU in Table A5, the premium is 77.5%. This second ratio compares a fixed-behavior expected-revenue quantity with an observed-revenue ANCOVA estimate, so it is a magnitude benchmark

[13] We report results for $w = 2$. The results are analogous for $w = 3$.

rather than a decomposition share. The corresponding 95% subject-bootstrap intervals are 88.1%–136.0% and 48.3%–168.1%, underscoring the uncertainty around both ratios. The all-linear-schedule robustness column confirms the point-estimate pattern.

Table 3: Level, shape, and burden-equivalent revenue incidence.

Panel A. Level and shape diagnostics

	T2 observations		All linear-schedule observations	
Quantity	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Average multiplier, positive evasion	3.182	[2.942, 3.426]	3.241	[3.090, 3.385]
Evasion-weighted multiplier	4.240	[4.065, 4.388]	4.223	[4.120, 4.315]
Burden-equivalent scale, c^*	0.767	[0.741, 0.800]	0.770	[0.753, 0.789]
Mechanical level premium, ECU	1.593	[1.200, 1.998]	1.668	[1.403, 1.929]
Premium / unadjusted difference of 1.467 ECU	108.5%	[88.1%, 136.0%]	113.6%	[91.8%, 150.7%]
Premium / adjusted difference of 2.055 ECU	77.5%	[48.3%, 168.1%]	81.1%	[51.6%, 183.8%]

Panel B. Burden-equivalent incidence (T2 behavior)

	$E = 0$	$(0, 25]$	$(25, 50]$	$(50, 75]$	$(75, 100]$	$(100, 125]$	**Aggregate**
Observation share	44.9%	15.6%	7.6%	5.8%	3.5%	22.6%	100.0%
Rescaled multiplier $c^* f_k$	0.000	0.767	1.533	2.300	3.066	3.833	
Estimate	0.000	-0.272	-0.205	-0.141	-0.024	0.642	0.000
CI_{95}	[0.000, 0.000]	[-0.350, -0.207]	[-0.254, -0.154]	[-0.181, -0.104]	[-0.040, -0.007]	[0.555, 0.721]	[-0.000, 0.000]

Notes. The primary Panel A sample contains T2 observations; the all-linear sample also includes the linear-schedule rounds from T3 and T4. Panel B is a fixed-behavior accounting exercise under the burden-equivalent scale, using observed T2 behavior; behavior under the rescaled schedule is unobserved, and the aggregate effect is zero by construction. The 95% intervals use 2,000 subject-bootstrap replications within treatment arm, with c^* and the Panel A ratios re-estimated in each draw. The table reports no hypothesis tests; 90% and 99% intervals are in the replication package.

The burden-equivalent benchmark of Subsection 4.1 removes this level premium by construction, with c^* estimated at observed T2 choices. Panel B of Table 3 shows that the top bracket gains 0.642 ECU per observation, exactly offset by losses across the four lower positive-evasion brackets. At fixed behavior, the rescaled schedule therefore reallocates expected revenue toward high evasion and has zero aggregate mechanical effect by construction.[14] Because no subject faced this schedule, the calculation does not identify behavioral responses to burden-equivalent progressivity. The fixed-behavior reallocation is the incidence shift in Prediction 1. Its direction holds by construction, so what the data supply is its magnitude.

5.5 Heterogeneity: revenue effects by predicted group and in aggregate

The incidence analysis conditions on the bracket a subject chose, so it says where the difference falls, not who produced it. We now ask who responds, since the predictions differ across the evasion distribution. The grouped results use the baseline covariate-adjusted

[14]The top-bracket gain is about 40% of the 1.593 ECU level premium. This ratio describes relative magnitude, not a share of the aggregate gain.

specification from Section 4 and the external ML-predicted groups from Subsection 4.2. Figure 3 and Table A4 report the grouped observed-scale differences, Figure 4 and Table A5 the aggregate.

5.5.1 Grouped effects

Prediction 2 signs revenue effects only for high evaders. Progressivity should raise their total and declared-tax revenue, while its effect on fine revenue remains ambiguous. In the predicted High group, the T1–T2 total-revenue difference is $\widehat{\beta}_{\text{High},R^{tot}} = 3.675$, or about 17%, with session-day RI $p = 0.037$.[15] The pooled T3/T4 estimate is $\widehat{\Delta}^{\text{T3/T4}}_{\text{High},R^{tot}} = 4.185$ ($p = 0.044$), corresponding under the interpretation of Subsection 4.3.2 to a per-switch average of 2.092, or about 10%.[16] The positive estimate within the predicted High group in both designs is consistent with the upper-tail prediction.

The baseline sequence-specific estimates are positive in both orders: 2.17 in T3 ($p = 0.35$) and 4.85 in T4 ($p = 0.03$), with stronger evidence from T4. Only T4 remains positive throughout the temporal checks.

Declared-tax revenue is the compliance channel, since it is the tax on what subjects declare. It moves in the predicted direction in both designs, with $\widehat{\beta}_{\text{High},R^{d}} = 0.612$ ($p = 0.661$) in T1–T2 and a pooled T3/T4 estimate of 0.460 ($p = 0.649$), though neither estimate is precise. In T1–T2, declared-tax and fine revenue are strongly negatively correlated, with an implied correlation of -0.71. Total revenue is therefore estimated more precisely than either component, as anticipated in Subsection 4.4. Fine revenue is the one component Prediction 2 leaves unsigned for high evaders, and the estimates are indeterminate accordingly.

The estimate for the predicted Low group is positive in both designs but stronger within subjects. Pooled T3/T4 total revenue rises by 3.646 ($p = 0.024$), corresponding under the interpretation of Subsection 4.3.2 to a per-switch average of 1.823 (about 8.4%), compared with 1.830 ($p = 0.055$) in T1–T2. Recall that a Low label means a subject is predicted to evade little, not that their response is small or negative. Prediction 2 leaves the Low group's response unsigned because weaker local fines tend to increase evasion, while threshold attraction can instead pull choices toward bracket cutoffs.

Within subjects, the Low result is concentrated in T4. The total-revenue estimate is 5.061 ($p = 0.011$) in T4 and 1.308 ($p = 0.576$) in T3. In T4, the clearest component is declared-tax revenue, which is estimated to be 2.353 ECU higher in the progressive first block than in the later linear block ($p = 0.006$). The fine-revenue difference is

[15] The percent conversions use the relevant linear-arm or linear-cell mean total revenue as the denominator throughout.

[16] The corresponding difference in subject-level pre-post changes is 4.143, or 2.072 per switch, showing that the magnitude is not driven by the RE-Tobit specification.

1.543 and imprecise ($p = 0.427$). A plausible reading is that threshold attraction is sharper when the progressive schedule comes first, and that choices made under the linear schedule carry over into T3's progressive block. Cutoff movement is indeed strongest in T4 (Subsection 5.3), but sequence, timing, and prior exposure are confounded within each order, so this stays a reading rather than a test. Adding sequence trends lowers the T4 estimate to roughly 2 ECU and reduces its precision, while the pooled Low estimate survives individual and round fixed effects (Subsection 6.2). We therefore treat the Low result as positive but order-sensitive.

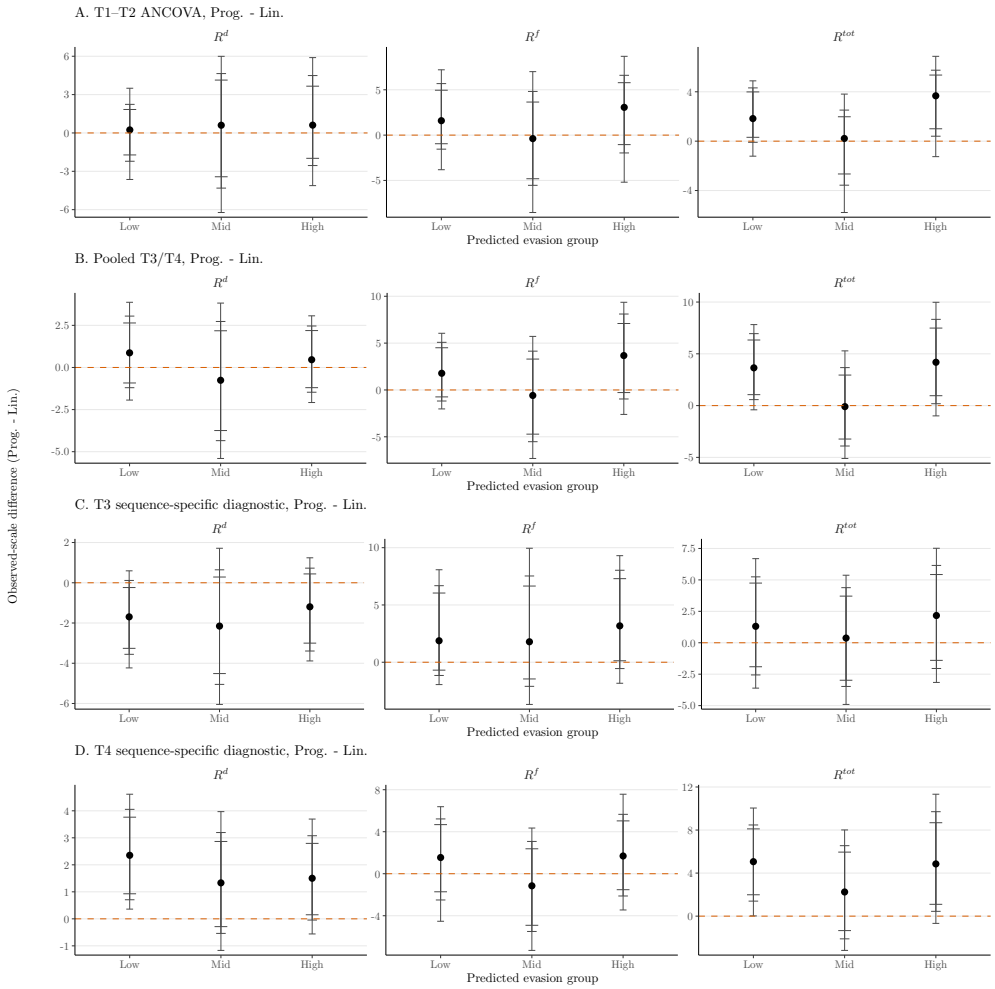

Figure 3: Grouped observed-scale differences with covariates (Prog. − Lin.). Panel A reports T1–T2 ANCOVA differences; Panel B reports the pooled T3/T4 randomized-sequence ITT on the pre-post change. Panels C and D report sequence-specific diagnostics. Within each panel, markers give the progressive-minus-linear difference for the Low, Mid, and High predicted groups across the three revenue outcomes (R^d, R^f, R^{tot}). Error bars show nested 90%, 95%, and 99% session-day RI intervals for Panel A T1-T2 and subject-bootstrap intervals for the T3/T4 and sequence-specific panels.

Estimates for the Mid group are near zero and imprecise in both designs. Prediction 2 leaves the Mid response unsigned as well. Their especially wide T1–T2 interval reflects the

combination of the smallest between-subject group and only thirteen assignment clusters: Mid contains 24 subjects in T1 and 30 in T2 (Table A1).

Online Appendix Section B.1 shows that the grouped point-estimate pattern is stable to changes in the covariate set and reports the corresponding T1–T2, pooled T3/T4, and sequence-specific estimates. Section B.4 reports the latent-scale T3/T4 coefficients underlying the grouped observed-scale results.

5.5.2 Aggregate effect

We estimate the aggregate regime effect on the full sample rather than reconstruct it from the group-specific estimates. With heterogeneous responses, its magnitude reflects the sample's type composition, as in Prediction 3. Figure 4 reports the full design results, with Table A5 reporting the same cells with standard errors and intervals. Total revenue is positive in both design blocks: $\widehat{\beta}_{R^{tot}} = 2.055$, or 9.2%, in the between-subject T1–T2 design[17] ($\text{CI}_{95} = [0.406, 3.825]$, $p = 0.022$), and $\widehat{\Delta}_{R^{tot}}^{\text{T3/T4}} = 2.874$ in the pooled T3/T4 design ($\text{CI}_{95} = [0.844, 5.298]$, $p = 0.005$). Under the interpretation of Subsection 4.3.2, half of the pooled estimate is the per-switch average, 1.437 ECU or about 7%. The positive sign is a feature of this sample's type composition rather than a general implication of that prediction.

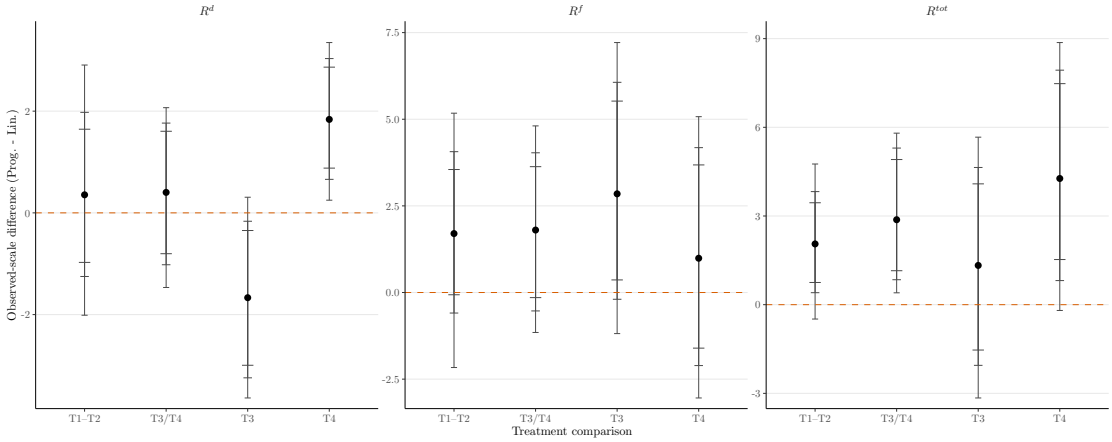

Figure 4: Aggregate observed-scale differences with covariates (Prog. − Lin.) across T1–T2, pooled T3/T4, and the T3 and T4 sequence-specific diagnostics. T1–T2 is an ANCOVA difference and pooled T3/T4 is the randomized-sequence ITT on the pre-post change. For T1–T2, error bars show nested 90%, 95%, and 99% session-day RI confidence intervals. For the pooled T3/T4, T3, and T4 series, error bars show subject-bootstrap intervals.

The internal decomposition is less precise. Aggregate declared-tax revenue is positive but small and imprecise, at 0.355 in T1–T2 ($p = 0.622$) and 0.403 in pooled T3/T4

[17] Online Appendix Section B.5 compares the primary session-day RI with a subject bootstrap across the grouped and aggregate T1–T2 estimates. The bootstrap gives tighter intervals in the headline cells, consistent with within-session-day dependence that subject resampling does not preserve.

($p = 0.582$). Fine-revenue estimates are also positive but imprecise.

The aggregate gain is not explained by overrepresentation of the predicted High group in T1. Their share is similar across arms, at 32.8% in T1 and 35.4% in T2.

6 Robustness and design diagnostics

This section addresses the distinct threats attached to the grouped and within-subject evidence. Subsection 6.1 asks whether the heterogeneity results depend on the measures used to form the groups or on where the classifier draws its boundaries, and Subsection 6.2 asks whether timing rather than the schedule drives the pooled pre–post difference. Pooled T3/T4 entries remain on the randomized-sequence scale of Subsection 4.3.2. Under the interpretation defined there, the estimated effect of one switch is half of the pooled T3/T4 estimate.

6.1 Grouping validity

The first robustness threat concerns the heterogeneity interpretation. The positive High estimate may reflect post-task predictors or a particular classifier boundary. We therefore compare the baseline risk set (macro-F1 = 0.371) with the demographic set introduced in Subsection 4.2, which uses pre-treatment characteristics. We also report the average-observed-evasion benchmark as a deliberately more endogenous but transparent sensitivity check.

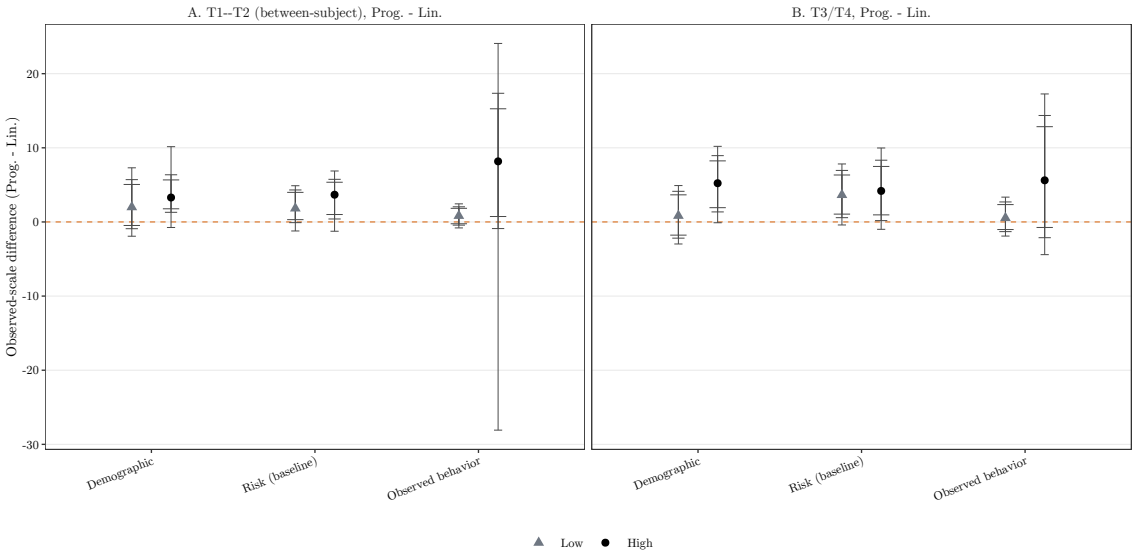

Figure 5: Grouped R^{tot} across the exogeneity–informativeness frontier. Panel A reports T1–T2 between-subject estimates with exact session-day randomization-inference intervals obtained by test inversion. Panel B reports T3/T4 within-subject estimates with subject-bootstrap percentile intervals. Triangles denote the Low group and circles denote the High group. Error bars show nested 90%, 95%, and 99% intervals.

Figure 5, with corresponding estimates in Appendix Table A6, shows that the positive total-revenue estimate for the High group is not tied to a single partition.[18] Using assignment-appropriate inference for each design, High R^{tot} remains positive across the partition frontier, with stronger precision in T1–T2 and more variable precision in pooled T3/T4. Low also remains positive but varies more in magnitude and precision. Because the demographic set uses only age and a female indicator, its agreement with the baseline addresses dependence on post-task risk measures. The demographic classifier nevertheless remains only weakly informative about the model's types (macro-F1 = 0.292). The observed-behavior benchmark produces a larger but much less precise High estimate, at 8.17 ECU between subjects and 5.62 ECU in pooled T3/T4, while its Low estimates remain small. This alignment supports the upper-tail pattern under the model's single-index reading of evasion ranks. Because the groups are based on realized behavior, the comparison is otherwise descriptive.

Appendix A.6.2 reports two additional classification-sensitivity checks. First, we re-estimate the grouped effects under the top 10 support-valid hyperparameter configurations, including the baseline (Figure A3). Second, we perturb group assignments through low-margin exclusion, boundary flips, and stochastic reassignment based on posterior probabilities (Figure A4). These checks show that the positive High estimate is not tied to one model specification or a single classification boundary. Precision remains stable

[18] Online Appendix Section B.2 gives the full four-panel figure and table sets for these partitions, with all outcomes and sequence-specific diagnostics.

across the between-subject specifications but is more variable in pooled T3/T4, where the largest assignment perturbations widen the interval across zero while preserving the sign. The Low estimate is also directionally stable but more sequence- and precision-sensitive. Sequence-specific estimates across alternative hyperparameter configurations preserve the sequence asymmetry documented in the main results.

6.2 Temporal diagnostics and sequence asymmetry

The second robustness question is whether stable individual differences or temporal changes common to the two sequences drive the pooled pre–post difference. Figure 6, with estimates in Appendix Table A7, addresses this concern with two checks. The first re-estimates the pooled specification on rounds 8–13, the three rounds either side of the switch, so that dynamics far from it cannot drive the difference. The second adds individual fixed effects to the baseline's round fixed effects, absorbing stable subject differences that the baseline captures only through a random intercept.

Pooled High, Low, and aggregate total revenue remain positive with individual fixed effects. The local-window estimates are also positive, although the Low estimate is imprecise. The larger estimates near the switch are suggestive of attenuation, but they cannot distinguish a fading response from sequence-specific dynamics or prior exposure. The separate estimates remain asymmetric. Round fixed effects would absorb the switch within a single sequence, so these diagnostics replace them with a linear round trend (Subsection 4.3.2). For Low, the T4 estimate falls from 5.061 to about 2 ECU and loses precision after adding sequence-specific trends, while T3 remains smaller and imprecise. T3 also turns negative for High in the trend specifications, whereas T4 remains positive. These patterns reinforce the order-sensitive interpretation in Subsection 5.5. The pooled estimates are still identified by the sequence randomization, but their interpretation as a common per-switch effect is limited. Appendix A.6 reports the full estimates.

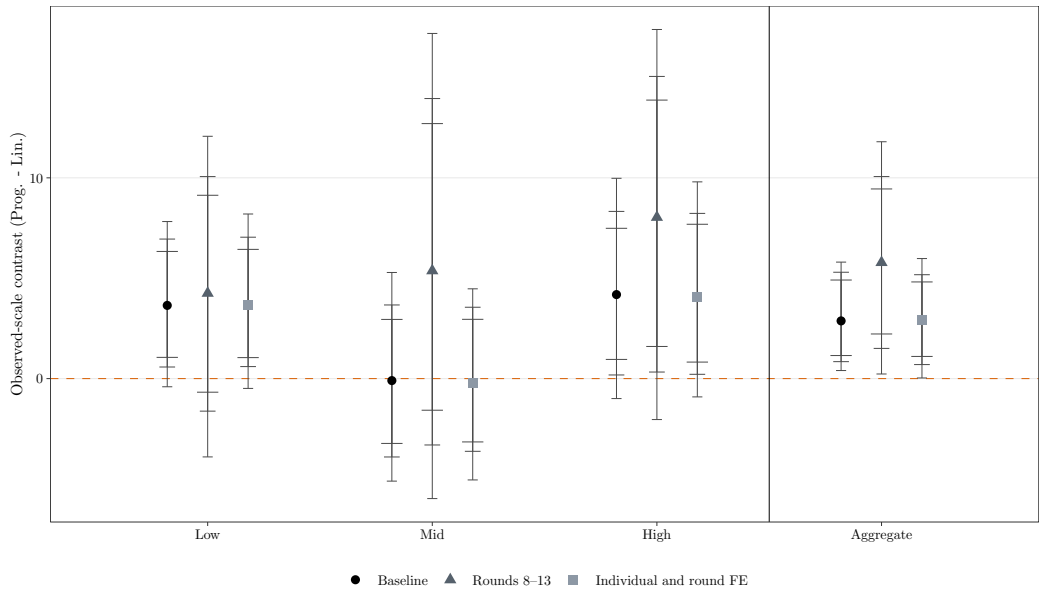

Figure 6: Temporal and individual fixed-effects checks for pooled T3/T4 total-revenue differences, by predicted group and in aggregate. Markers are point estimates; bars show nested 90%, 95%, and 99% subject-bootstrap confidence intervals.

7 Conclusion

Progressive penalties rise more than proportionally with offense severity, yet how their shape affects revenue and behavior remains unclear. This paper asks the incidence question: when sanction intensity is reallocated from small to large offenses, where does expected enforcement revenue move, and who responds? A simple economics-of-crime model shows that the schedule's sharpest implication is this incidence object. Its fixed-behavior mechanical component is signed bracket by bracket, whereas the behavioral and total-revenue predictions are signed only in the upper tail. The experiment estimates the effect of the implemented schedule, and the accounting exercises trace where that difference appears.

The two designs used in the study provide complementary evidence on the implemented progressive schedule. In the between-subject design, facing the progressive schedule throughout raises total revenue by about 9%. We detect no aggregate evasion difference in T1–T2. Any evasion response the equivalence analysis still allows is small enough to move revenue by less than 40% of the unadjusted difference. The pooled randomized pre–post difference is also positive, but the T3–T4 asymmetry limits a common per-switch interpretation because switch direction, timing, and carryover cannot be separated.

The level–shape decomposition explains why the revenue effect should not be attributed to curvature alone. At observed T2 behavior, the progressive schedule applies a substantially higher evasion-weighted fine level, and the corresponding fixed-behavior

premium is large relative to the aggregate point estimate. Once the progressive schedule is rescaled to match the flat schedule's evasion-weighted burden, expected revenue moves from lower brackets to the top and sums to zero by construction. This calculation identifies the incidence of schedule shape at fixed behavior, not the behavioral effect of an unimplemented burden-equivalent schedule.

The grouped and notch results refine this conclusion. In both designs total revenue rises for the predicted High group, where the model signs the effect (Prediction 2), and for Low, where it leaves the sign open. The Low gain comes almost entirely from T4, the sequence where cutoff movement is also strongest, while Mid stays near zero. The pooled within-subject estimates show movement toward interior cutoffs, although the between-subject notch evidence is less conclusive.

Our evidence qualifies a strong behavioral interpretation of marginal deterrence. Although the implemented progressive schedule imposes substantially higher fines on larger amounts of evasion, the between-subject evidence shows higher total revenue without a detectable reduction in aggregate evasion. Progressive fines may therefore serve distributional and fiscal-recovery goals by shifting the enforcement burden toward more severe offenses and helping restore public resources lost to evasion. Policymakers seeking to reduce evasion should not treat a steeper fine schedule as sufficient. The resulting revenue could finance more frequent audits and improved detection technology, both of which raise the probability that evasion is uncovered.

As with any laboratory evidence, external validity requires care. Stakes are bounded, so our magnitudes are evidence on the direction and structure of the response rather than on its size in the field. Whether these patterns extend to actual tax environments depends on punishment salience, liability size, taxpayer composition, and the shape and level of the implemented schedule.

Data availability

The replication package reproduces the analysis in full. It is at `https://github.com/dgdi/evadeProgFine` and archived at `https://doi.org/10.5281/zenodo.22677551`. The data are released under Creative Commons Attribution 4.0 (CC BY 4.0) and the code under the BSD 3-Clause licence.

Ethics

The experiment was conducted online under the auspices of the Laboratory for Experimentation in Social Science and Behavioral Analysis (LESSAC) at the Burgundy School

of Business (Dijon, France). Participation in the present study was voluntary, and participants were recruited from the laboratory's subject pool, which adults join to receive invitations to paid economic experiments. Before starting, participants received information about the procedures and the payment arrangements and consented to take part. The host institution had no ethics committee at the time of the experiment, so the study did not obtain formal ethics approval.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Anthropic's Claude and OpenAI's ChatGPT to draft and revise manuscript prose, to check the internal consistency of the theoretical results, and to write and review parts of the analysis code and the replication package documentation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Abadie, A., Chingos, M. M., and West, M. R. (2018). Endogenous stratification in randomized experiments. *The Review of Economics and Statistics*, 100(4):567–580.

Advani, A., Elming, W., and Shaw, J. (2023). The dynamic effects of tax audits. *Review of Economics and Statistics*, 105(3):545–561.

Allingham, M. G. and Sandmo, A. (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics*, 1(3-4):323–338.

Alm, J. and Malézieux, A. (2021). 40 years of tax evasion games: A meta-analysis. *Experimental Economics*, 24(3):699–750.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Caffera, M., Chávez, C., and Ardente, A. (2021). The deterrence effect of linear versus convex fines: Laboratory evidence. *Journal of Applied Economics*, 24(1):1–26.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2025). Fisher–Schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. *Econometrica*, 93(4):1121–1164.

Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261.

Cowell, F. A. (1985). The economic analysis of tax evasion. *Bulletin of Economic Research*, 37(3):163–193.

Crinò, R., Immordino, G., and Piccolo, S. (2019). Marginal deterrence at work. *Journal of Economic Behavior & Organization*, 166:586–612.

Direction Générale des Finances Publiques (2025). Les Finances Publiques: Rapport d'Activité 2024. Ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique, Paris. Accessed July 3, 2026.

Falkinger, J. and Walther, H. (1991). Reward versus penalties: on a new policy against tax evasion. *Public Finance Quarterly*, 19(1):67–79.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Friedland, N., Maital, S., and Rutenberg, A. (1978). A simulation study of income tax evasion. *Journal of Public Economics*, 10(1):107–116.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2):631–645.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the most out of programme evaluation and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies*, 64(4):487–535.

Heinemann, F. and Kocher, M. G. (2013). Tax compliance under tax regime changes. *International Tax and Public Finance*, 20(2):225–246.

Imai, K. and Li, M. L. (2025). Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. *Journal of Business & Economic Statistics*, 43(1):256–268.

Internal Revenue Service (2025). Internal Revenue Service Data Book, 2024. Publication 55-B, U.S. Department of the Treasury, Washington, DC. Revised May 2025.

Kaila, M. (2026). How Do People React to Income-Based Fines? Evidence from Speeding Tickets Discontinuities. *Review of Economic Studies*. Accepted for publication.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–464.

Kleven, H. J. and Waseem, M. (2013). Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *Quarterly Journal of Economics*, 128(2):669–723.

Koskela, E. (1983). A note on progression, penalty schemes, and tax evasion. *Journal of Public Economics*, 22:127–133.

Landsberger, M. and Meilijson, I. (1982). Incentive generating state dependent penalty system: The case of income tax evasion. *Journal of Public Economics*, 19(3):333–352.

Malézieux, A. (2018). A practical guide to setting up your tax evasion game. *Journal of Tax Administration*, 4(1):107–127.

Mello, S. (2025). Fines and Financial Wellbeing. *Review of Economic Studies*, 92(5):3340–3374.

Milgrom, P. and Shannon, C. (1994). Monotone comparative statics. *Econometrica*, 62(1):157–180.

Ministero dell'Economia e delle Finanze, Dipartimento delle Finanze (2025). Rapporto di verifica 2024: Agenzia delle entrate e Agenzia delle entrate-Riscossione. Ministero dell'Economia e delle Finanze, Rome. Signed June 2025.

Mookherjee, D. and Png, I. P. L. (1994). Marginal Deterrence in Enforcement of Law. *Journal of Political Economy*, 102(5):1039–1066.

Norris, S. and Rose, E. K. (2024). Laffer's day in court: The revenue effects of criminal justice fees and fines. *Journal of Public Economics*, 240:105249.

Pencavel, J. H. (1979). A note on income tax evasion, labor supply, and nonlinear tax schedules. *Journal of Public Economics*, 12(1):115–124.

Polinsky, A. M. and Shavell, S. (2000). The Economic Theory of Public Enforcement of Law. *Journal of Economic Literature*, 38(1):45–76.

Rickard, J. A., Russell, A. M., and Howroyd, T. D. (1982). A Tax Evasion Model with Allowance for Retroactive Penalties. *Economic Record*, 58(4):379–385.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):656–666.

Saez, E. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy*, 2(3):180–212.

Shavell, S. (1992). A note on marginal deterrence. *International Review of Law and Economics*, 12(3):345–355.

Srinivasan, T. N. (1973). Tax evasion: A model. *Journal of Public Economics*, 2(4):339–346.

Stigler, G. J. (1970). The Optimum Enforcement of Laws. *Journal of Political Economy*, 78(3):526–536.

Traxler, C. and Dušek, L. (2025). Fines, Nonpayment, and Revenues: Evidence from Speeding Tickets. *The Journal of Law, Economics, and Organization*, 41(2):381–401.

Traxler, C., Westermaier, F. G., and Wohlschlegel, A. (2018). Bunching on the Autobahn? Speeding Responses to a 'Notched' Penalty Scheme. *Journal of Public Economics*, 157:78–94.

Yitzhaki, S. (1974). A Note on "Income tax evasion: A theoretical analysis". *Journal of Public Economics*, 3(2):201–202.

Appendix

Appendix section numbers mirror the corresponding main-text sections, and there is no Appendix A.1.

A.2 Experimental design

A.2.1 Instructions

Instructions translated from the French original. The base text is the progressive-schedule version. Square brackets give the corresponding wording or values for the linear arm (T2) and for the two switching sequences, Linear-Progressive (T3) and Progressive-Linear (T4).

GENERAL

This is an experiment in which you will earn "real" money. The total amount you will earn depends on your decisions.

The experiment consists of 2 independent parts. You will receive the proposition for part 2 after the end of part 1. At the end of the experiment, your earnings will be the sum of what you earn in part 1 and part 2. All your decisions will be anonymous.

Thank you for your participation.

PART 1

This part consists of 20 periods with identical course of action. The currency used during the experiment is ECU (Experimental Currency Units). Your earning for this part will be converted into euro at the exchange rate of

$$25 \text{ ECU} = 1 \text{ euro}$$

At the end of part 1, 2 periods will be randomly drawn and your decision for those two periods will be used for your payments of part 1.

DECISIONS

This experiment is a **tax simulation**.

At the beginning of each period you will receive an initial income equal to 125 ECU. After receiving your income, you will be asked to declare it for tax purposes. Only declared income will be taxed. The tax rate is equal to 20% and will be the same in all the periods.

CALCULATION OF PROFITS

Your profit in each round will be equal to:

Your income − (the declared income × 0,2)

The total amount of tax collected during all the 20 periods of the experiment will be invested in a research fund to run more experiments. You will therefore not be getting anything back from those taxes.

Your *declared income* does not have to be equal to your *initial income*. It can be equal or lower.

However, there is a 20% chance that you will be audited. The probability of an audit is fixed and does not change during all the 20 periods. It means that if there are 40 participants, 8 participants will be audited in each period.

If an audit occurs, then the computer will examine your declared income comparing it with your initial income. If there is a difference between the declared income and the initial income, you will pay a fine on the unpaid tax.

So, there are three possibilities:

- *No audit.*
  A message will inform you that you were not audited in that period.
- *An audit occurs and there is no difference between the declared income and the initial income.*
  A message will inform you that the audit did not entail any fine.
- *An audit occurs and there is a difference between the declared income and the initial income.*
  A message will inform you that the income declared is not equal to the initial income of that period. A fine will be applied to your earnings.

The fine is determined in the following way:

Fine = q× unpaid tax

q is the fine rate and can vary between 0 and 5 times the unpaid tax. [Linear: *It is equal to 3.25 and is fixed in each period.* Progressive-Linear: *Between periods 1 to 10, the fine rate varies between 0 and 5 times according to the income declared. Between periods 11 to 20, the fine is equal to 3.25.* Linear-Progressive: *Between periods 1 to 10, the fine rate is equal to q=3.25. Between periods 11 to 20, the fine rate varies between 0 and 5 times, according to the income declared.*]

Declared Income	**Level of *q***	**Fine**
125	q=0	0
124 - 100	q=1 [=*3.25*]	1 [*3.25*] ×unpaid tax
99 - 75	q=2 [=*3.25*]	2 [*3.25*] ×unpaid tax
74 - 50	q=3 [=*3.25*]	3 [*3.25*] ×unpaid tax
49 - 25	q=4 [=*3.25*]	4 [*3.25*] ×unpaid tax
24 - 0	q=5 [=*3.25*]	5 [*3.25*] ×unpaid tax

In this case your profit will be:

Your income − (the declared income × 0,2) − (q × unpaid tax)

Please complete the following questionnaire to verify the comprehension of the instructions of part 1 :

1. The income that I declare will be taxed at the rate of 20%.
2. In each period I will receive an income equal to 125 ECU.
3. The audit probability is fixed at 20%.
4. The fine can vary between 0 and 5 times the unpaid tax. [Linear: *The fine is equal to 3.25 and is fixed in each period.* Progressive-Linear: *Between periods 1 to 10, the fine varies between 0 and 5 times the unpaid tax. Between periods 11 to 20, the fine is equal to 3.25 times the unpaid tax.* Linear-Progressive: *Between periods 1 to 10, the fine is equal to 3.25 times the unpaid tax. Between periods 11 to 20, the fine varies between 0 and 5 times the unpaid tax.*]

PART 2

You have an endowment of 10 euros. You need to decide how much you invest in a risky option. You can choose how much to invest (in whole number) between 0 and 10, i.e. 0, 1, 2,..., 9, or 10.

The risky option is a coin toss game. The risky option will bring you 3 times the amount invested if heads is drawn and 0 if tails is drawn.

Your final gain is equal to the number of euros kept + the gain of your investment.

Examples

1. You invest 5 euros. Your final win will be 20 euros if heads are drawn (5 euros kept + 5×3 euros of your 5 euros investment in the risky option) and 5 euros if tails is drawn (5 euros kept + 5×0 euro of your investment in the risky option).

2. You invest 0 euros, your gain will be 10 euros regardless of the outcome of the coin toss game.

3. You invest 10 euros. Your final win will be 30 euros if heads are drawn (0 euros kept + 10×3 euros of your investment in the risky option) and 0 euros if tails is drawn (0 euros kept + 10×0 euros of your investment in the risky option).

To avoid calculations, the table below directly reports the final gain according to the amount invested and the result of the draw.

Amount invested	Gain	
	if heads	if tails
0	10	10
1	12	9
2	14	8
3	16	7
4	18	6
5	20	5
6	22	4
7	24	3
8	26	2
9	28	1
10	30	0

Questions

1. Age
2. Gender
3. Year of study
4. Nationality
5. Field of study
6. Is this your first experience in a lab experiment?
7. From 1 (completely unwilling to take risks) to 10 (completely willing to take risks), are you generally a person who is completely willing to take risks, or completely unwilling?
8. How much do you spend each month (on average) in leisure/recreational activities? (Examples of expenses in leisure/recreational activities are: subscriptions such as Netflix and Spotify, money spent for your hobbies and in fitness/sport, money spent at coffee shops, at restaurants etc.)
9. How much do you save each month (on average)? (Please consider all the possible forms of savings: cash, savings account, deposits, etc.)

A.3 Theoretical framework

A.3.1 Proof of Proposition 1

The proof first sets up four intermediate results and then verifies the proposition. We first compute the second derivative of expected utility on a flat segment. Under the maintained restrictions it is negative, so each bracket admits at most one interior optimum. We then state the bracketing conditions under which the type cutoffs, the values of ρ at which a corner or a bracket threshold is exactly optimal, exist and are unique. Next, we consider the conditions under which a bracket threshold is locally optimal, the mechanism behind bunching. Finally, we assemble the finite set of candidate optima under the progressive schedule. The proposition then follows by comparing marginal returns at the bracket boundaries.

For any flat segment, differentiating M with respect to E gives

$$M_E(E;\rho,\bar{f}) = \tau^2\Big[(1-p)U_{cc}\big(C^{na}(E);\rho\big) + p(1-\bar{f})^2 U_{cc}\big(C^{a}(E;\bar{f});\rho\big)\Big], \tag{10}$$

the second derivative of expected utility with respect to evasion on the segment.

The cutoffs we use exist when the following sign conditions hold on a type interval $[\rho_L, \rho_H]$:

$$\begin{aligned} &M(0;\rho_L, f_{lin})\, M(0;\rho_H, f_{lin}) \le 0, \qquad M(Y;\rho_L, f_{lin})\, M(Y;\rho_H, f_{lin}) \le 0, \\ &M(0;\rho_L, f_1)\, M(0;\rho_H, f_1) \le 0, \qquad M(Y;\rho_L, f_5)\, M(Y;\rho_H, f_5) \le 0, \\ &M(B_k;\rho_L, f_k)\, M(B_k;\rho_H, f_k) \le 0, \qquad k = 1, \ldots, 4. \end{aligned} \tag{11}$$

Each condition in (11) implies a sign change. By continuity the corresponding cutoff exists, and since $M_\rho < 0$, it is unique.

Fix a threshold B_k with $k \in \{1, \ldots, 4\}$ and suppose that:

$$M(B_k^-;\rho, f_k) \ge 0, \qquad M(B_k^+;\rho, f_{k+1}) \le 0.$$

Because utility is strictly concave on each adjacent flat segment, the left inequality implies that it is weakly increasing up to B_k from the left, whereas the right inequality implies that it is weakly decreasing immediately to the right. Hence B_k is locally optimal. If the type distribution has positive density on a neighborhood of types for which these inequalities hold, the local threshold selection appears as bunching at B_k.

For the progressive schedule, define for each bracket k the within-bracket candidate E_k^I as the unique solution to $M(E;\rho, f_k) = 0$ when it lies in (B_{k-1}, B_k). The finite candidate set is

$$\mathcal{C}(\rho) = \{B_0, \ldots, B_5\} \cup \{E_k^I : \ E_k^I \text{ exists for } k = 1, \ldots, 5\}.$$

The progressive optimizer correspondence is

$$\mathcal{E}^{prog}(\rho) := \arg\max_{E \in \mathcal{C}(\rho)} \mathcal{U}(E;\rho, f_{prog}).$$

We write $E^{prog}(\rho)$ for an arbitrary element of $\mathcal{E}^{prog}(\rho)$. A type can be indifferent between several candidates in $\mathcal{C}(\rho)$, in which case the optimizer correspondence is not single-valued and some selection must be made. The weak sign predictions in the main text hold for every element of the tied set, so no tie-breaking assumption is needed and none is imposed.

We now use these results to prove the three parts of the proposition.

(i) Linear schedule. With $f(E) \equiv f_{lin}$, strict concavity follows from (10) evaluated at $\bar{f} = f_{lin}$, so $M_E < 0$ on $[0, Y]$. Hence there is a unique maximizer in $[0, Y]$. On the interior, the implicit-function theorem gives the comparative statics in the type and, for

flat-rate comparisons, in the fine level:

$$E_\rho = -\frac{M_\rho}{M_E} < 0, \qquad E_f = -\frac{M_f}{M_E} \leq 0.$$

The corner cutoffs ρ^0_{lin} and ρ^Y_{lin} are defined by the boundary equalities $M(0;\rho^0_{lin}, f_{lin}) = 0$ and $M(Y;\rho^Y_{lin}, f_{lin}) = 0$. Since $M_E < 0$, we have $M(0;\rho, f_{lin}) \geq M(Y;\rho, f_{lin})$ for every ρ. Evaluating this inequality at $\rho = \rho^Y_{lin}$ gives $M(0;\rho^Y_{lin}, f_{lin}) \geq 0$, so the zero of $M(0;\rho, f_{lin})$, which is decreasing in ρ, lies weakly above ρ^Y_{lin}, giving $\rho^Y_{lin} \leq \rho^0_{lin}$. Because $M_\rho < 0$, the sign of $M(0;\rho, f_{lin})$ and $M(Y;\rho, f_{lin})$ partitions the full-evasion, interior, and full-compliance cases exactly as stated in Proposition 1(i).

(ii) Progressive global selection and local threshold attraction. Within each bracket, (10) with $\bar{f} = f_k$ implies strict concavity, hence at most one interior candidate E^I_k. Since $[0, Y]$ is compact and the schedule has finitely many thresholds, the global solution is obtained by finite comparison over $\mathcal{C}(\rho)$. A threshold B_k that is locally optimal by the conditions above can therefore also be globally selected, and by continuity they hold on a neighborhood of such types, so bunching follows once the type distribution places positive density there.

(iii) Type-space mapping under the progressive schedule. Define the corner cutoffs

$$M\big(0;\rho^0_{prog}, f_1\big) = 0, \qquad M\big(Y;\rho^Y_{prog}, f_5\big) = 0,$$

and, for $k = 1, \ldots, 4$, the interior-threshold cutoffs

$$M\big(B_k;\rho^{B_k}_{prog}, f_k\big) = 0. \tag{12}$$

The sign-bracketing conditions in (11) guarantee that these cutoffs exist, and $M_\rho < 0$ implies that each is unique. To order the cutoffs, define

$$\begin{aligned} H_0(\rho) &:= M(0;\rho, f_1), \\ H_k(\rho) &:= M(B_k;\rho, f_k), \quad k = 1, \ldots, 4, \\ H_5(\rho) &:= M(Y;\rho, f_5). \end{aligned}$$

Because $M_E < 0$ and $M_f \leq 0$ on flat segments, and because both thresholds and fine levels increase across brackets, we have

$$H_0(\rho) \geq H_1(\rho) \geq H_2(\rho) \geq H_3(\rho) \geq H_4(\rho) \geq H_5(\rho) \quad \text{for every } \rho.$$

Since each H_j is decreasing in ρ, their unique zeros must satisfy

$$\rho^0_{prog} \geq \rho^{B_1}_{prog} \geq \rho^{B_2}_{prog} \geq \rho^{B_3}_{prog} \geq \rho^{B_4}_{prog} \geq \rho^{Y}_{prog}.$$

Let $\mathcal{G}_k$ collect the types whose progressive choice falls in bracket k,

$$\begin{aligned}
\mathcal{G}_1 &:= \{\rho : E^{prog}(\rho) \leq B_1\}, \\
\mathcal{G}_k &:= \{\rho : B_{k-1} < E^{prog}(\rho) \leq B_k\}, \quad k = 2, \ldots, 5.
\end{aligned}$$

These sets partition the type space, since every choice lies in $[0, Y]$. The cutoffs bound which of them a type can belong to. Now fix ρ. If $\rho > \rho^{B_1}_{prog}$, then $M(B_1; \rho, f_1) < 0$, so the optimum cannot lie above B_1 and $E^{prog}(\rho) \in [0, B_1]$. If, in addition, $\rho \geq \rho^0_{prog}$, then $M(0; \rho, f_1) \leq 0$ and the corner $E = 0$ is selected. For $k = 2, 3, 4$, if $\rho^{B_k}_{prog} < \rho \leq \rho^{B_{k-1}}_{prog}$, then

$$M(B_{k-1}; \rho, f_{k-1}) \geq 0, \qquad M(B_k; \rho, f_k) < 0.$$

The second inequality bounds the choice from above. Evading more is already unattractive at B_k. Above B_k the multiplier is higher and every crossing lowers utility further, so the optimum does not exceed B_k.

The first inequality shapes the choice below. Since $H_j(\rho) \geq H_{k-1}(\rho) \geq 0$ for every $j \leq k-1$, expected utility increases across each lower bracket. No interior point of a lower bracket is selected, so the only candidates below B_{k-1} are the thresholds. Marginal signs cannot rank those thresholds against the bracket-k candidate. Crossing a threshold applies the higher multiplier to the whole evaded amount, not only to the extra evasion. The comparison is therefore between utility levels on either side of a notch. A type whose gain within bracket k does not cover that loss stops at the lower threshold, the local-optimality case of part (ii).

We conclude that $E^{prog}(\rho) \leq B_k$. The choice falls inside bracket k or at a threshold at or below B_{k-1}. The same argument applies at $k = 5$ when $\rho \leq \rho^{B_4}_{prog}$. Because the cutoffs are ordered, these bounds fall as ρ rises, the ordering stated in Proposition 1(iii).

The compliance corner above has no counterpart at full evasion. Reaching $E = 0$ crosses no threshold. Reaching $E = Y$ crosses all of them. ■

A.3.2 Revenue-sign derivations

This subsection derives the predicted revenue effects of the regime switch by bracket. For high-evasion types it signs declared-tax and total revenue while leaving expected penalty revenue ambiguous, and all revenue effects are ambiguous for low- and mid-evasion types. It closes with the aggregate prediction decomposition. For the upper tail,

the derivations rest on the inequality $E^{prog}(\rho) \leq E^{lin}(\rho)$. Signing it is not immediate: the progressive optimizer can be an interior point, a bracket threshold, or a corner, and the two optima can fall in different brackets, which rules out comparing first-order conditions on a common flat segment. The lemma below handles all these cases.

Lemma 1 *Assume $U_c > 0$ and, on each flat segment, $M_E < 0$ and $M_f \leq 0$. For every type ρ, if $E^{prog}(\rho) > B_3$, then $E^{lin}(\rho) > B_3$ and $E^{prog}(\rho) \leq E^{lin}(\rho)$. In particular, $\Delta^{lin \to prog} E(\rho) \leq 0$ for every $\rho \in \mathcal{G}^H$.*

Proof. Fix a type ρ, write $\mathcal{U}^P(E) := \mathcal{U}(E; \rho, f_{prog})$, and let $m := \max\{E^{lin}(\rho), B_3\}$ denote the type-specific cutoff. The proof shows that, above m, evading more always lowers expected utility, so each type's progressive choice is capped at that type's m. For the high-evasion types the lemma concerns, m will turn out to equal $E^{lin}(\rho)$, so their evasion under the progressive schedule cannot exceed their evasion under the linear one. Each piece of the cutoff removes one reason utility could still rise: above $E^{lin}(\rho)$ the type did not want to evade more even at the flat rate f_{lin}, and above B_3 the progressive multipliers exceed the flat rate, making evasion still less attractive.

Formally, two facts hold above m (if $m = Y$ there is nothing to prove). First, the marginal return to evasion is negative. If $E^{lin}(\rho) < Y$, optimality requires $M(E^{lin}; \rho, f_{lin}) \leq 0$, and $M_E < 0$ gives $M(E; \rho, f_{lin}) < 0$ for all $E > E^{lin}(\rho)$. Any $E \in (m, Y]$ also satisfies $E > B_3$, so $f_{prog}(E) \in \{f_4, f_5\}$ exceeds f_{lin}, and $M_f \leq 0$ yields

$$M\big(E; \rho, f_{prog}(E)\big) \;\leq\; M(E; \rho, f_{lin}) \;<\; 0,$$

so $\mathcal{U}^P$ strictly decreases along every flat segment inside $(m, Y]$. Second, utility drops at every threshold. As E crosses B_k from the left, the higher multiplier applies to the entire evaded amount: audit-state consumption falls by $\tau(f_{k+1} - f_k)B_k > 0$ while non-audit consumption is continuous, and $U_c > 0$ with $p > 0$ gives $\lim_{E \downarrow B_k} \mathcal{U}^P(E) < \mathcal{U}^P(B_k)$. Together with $\lim_{E \downarrow m} \mathcal{U}^P(E) \leq \mathcal{U}^P(m)$, which holds with equality when m is interior to a bracket and strictly when m is a threshold, these facts give $\mathcal{U}^P(E) < \mathcal{U}^P(m)$ for every $E > m$. No maximizer of $\mathcal{U}^P$ lies strictly above m.

The lemma follows. For $\rho \in \mathcal{G}^H$, $E^{prog}(\rho) > B_3$ and, by the claim just shown, $E^{prog}(\rho) \leq m$. If $E^{lin}(\rho) \leq B_3$, then $m = B_3$, contradicting $E^{prog}(\rho) > B_3$. Hence $E^{lin}(\rho) > B_3$, so $m = E^{lin}(\rho)$ and $E^{prog}(\rho) \leq E^{lin}(\rho)$: evasion under the progressive schedule cannot exceed evasion under the linear one, whether the progressive choice is interior, a threshold, or the full-evasion corner. ■

The lemma also shows that types in $\mathcal{G}^H$ are upper-tail types under both regimes: their linear optimum already lies above B_3. The main-text comparison for high-evasion types is therefore one within the upper tail.

We can now sign the revenue changes. For high-evasion types $\mathcal{G}^H$ ($k = 4, 5$), $f_k > f_{lin}$ and Lemma 1 gives $\Delta^{lin\to prog} E_k \leq 0$. Declared-tax revenue then changes by

$$\Delta^{lin\to prog} R_k^d = \tau\big[(Y - E^{prog}) - (Y - E^{lin})\big] = -\tau \Delta^{lin\to prog} E_k \geq 0,$$

and expected penalty revenue changes by

$$\Delta^{lin\to prog} R_k^f = p\tau\big(f_k E^{prog} - f_{lin} E^{lin}\big) = p\tau\Big[f_k \Delta^{lin\to prog} E_k + (f_k - f_{lin}) E^{lin}\Big].$$

The second term is strictly positive when $E^{lin} > 0$, whereas the first term is weakly non-positive because $\Delta^{lin\to prog} E_k \leq 0$, so the sign of $\Delta^{lin\to prog} R_k^f$ is ambiguous.

Total expected revenue satisfies (6),

$$\Delta^{lin\to prog} R_k^{tot} = \tau\Big[(pf_k - 1)\Delta^{lin\to prog} E_k + p(f_k - f_{lin}) E^{lin}\Big].$$

Under the experimental range ($p = 0.2$, $f_k \leq 5$), $pf_k - 1 \leq 0$. The first term is therefore weakly non-negative given $\Delta^{lin\to prog} E_k \leq 0$, and the second term is strictly positive when $E^{lin} > 0$. Hence $\Delta^{lin\to prog} R_k^{tot}$ is weakly positive in the upper brackets. In the knife-edge case $k = 5$ ($pf_5 - 1 = 0$), only the strictly positive second term remains.

For low- and mid-evasion types $\mathcal{G}^L \cup \mathcal{G}^M$ ($k = 1, 2, 3$), the within-bracket fine level under progressivity is weakly below the linear benchmark, whereas local threshold attraction can generate bunching at the bracket cutoffs. Because the model does not rank these forces globally, $\Delta^{lin\to prog} E_k$ is not signed, and neither are $\Delta^{lin\to prog} R_k^d$, $\Delta^{lin\to prog} R_k^f$, and $\Delta^{lin\to prog} R_k^{tot}$. For $\Delta^{lin\to prog} R_k^f$ the ambiguity is double: in the decomposition above, $(f_k - f_{lin}) E^{lin} \leq 0$, strictly when $E^{lin} > 0$, while $f_k \Delta^{lin\to prog} E_k$ carries the sign of the behavioral response, so the two terms conflict whenever evasion rises. For the reverse switch,

$$\Delta^{prog\to lin} E_k := E^{lin} - E^{prog} = -\Delta^{lin\to prog} E_k, \qquad \Delta^{prog\to lin} R_k^{tot} := -\Delta^{lin\to prog} R_k^{tot}.$$

The signs for compliance, declared-tax revenue, and total revenue therefore reverse, while expected penalty revenue remains ambiguous.

Substituting (6) into (7) gives

$$\Delta^{lin\to prog} R^{tot} = \tau \sum_{k=1}^{5} \left[(pf_k - 1) \int_{\mathcal{G}_k} \Delta^{lin\to prog} E_k \, dF_\rho + p(f_k - f_{lin}) \int_{\mathcal{G}_k} E^{lin} \, dF_\rho \right],$$

showing that aggregate gains depend both on behavioral responses within brackets and on the distribution of types across brackets.

A.4 Empirical strategy and identification

A.4.1 Details on predictive grouping

For the training target, the baseline specification uses progressive observations from T1 and the first ten rounds of T4, avoiding the late-switch progressive block where adaptation, learning, and session dynamics are most likely to contaminate it. Final group labels never use a subject's own realized behavior. Subjects whose behavior enters the target receive out-of-fold labels from 10-fold cross-fitting: each is labeled by a classifier trained only on the other folds. Subjects outside the training target receive labels from the corresponding full-sample fit. The same external-feature predictive rule therefore defines the groups in every arm.

We abbreviate the four model families of Subsection 4.2 as logit, nnet, forest and boost. For the two main feature sets, the weighted classification grids contain 505 configurations for logit, 512 for nnet, 600 for forest, and 720 for boost. The grids vary model complexity and regularization while holding numerical and convergence controls fixed. Each model grid is evaluated using repeated stratified out-of-fold prediction (10 folds $\times$ 10 repeats). The lexicographic ranking prioritizes minority-class performance before overall-fit tie-breakers.

A.4.2 Average-observed-evasion benchmark

Let $Prog_{it}$ indicate that subject i faces the progressive schedule in round t. The average progressive-round evasion benchmark is

$$\overline{E}_i^{\text{prog}} := \frac{\sum_t Prog_{it}\, E_{it}}{\sum_t Prog_{it}}.$$

For T3/T4, groups use bracket-consistent evasion cutoffs (< 50, $[50, 75]$, > 75). For T1–T2, progressive-arm group shares are transported to the linear arm by rank-order assignment, so quantile shares are aligned across arms.

A structural reading of this percentile-share transport requires optimal evasion to be monotone in a one-dimensional latent index under both schedules. It also requires single-valued optimal choices, or a common deterministic tie-breaking rule. In the model of Section 3, these are natural regularity conditions. Because the comparison is between subjects, randomization equates the type distribution across arms, and no individual rank has to be preserved across schedules. Under these assumptions, aligned quantile shares identify comparable latent-index groups, so the unadjusted grouped difference $\widehat{\Delta}_g^{T1-T2}$ of Subsection 4.3.1 converges to latent-quantile-specific treatment effects. Without these as-

sumptions, the same estimates remain informative descriptive heterogeneity comparisons, but they need not recover latent-type-specific causal parameters.

A.4.3 Observed-scale RE-Tobit differences

This subsection gives the sequence-specific specification and the mapping from latent indices to observed-scale means that produces the differences of Subsection 4.3.2. The sequence-specific models are estimated separately for T3 and T4, replacing the pooled sequence-by-post terms of (9) with $Prog_{it}$, a linear round trend, group indicators, and interactions between $Prog_{it}$ and the High and Low groups. Mid is the omitted group. Sequence-specific aggregate models omit the group terms.

To express the estimates in revenue units, let μ denote a latent index excluding the random intercept, σ_ε the idiosyncratic-error standard deviation, and σ_u the random-intercept standard deviation. The conditional censored mean is

$$\begin{aligned} m(\mu, \sigma_\varepsilon; a, b) = a\,\Phi\!\left(\frac{a-\mu}{\sigma_\varepsilon}\right) + \mu\left[\Phi\!\left(\frac{b-\mu}{\sigma_\varepsilon}\right) - \Phi\!\left(\frac{a-\mu}{\sigma_\varepsilon}\right)\right] \\ + \sigma_\varepsilon\left[\phi\!\left(\frac{a-\mu}{\sigma_\varepsilon}\right) - \phi\!\left(\frac{b-\mu}{\sigma_\varepsilon}\right)\right] + b\left[1 - \Phi\!\left(\frac{b-\mu}{\sigma_\varepsilon}\right)\right]. \end{aligned}$$

$$\tilde{m}(\mu, \sigma_\varepsilon, \sigma_u; a, b) = \mathbb{E}_u[m(\mu + u, \sigma_\varepsilon; a, b)], \qquad u \sim \mathcal{N}(0, \sigma_u^2). \tag{13}$$

The censoring bounds are $(a, b) = (0, 25)$ for R^d and $(a, b) = (0, 125)$ for R^f and R^{tot}.

Applying this mapping to each sequence-block cell yields the four means $m^{\text{pre}}_{T3,g}$, $m^{\text{post}}_{T3,g}$, $m^{\text{pre}}_{T4,g}$, and $m^{\text{post}}_{T4,g}$. Substituting their estimates into the difference of Subsection 4.3.2 gives $\widehat{\Delta}^{\text{T3/T4}}_g$. Estimation on the full sample gives the aggregate counterpart $\widehat{\Delta}^{\text{T3/T4}}$.

Let $\mu^T_{it,g}(z)$ denote the sequence-specific latent index after setting $Prog_{it} = z \in \{0, 1\}$. The grouped sequence-specific observed-scale target is

$$\Delta^T_g := \mathbb{E}\Big[\tilde{m}\big(\mu^T_{it,g}(1), \sigma_\varepsilon, \sigma_u; a, b\big) - \tilde{m}\big(\mu^T_{it,g}(0), \sigma_\varepsilon, \sigma_u; a, b\big) \mid \widehat{G}_i = g,\ T_i = T\Big],$$

where T_i is the sequence assigned to i. The plug-in estimator is

$$\widehat{\Delta}^T_g = \frac{1}{n_{gT}} \sum_{(i,t):\,\widehat{G}_i = g,\ T_i = T} \Big[\tilde{m}\big(\widehat{\mu}^T_{it,g}(1), \widehat{\sigma}_\varepsilon, \widehat{\sigma}_u; a, b\big) - \tilde{m}\big(\widehat{\mu}^T_{it,g}(0), \widehat{\sigma}_\varepsilon, \widehat{\sigma}_u; a, b\big)\Big],$$

where n_{gT} is the number of subject-round pairs in the sum. The corresponding sequence-specific aggregate observed-scale target is

$$\Delta^T := \mathbb{E}\big[\tilde{m}\big(\mu^T_{it}(1), \sigma_\varepsilon, \sigma_u; a, b\big) - \tilde{m}\big(\mu^T_{it}(0), \sigma_\varepsilon, \sigma_u; a, b\big) \mid T_i = T\big],$$

with plug-in estimator $\widehat{\Delta}^T$.

A.5 Results

A.5.1 Descriptive diagnostics

Table A1 first reports descriptive patterns and cell sizes within the predicted groups, and Figures A1 and A2 then display round-by-round declarations by treatment and group. Treatment-arm summary statistics are reported in Table A2.

Table A1: Descriptive statistics by predicted evasion group, treatment arm, and sequence.

	Between variation					Within variation			
	Progressive (T1)		Linear (T2)		Fisher RI test	Linear → Progressive (T3)		Progressive → Linear (T4)	
Variable	Mean	SE	Mean	SE	Fisher p-value	Mean	SE	Mean	SE
Panel A: Low predicted evasion									
N subjects	64		52			49		48	
N rounds obs	1,280		1,040			980		960	
Age	20.28	0.23	20.79	0.22	0.339	21.20	0.31	21.33	0.43
Female	0.734	0.056	0.712	0.063	0.779	0.592	0.071	0.688	0.068
French	0.969	0.022	0.942	0.033	0.414	0.959	0.029	0.979	0.021
Experienced	0.828	0.048	0.596	0.069	0.018**	0.714	0.065	0.688	0.068
Risk preference (self)	4.34	0.17	4.44	0.24	0.605	4.78	0.24	4.81	0.24
Risk preference (task)	3.12	0.24	3.48	0.30	0.382	3.12	0.31	3.25	0.25
Spending	129.66	14.19	182.79	25.28	0.177	160.31	21.37	130.31	20.81
Saving	137.95	24.52	701.54	575.09	0.316	160.82	22.38	121.46	18.71
Full compliance ($X = Y$)	0.438	0.014	0.524	0.015	0.266	0.413	0.016	0.343	0.015
Full evasion ($X = 0$)	0.110	0.009	0.135	0.011	0.558	0.073	0.008	0.103	0.010
R^d (declaration)	17.80	0.26	18.08	0.30	0.746	17.57	0.29	16.44	0.30
R^f (fines)	7.35	0.75	4.69	0.52	0.114	5.66	0.68	5.38	0.61
R^{tot} (total)	25.14	0.68	22.77	0.49	0.031**	23.23	0.63	21.81	0.59
Decision time	6.58	0.35	6.73	0.46	0.872	5.64	0.23	7.01	0.44
Panel B: Mid predicted evasion									
N subjects	24		30			20		31	
N rounds obs	480		600			400		620	
Age	21.08	0.45	21.33	0.37	0.748	23.25	0.95	20.39	0.27
Female	0.667	0.098	0.733	0.082	0.594	0.400	0.112	0.645	0.087
French	0.917	0.058	1.000	0.000	0.207	0.950	0.050	1.000	0.000
Experienced	0.875	0.069	0.567	0.092	0.040**	0.650	0.109	0.581	0.090
Risk preference (self)	6.58	0.37	5.93	0.33	0.247	5.80	0.53	7.32	0.26
Risk preference (task)	6.25	0.30	6.67	0.29	0.252	6.45	0.54	5.23	0.34
Spending	213.38	56.71	161.53	19.41	0.814	261.25	48.19	189.35	25.33
Saving	124.96	27.91	111.17	23.06	0.536	239.00	72.20	165.16	31.84
Full compliance ($X = Y$)	0.271	0.020	0.367	0.020	0.203	0.360	0.024	0.334	0.019
Full evasion ($X = 0$)	0.062	0.011	0.082	0.011	0.766	0.095	0.015	0.116	0.013
R^d (declaration)	14.92	0.44	15.85	0.41	0.872	16.51	0.45	15.15	0.38
R^f (fines)	7.20	1.21	5.91	0.77	0.649	4.85	0.92	7.21	0.89
R^{tot} (total)	22.12	1.14	21.76	0.74	0.647	21.36	0.89	22.36	0.84
Decision time	5.44	0.26	6.26	0.41	0.250	6.90	0.63	6.30	0.41
Panel C: High predicted evasion									
N subjects	43		45			40		44	
N rounds obs	860		900			800		880	
Age	20.02	0.24	21.18	0.24	0.034**	23.27	1.20	20.61	0.27
Female	0.651	0.074	0.556	0.075	0.146	0.525	0.080	0.545	0.076
French	0.907	0.045	0.933	0.038	0.674	0.975	0.025	0.955	0.032
Experienced	0.814	0.060	0.578	0.074	0.010**	0.725	0.071	0.705	0.070
Risk preference (self)	6.91	0.28	7.31	0.23	0.102	7.42	0.27	7.41	0.17
Risk preference (task)	9.26	0.24	8.42	0.34	0.075*	8.88	0.35	9.05	0.30
Spending	173.19	27.87	174.78	19.87	0.939	198.00	39.04	186.36	34.23
Saving	129.88	26.37	156.89	26.81	0.326	290.25	44.58	232.05	69.12
Full compliance ($X = Y$)	0.455	0.017	0.418	0.016	0.565	0.395	0.017	0.360	0.016
Full evasion ($X = 0$)	0.205	0.014	0.089	0.009	0.046**	0.136	0.012	0.131	0.011
R^d (declaration)	15.65	0.37	16.40	0.33	0.568	15.94	0.36	14.61	0.34
R^f (fines)	9.76	1.09	5.58	0.61	0.110	6.35	0.79	7.13	0.81
R^{tot} (total)	25.41	1.01	21.98	0.58	0.072*	22.29	0.76	21.74	0.77
Decision time	6.39	0.84	5.28	0.29	0.311	6.26	0.53	5.32	0.27

Notes. Predicted groups are assigned before the causal comparisons using the baseline risk-set specification. T1–T2 p-values use two-sided session-day randomization inference; when a reassignment leaves a within-group arm empty, the test conditions on assignments with both arms represented. Decision-level outcomes are first averaged within subject. Stars on p-values use $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (unadjusted). The T3 and T4 columns are descriptive sequence summaries. Baseline T3/T4 identification uses within-subject regime-switch variation, so no between-sequence p-values are reported for these columns. Rows *N subjects* and *N rounds obs* are counts and therefore have no SE or p-value.

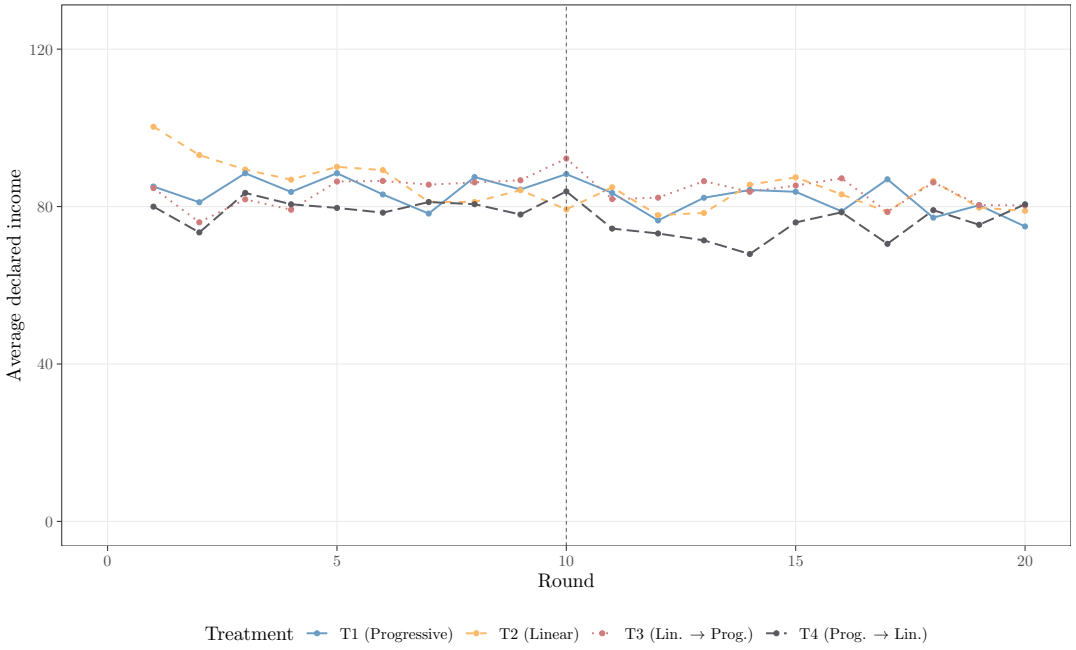

Figure A1: Average declaration by round and treatment.

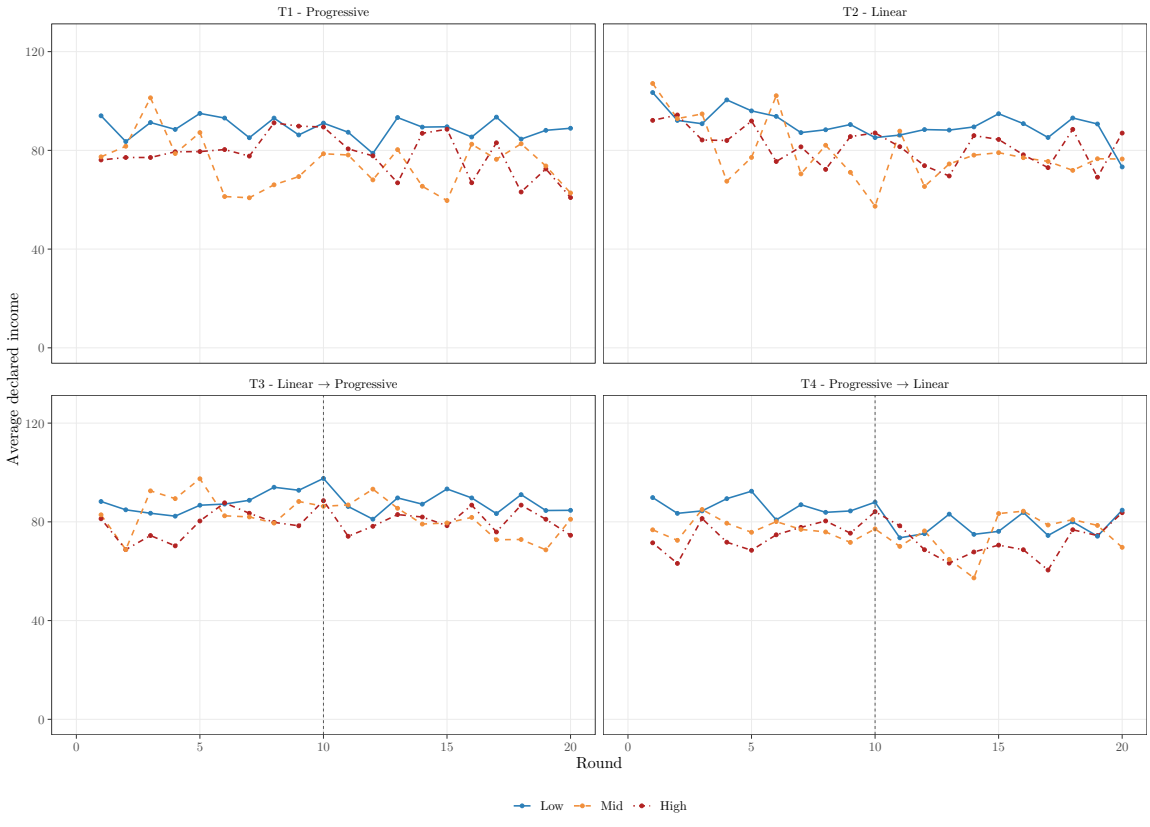

Figure A2: Average declaration by round, treatment, and predicted group (group means).

Table A2: Descriptive statistics by treatment arm and sequence.

	T1 (Progressive)		**T2 (Linear)**		**Fisher RI**	**T3 (Lin. → Prog.)**		**T4 (Prog. → Lin.)**	
Variable	Mean	SE	Mean	SE	p(T1-T2)	Mean	SE	Mean	SE
N subjects	131		127			109		123	
N rounds obs	2,620		2,540			2,180		2,460	
Age	20.34	0.16	21.06	0.15	0.036**	22.34	0.50	20.84	0.21
Female	0.695	0.040	0.661	0.042	0.506	0.532	0.048	0.626	0.044
French	0.939	0.021	0.953	0.019	0.530	0.963	0.018	0.976	0.014
Experienced	0.832	0.033	0.583	0.044	0.002***	0.706	0.044	0.667	0.043
Risk preference (self)	5.60	0.18	5.81	0.19	0.213	5.94	0.21	6.37	0.17
Risk preference (task)	5.71	0.28	5.98	0.27	0.291	5.84	0.33	5.82	0.28
Spending	159.28	15.56	174.93	13.26	0.641	192.66	19.51	165.24	16.09
Saving	132.92	15.54	369.09	235.67	0.369	222.66	23.74	172.03	27.14
Full compliance ($X = Y$)	0.413	0.028	0.449	0.026	0.402	0.397	0.028	0.347	0.026
Full evasion ($X = 0$)	0.132	0.007	0.106	0.006	0.459	0.100	0.006	0.116	0.006
R^d (declaration)	16.57	0.50	16.96	0.49	0.601	16.78	0.58	15.46	0.56
R^f (fines)	8.11	0.76	5.29	0.44	0.038**	5.77	0.59	6.46	0.56
R^{tot} (total)	24.68	0.48	22.25	0.40	0.012**	22.54	0.43	21.93	0.47
Decision time	6.31	0.33	6.10	0.24	0.698	6.10	0.25	6.23	0.22

Notes. p(T1–T2) is the two-sided session-day randomization p-value obtained by enumerating all $\binom{13}{6} = 1{,}716$ admissible assignments. Full compliance and revenue means, standard errors, and randomization tests use one subject mean across rounds per subject. Other decision-level randomization tests also use subject means. Stars on p(T1–T2): $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (unadjusted). Rows *N subjects* and *N rounds obs* are counts and therefore have no SE or p-value.

A.5.2 Distributional evidence

Table A3 complements the main-text ECDF analysis with the full enumeration of quantile treatment effects.

Table A3: Quantile treatment effects on expected total revenue in T1–T2.

Quantile u	$Q_{T1}(u)$	$Q_{T2}(u)$	**QTE**	CI_{90}	CI_{95}	CI_{99}	p**-value**
0.10	21.700	19.131	2.570***	[1.551, 3.105]	[1.367, 3.201]	[1.080, 3.418]	< 0.001
0.25	22.662	21.154	1.508***	[1.030, 2.329]	[0.955, 2.461]	[0.840, 2.775]	< 0.001
0.50	24.052	22.372	1.681***	[1.183, 2.247]	[1.061, 2.352]	[0.840, 2.502]	< 0.001
0.75	24.800	23.670	1.130***	[0.759, 1.494]	[0.703, 1.549]	[0.498, 1.736]	< 0.001
0.90	25.000	24.755	0.245**	[0.017, 0.689]	[0.014, 0.721]	[0.000, 0.962]	0.012

Notes. Entries report T1–T2 quantile differences in subject-level mean expected total revenue. Quantiles are obtained by inverting each arm's empirical distribution without interpolation. T1 has 131 subjects and T2 has 127 subjects. Confidence intervals and plus-one two-sided p-values use a subject bootstrap resampling subjects with replacement within arm ($B = 2{,}000$). Significance stars use these p-values: * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$. Because the underlying outcomes have mass points and the empirical quantile functions have flat segments, some QTEs can be exactly zero by construction; the ECDF bands in Figure 2 are therefore the more informative distributional object.

A.5.3 Grouped and aggregate observed-scale differences

The following tables report the complete grouped and aggregate observed-scale estimates underlying the compact figures and selected comparisons in Section 5.

Table A4: Grouped observed-scale revenue differences in T1–T2 and T3/T4 (Prog. − Lin.).

	R^d			R^f			R^{tot}		
	Low	Mid	High	Low	Mid	High	Low	Mid	High
Panel A: T1 Progressive, T2 Linear. Difference: Prog. - Lin.									
Shared column layout for Panels A–D: outcomes (R^d, R^f, R^{tot}) by group (Low, Mid, High).									
$\widehat{\beta}$	0.239	0.603	0.612	1.591	-0.385	3.063	1.830*	0.218	3.675**
SE	(1.172)	(1.521)	(1.395)	(1.279)	(1.974)	(1.604)	(0.868)	(1.352)	(1.142)
p	0.799	0.783	0.661	0.293	0.893	0.206	0.055	0.840	0.037
CI_{99}	[-3.639, 3.498]	[-6.223, 5.995]	[-4.121, 5.893]	[-3.818, 7.206]	[-8.539, 7.003]	[-5.194, 8.694]	[-1.217, 4.887]	[-5.782, 3.813]	[-1.255, 6.874]
CI_{95}	[-2.215, 2.236]	[-4.315, 4.644]	[-2.551, 4.490]	[-1.550, 5.683]	[-5.541, 4.819]	[-1.967, 6.599]	[-0.093, 4.316]	[-3.569, 2.517]	[0.401, 5.750]
CI_{90}	[-1.720, 1.835]	[-3.427, 4.136]	[-1.985, 3.658]	[-0.950, 4.953]	[-4.813, 3.648]	[-1.047, 5.786]	[0.312, 3.990]	[-2.659, 1.977]	[1.003, 5.355]
Panel B: Pooled T3/T4. Difference: Prog. - Lin.									
$\widehat{\Delta}$	0.868	-0.760	0.460	1.781	-0.589	3.670	3.646**	-0.099	4.185**
SE	(1.096)	(1.823)	(1.021)	(1.605)	(2.497)	(2.281)	(1.619)	(1.921)	(2.060)
p	0.414	0.712	0.649	0.268	0.791	0.135	0.024	0.974	0.044
CI_{99}	[-1.936, 3.869]	[-5.402, 3.822]	[-2.081, 3.066]	[-2.024, 6.051]	[-7.314, 5.716]	[-2.612, 9.358]	[-0.407, 7.823]	[-5.106, 5.286]	[-0.994, 9.977]
CI_{95}	[-1.201, 3.050]	[-4.344, 2.731]	[-1.470, 2.462]	[-1.182, 5.084]	[-5.523, 4.141]	[-0.963, 8.106]	[0.577, 6.950]	[-3.903, 3.669]	[0.182, 8.328]
CI_{90}	[-0.920, 2.642]	[-3.744, 2.179]	[-1.199, 2.191]	[-0.741, 4.506]	[-4.718, 3.295]	[-0.279, 7.094]	[1.058, 6.334]	[-3.230, 2.948]	[0.955, 7.486]
Panel C: T3 Linear → Progressive. Difference: Prog. - Lin.									
$\widehat{\Delta}$	-1.694*	-2.153	-1.195	1.876	1.786	3.176*	1.308	0.374	2.166
SE	(0.933)	(1.475)	(1.053)	(2.029)	(2.532)	(2.174)	(1.995)	(1.978)	(2.077)
p	0.066	0.150	0.249	0.220	0.336	0.089	0.576	0.863	0.349
CI_{99}	[-4.232, 0.593]	[-6.047, 1.715]	[-3.885, 1.240]	[-1.951, 8.074]	[-3.685, 9.950]	[-1.836, 9.303]	[-3.609, 6.690]	[-4.907, 5.371]	[-3.156, 7.517]
CI_{95}	[-3.557, 0.115]	[-5.051, 0.644]	[-3.392, 0.727]	[-1.156, 6.683]	[-2.105, 7.530]	[-0.554, 8.030]	[-2.560, 5.242]	[-3.475, 4.382]	[-2.047, 6.150]
CI_{90}	[-3.260, -0.237]	[-4.513, 0.285]	[-3.001, 0.440]	[-0.691, 6.041]	[-1.460, 6.651]	[0.127, 7.296]	[-1.908, 4.746]	[-2.983, 3.709]	[-1.401, 5.427]
Panel D: T4 Progressive → Linear. Difference: Prog. - Lin.									
$\widehat{\Delta}$	2.353***	1.332	1.503*	1.543	-1.153	1.689	5.061**	2.245	4.852**
SE	(0.851)	(0.972)	(0.811)	(1.987)	(2.217)	(1.991)	(1.872)	(2.237)	(2.336)
p	0.006	0.181	0.059	0.427	0.579	0.384	0.011	0.320	0.026
CI_{99}	[0.360, 4.614]	[-1.169, 3.970]	[-0.559, 3.694]	[-4.527, 6.383]	[-7.302, 4.351]	[-3.452, 7.572]	[0.021, 10.045]	[-3.185, 8.010]	[-0.674, 11.328]
CI_{95}	[0.707, 4.054]	[-0.539, 3.195]	[-0.045, 3.077]	[-2.499, 5.219]	[-5.499, 3.075]	[-2.120, 5.656]	[1.395, 8.471]	[-2.110, 6.548]	[0.446, 9.705]
CI_{90}	[0.928, 3.764]	[-0.288, 2.864]	[0.151, 2.794]	[-1.719, 4.676]	[-4.912, 2.375]	[-1.520, 5.033]	[1.986, 8.107]	[-1.343, 5.949]	[1.098, 8.680]

Notes. Estimates use the baseline covariate specification. Panel A reports T1–T2 between-subject ANCOVA differences. Panel B reports observed-scale margins from the pooled RE-Tobit with full round fixed effects and is the randomized-sequence ITT on the pre-post change. Panels C and D report observed-scale margins from separate sequence RE-Tobits with a linear round trend; they are diagnostics, not standalone schedule effects. All differences are Prog. − Lin. Within each panel and outcome, rows report the estimate, SE, p-value, and confidence intervals (CI_{99}, CI_{95}, CI_{90}). Standard errors are from the subject bootstrap. The T1–T2 p-values and confidence intervals use exact session-day randomization inference. The pooled T3/T4 p-values and confidence intervals use the subject bootstrap. Panel B uses subject-bootstrap SEs, plus-one p-values, and percentile confidence intervals. Panels C and D use subject-bootstrap SEs and p-values.

Table A5: Aggregate observed-scale revenue differences in T1–T2 and T3/T4 (Prog. − Lin.).

	T1 (Prog.) vs. T2 (Lin.) (1)	Pooled T3/T4 (2)	T3 (Lin. → Prog.) (3)	T4 (Prog. → Lin.) (4)
Baseline covariate specification				
R^d	0.355	0.403	-1.668**	1.838***
	(0.767)	(0.716)	(0.789)	(0.607)
	p=0.622	p=0.582	p=0.039	p=0.004
	CI_{99}=[-2.013, 2.907]	CI_{99}=[-1.467, 2.068]	CI_{99}=[-3.639, 0.307]	CI_{99}=[0.249, 3.349]
	CI_{95}=[-1.251, 1.976]	CI_{95}=[-1.020, 1.765]	CI_{95}=[-3.243, -0.165]	CI_{95}=[0.659, 3.034]
	CI_{90}=[-0.974, 1.645]	CI_{90}=[-0.803, 1.604]	CI_{90}=[-2.997, -0.347]	CI_{90}=[0.880, 2.865]
R^f	1.701	1.803	2.848*	0.989
	(0.862)	(1.172)	(1.576)	(1.623)
	p=0.107	p=0.130	p=0.069	p=0.538
	CI_{99}=[-2.165, 5.175]	CI_{99}=[-1.153, 4.808]	CI_{99}=[-1.186, 7.213]	CI_{99}=[-3.043, 5.074]
	CI_{95}=[-0.592, 4.064]	CI_{95}=[-0.531, 4.032]	CI_{95}=[-0.196, 6.065]	CI_{95}=[-2.108, 4.181]
	CI_{90}=[-0.065, 3.550]	CI_{90}=[-0.147, 3.632]	CI_{90}=[0.363, 5.524]	CI_{90}=[-1.605, 3.682]
R^{tot}	2.055**	2.874***	1.328	4.269**
	(0.609)	(1.122)	(1.694)	(1.823)
	p=0.022	p=0.005	p=0.440	p=0.017
	CI_{99}=[-0.486, 4.759]	CI_{99}=[0.401, 5.802]	CI_{99}=[-3.156, 5.667]	CI_{99}=[-0.196, 8.867]
	CI_{95}=[0.406, 3.825]	CI_{95}=[0.844, 5.298]	CI_{95}=[-2.052, 4.641]	CI_{95}=[0.815, 7.933]
	CI_{90}=[0.751, 3.448]	CI_{90}=[1.149, 4.910]	CI_{90}=[-1.535, 4.087]	CI_{90}=[1.527, 7.477]

Notes. Estimates use the baseline covariate specification. Column (1) reports T1–T2 between-subject ANCOVA differences. Column (2) reports observed-scale margins from the pooled RE-Tobit with full round fixed effects and is the randomized-sequence ITT on the pre-post change. Columns (3) and (4) report observed-scale margins from separate sequence RE-Tobits with a linear round trend; they are diagnostics, not standalone schedule effects. All differences are Prog. − Lin. For each outcome and design, rows report the estimate, SE, two-sided p-value, and confidence intervals (CI_{99}, CI_{95}, CI_{90}). Standard errors are from the subject bootstrap. The T1–T2 p-values and confidence intervals use exact session-day randomization inference. The pooled T3/T4 p-values and confidence intervals use the subject bootstrap. The pooled T3/T4 column uses subject-bootstrap SEs, plus-one p-values, and percentile confidence intervals. The separate T3 and T4 diagnostics use subject-bootstrap SEs, plus-one p-values, and percentile confidence intervals.

A.6 Robustness and design diagnostics

A.6.1 Grouping validity

Table A6 reports the grouped R^{tot} differences behind the partition frontier of Subsection 6.1, in both designs, for the demographic set and the average-observed-evasion benchmark alongside the baseline risk set.

Table A6: Grouped total-revenue differences across the partition frontier in T1–T2 and pooled T3/T4 (Prog. − Lin.).

	T1–T2 (exact session-day RI)		**Pooled T3/T4 (Prog. − Lin.)**	
	Estimate: $\widehat{\beta}_{g,R^{tot}}$		**Estimate:** $\widehat{\Delta}^{\mathrm{T3/T4}}_{g,R^{tot}}$	
Partition design	**Low**	**High**	**Low**	**High**
Demographic	2.00 ([-0.91, 5.71])	3.30** ([1.30, 6.36])	0.84 ([-2.19, 4.13])	5.23** ([1.35, 8.94])
Risk (baseline)	1.83* ([-0.09, 4.32])	3.67** ([0.40, 5.75])	3.65** ([0.58, 6.95])	4.18** ([0.18, 8.33])
Observed behavior	0.83 ([-0.43, 2.04])	8.17* ([-0.89, 17.35])	0.54 ([-1.30, 2.70])	5.62 ([-2.12, 14.36])

Notes. All estimates use the baseline covariate adjustment. The T1–T2 columns report between-subject differences. Their intervals and p-values use exact session-day randomization inference, with test inversion over all $\binom{13}{6} = 1{,}716$ reassignments. The pooled T3/T4 columns report observed-scale margins from the pooled RE-Tobit with full round fixed effects and are randomized-sequence ITT estimates on the pre-post change (Prog. − Lin.). Their intervals and p-values use the subject bootstrap. Brackets contain 95% intervals. Stars use the two-sided p-value from the indicated inference method: * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

A.6.2 Alternative hyperparameter selection

The first check assesses sensitivity to model tuning while holding the causal estimators fixed. We recompute the groups and effects under the top 10 support-valid configurations, including the baseline. All are probability-forest specifications, so the comparison varies forest tuning within a single model family. Figure A3 reports the resulting effect envelope. Online Appendix Section B.3 reports the selected configurations, fixed search grids, cell-level summaries, and partition similarity.

Across the 10 selected configurations, the main cells are stable, and the reported baseline intervals use the subject bootstrap. For T1–T2 High total revenue, the baseline estimate 3.67 ($[1.57, 6.01]$) is close to the median across selected configurations, 3.73, with range $[3.53, 3.89]$. All 10 retain the baseline sign. Pooled T3/T4 High total revenue is similarly stable: the baseline estimate is 4.18 ($[0.18, 8.33]$), the median is 4.22, the range is $[4.02, 4.39]$, and all 10 retain the baseline sign. Pooled T3/T4 Low also preserves its sign (baseline 3.65, $[0.58, 6.95]$; median 3.97, range $[3.65, 4.27]$), while pooled T3/T4 Mid remains small relative to its uncertainty. Classification agreement is high (median ARI 0.91). The baseline lies near the center of these within-family estimates, so the point-estimate pattern is not tied to a single forest specification.

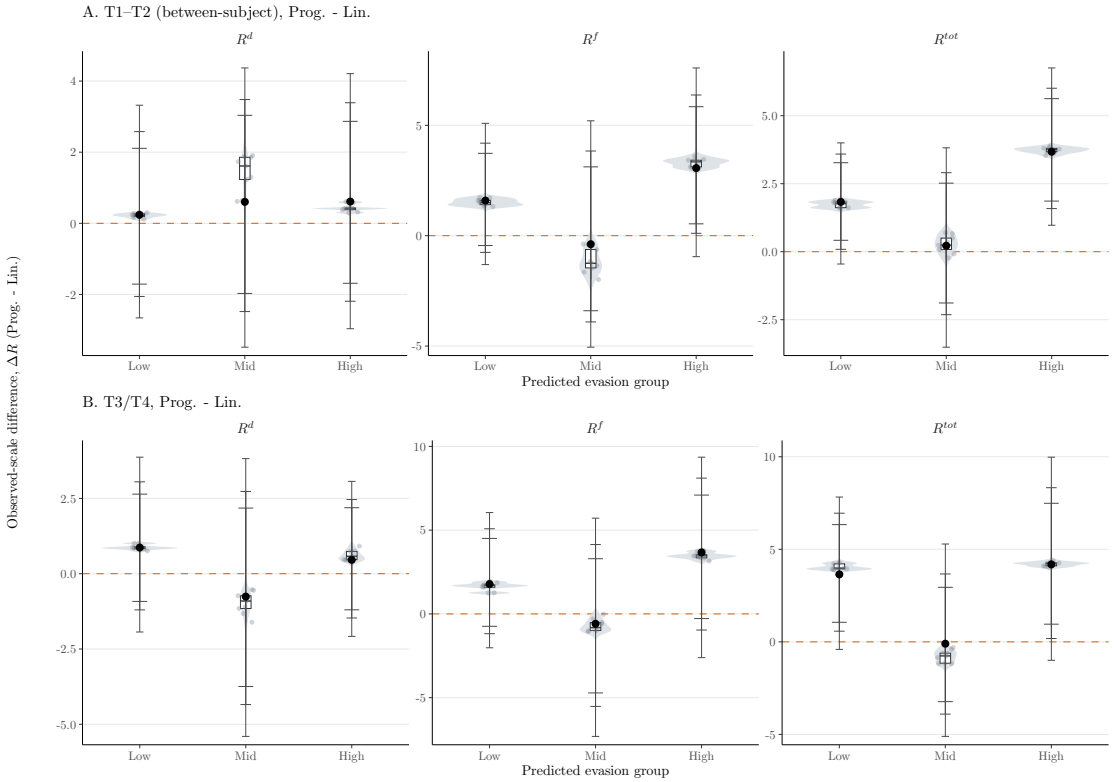

Figure A3: Alternative hyperparameter selection for the risk set (baseline specification). Panel A reports T1–T2 between-subject differences; Panel B reports pooled T3/T4 randomized-sequence ITT estimates on the pre-post change (Prog. − Lin.). Black markers report the baseline estimates with nested 90%, 95%, and 99% subject-bootstrap confidence intervals. Violins, embedded boxplots, and gray points summarize the other nine selected configurations.

A.6.3 Assignment uncertainty

Because the first-stage classifier assigns each subject to a single group, grouped effects may be sensitive to hard class boundaries. We test this by holding the downstream causal estimators fixed and perturbing only assignment through three complementary checks: *low-margin exclusion*, *boundary flip*, and *stochastic reassignment based on posterior probabilities*.

Low-margin exclusion ranks subjects by the gap between the two largest $\widehat{\pi}_i(g)$, and drops the bottom 10% or 20% of the sample, so estimates are recomputed on a smaller but higher-confidence subset.

Boundary flip is run on exactly the subjects selected in the low-margin step. For each, we replace the mode of $\widehat{\pi}_i$ with the runner-up, providing an adversarial boundary-misclassification check.

Stochastic reassignment based on posterior probabilities performs fifty replications. In each, a single class is drawn per subject from $\widehat{\pi}_i$, and the grouped effects are re-estimated under the resulting partition. Unlike the exclusion and boundary-flip checks, which per-

turb only the most ambiguous units, this exercise perturbs assignments throughout the sample and assesses sensitivity to global classification uncertainty.

Figure A4 visualizes the full assignment-uncertainty effect envelope. The perturbation-design shares, baseline-to-scenario reassignment summaries, and cell-level effect summaries for the deterministic perturbations and the stochastic reassignment exercise are reported in Online Appendix Section B.3.

The key total-revenue cells remain directionally stable. In T1–T2 High, all stochastic reassignments retain the baseline sign. For the pooled T3/T4 High and Low baselines of Subsection A.6.2, the corresponding shares are 100% and 96%. Pooled T3/T4 Mid is instead near zero and directionally unstable.

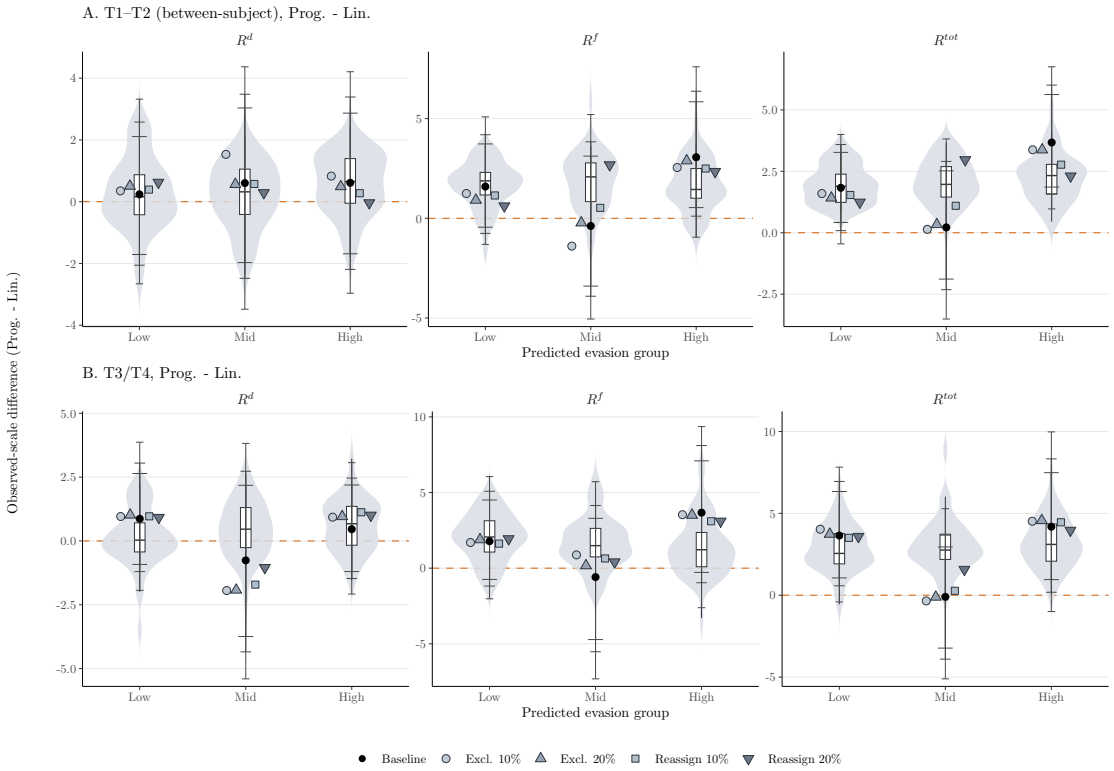

Figure A4: Assignment-uncertainty robustness envelope for the risk set (baseline specification). The upper row reports T1–T2 ANCOVA differences and the lower row reports pooled T3/T4 randomized-sequence ITT estimates on the pre-post change; columns report outcomes (R^d, R^f, R^{tot}). In each panel, the black marker reports the baseline estimate with nested 90%, 95%, and 99% subject-bootstrap confidence intervals. The violins summarize stochastic reassignment based on posterior probabilities, while the remaining markers report the exclusion and boundary-flip checks.

A.6.4 Temporal and individual fixed-effects diagnostics

Table A7 reports the estimates underlying Figure 6. Table A8 and Figure A5 report the corresponding sequence-specific diagnostics.

These tables add two details to the pattern described in Subsection 6.2. Under the trend specifications, T3 is negative or near zero for the aggregate as well as for High. Mid remains uninformative throughout.

Table A7: Pooled T3/T4 total-revenue differences under temporal and individual fixed-effects checks (Prog. − Lin.).

Group	Statistic	Baseline	Rounds 8–13	Individual and round FE
Low	Estimate	3.65**	4.25	3.67**
	p-value	0.024	0.178	0.022
	CI_{90}	[1.06, 6.33]	[-0.68, 9.13]	[1.05, 6.44]
	CI_{95}	[0.58, 6.95]	[-1.62, 10.06]	[0.60, 7.04]
	CI_{99}	[-0.41, 7.82]	[-3.90, 12.07]	[-0.49, 8.20]
Mid	Estimate	-0.10	5.37	-0.24
	p-value	0.974	0.195	0.889
	CI_{90}	[-3.23, 2.95]	[-1.58, 12.70]	[-3.15, 2.95]
	CI_{95}	[-3.90, 3.67]	[-3.31, 13.95]	[-3.62, 3.55]
	CI_{99}	[-5.11, 5.29]	[-5.98, 17.19]	[-5.05, 4.47]
High	Estimate	4.18**	8.03**	4.08**
	p-value	0.044	0.040	0.039
	CI_{90}	[0.96, 7.49]	[1.60, 13.87]	[0.82, 7.69]
	CI_{95}	[0.18, 8.33]	[0.33, 15.05]	[0.21, 8.23]
	CI_{99}	[-0.99, 9.98]	[-2.04, 17.39]	[-0.91, 9.80]
Aggregate	Estimate	2.87***	5.78***	2.92***
	p-value	0.005	0.007	0.010
	CI_{90}	[1.15, 4.91]	[2.22, 9.45]	[1.10, 4.82]
	CI_{95}	[0.84, 5.30]	[1.50, 10.06]	[0.70, 5.17]
	CI_{99}	[0.40, 5.80]	[0.23, 11.80]	[0.03, 5.98]

Notes. Entries report observed-scale pooled T3/T4 randomized-sequence ITT estimates on the pre-post change for R^{tot} (Prog. − Lin.). The baseline is the full-sample RE-Tobit with round fixed effects and standard covariates. The second specification restricts the sample to rounds 8–13, while the third includes individual and round fixed effects. Rows cover the Low, Mid, and High predicted groups and the aggregate sample. For each row, the table reports the point estimate, two-sided subject-bootstrap plus-one p-value, and subject-bootstrap percentile confidence intervals (CI_{90}, CI_{95}, CI_{99}). Significance stars use subject-bootstrap p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Separate T3 and T4 total-revenue diagnostics under temporal and individual fixed-effects checks (Prog. − Lin.).

Group	Statistic	Separate RE-Tobit	Pooled trend	Individual FE + trend	Separate RE-Tobit	Pooled trend	Individual FE + trend
		T3 (Prog-Lin)			T4 (Prog-Lin)		
Low	Estimate	1.31	1.71	1.52	5.06**	1.97	2.09
	p-value	0.576	0.385	0.426	0.011	0.188	0.189
	CI_{90}	[-1.91, 4.75]	[-1.61, 4.75]	[-1.68, 4.72]	[1.99, 8.11]	[-0.51, 4.71]	[-0.47, 4.80]
	CI_{95}	[-2.56, 5.24]	[-2.19, 5.48]	[-2.40, 5.75]	[1.40, 8.47]	[-1.05, 5.24]	[-1.07, 5.33]
	CI_{99}	[-3.61, 6.69]	[-3.24, 6.71]	[-3.60, 6.88]	[0.02, 10.05]	[-2.24, 6.30]	[-2.22, 6.40]
Mid	Estimate	0.37	1.21	1.49	2.25	-1.24	-1.75
	p-value	0.863	0.627	0.539	0.320	0.647	0.492
	CI_{90}	[-2.98, 3.71]	[-2.62, 5.23]	[-2.22, 5.51]	[-1.34, 5.95]	[-5.47, 2.99]	[-5.81, 2.38]
	CI_{95}	[-3.48, 4.38]	[-3.45, 5.92]	[-3.00, 6.35]	[-2.11, 6.55]	[-6.07, 3.87]	[-6.84, 3.21]
	CI_{99}	[-4.91, 5.37]	[-4.67, 7.83]	[-3.78, 7.40]	[-3.18, 8.01]	[-7.41, 5.22]	[-8.56, 4.32]
High	Estimate	2.17	-2.91	-2.77	4.85**	7.02***	6.60***
	p-value	0.349	0.293	0.272	0.026	0.008	0.008
	CI_{90}	[-1.40, 5.43]	[-7.37, 1.72]	[-7.35, 1.35]	[1.10, 8.68]	[2.53, 11.99]	[2.33, 11.37]
	CI_{95}	[-2.05, 6.15]	[-8.17, 2.51]	[-8.14, 2.24]	[0.45, 9.71]	[1.70, 12.80]	[1.60, 12.24]
	CI_{99}	[-3.16, 7.52]	[-9.92, 4.09]	[-9.89, 3.72]	[-0.67, 11.33]	[0.40, 15.52]	[0.35, 14.28]
Aggregate	Estimate	1.33	-0.10	-0.02	4.27**	3.13**	2.87**
	p-value	0.440	0.946	0.995	0.017	0.028	0.032
	CI_{90}	[-1.53, 4.09]	[-2.27, 2.11]	[-2.28, 2.16]	[1.53, 7.48]	[0.87, 5.35]	[0.58, 5.14]
	CI_{95}	[-2.05, 4.64]	[-2.74, 2.50]	[-2.73, 2.62]	[0.82, 7.93]	[0.40, 5.79]	[0.19, 5.58]
	CI_{99}	[-3.16, 5.67]	[-3.83, 3.52]	[-3.62, 3.36]	[-0.20, 8.87]	[-0.25, 6.98]	[-0.61, 6.64]

Notes. Entries report separate T3 and T4 sequence diagnostics for R^{tot} (Prog. − Lin.), not standalone schedule effects. Rows cover the Low, Mid, and High predicted groups and the aggregate sample. For each row, the table reports the point estimate, two-sided subject-bootstrap plus-one p-value, and subject-bootstrap percentile confidence intervals (CI_{90}, CI_{95}, CI_{99}). Within each sequence, the columns report the separate-sequence RE-Tobit, the pooled linear-trend diagnostic, and the specification with individual fixed effects and a linear round trend. Significance stars use subject-bootstrap p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

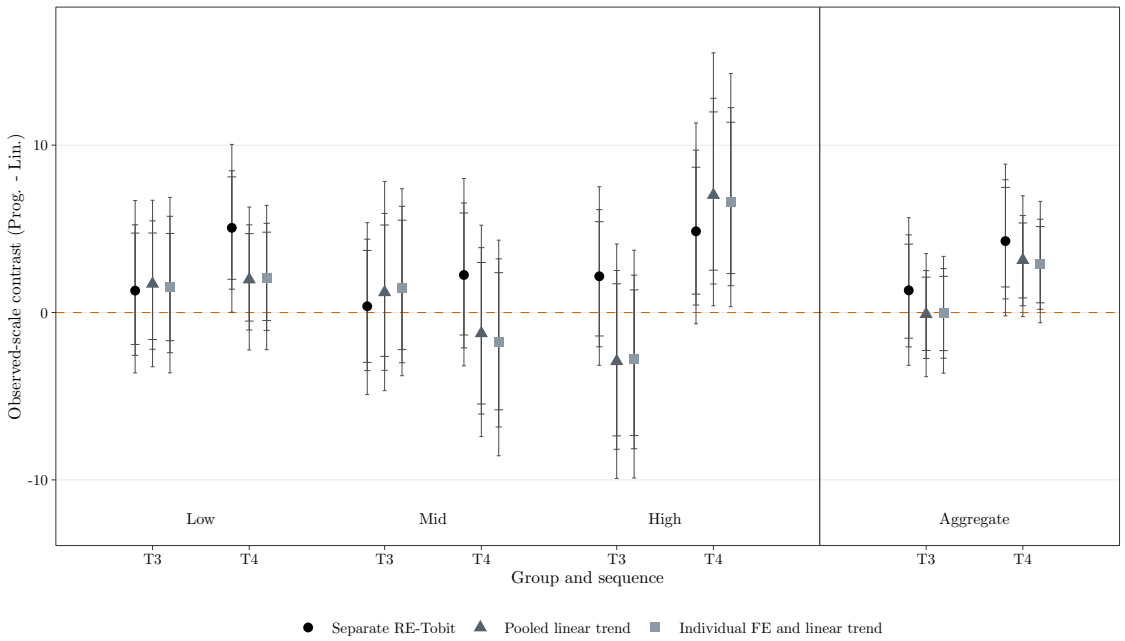

Figure A5: Sequence-specific T3 and T4 trend and individual-fixed-effects diagnostics for R^{tot} by group and aggregate. Markers are point estimates; bars show nested 90%, 95%, and 99% subject-bootstrap confidence intervals.