

Cost-Sensitive Calibration of Escalation Thresholds for Retirement Account Withdrawals and the Trade-Off Between Fraud Loss and Customer Friction

Olasunbo O Fagbore [1*], Laura B Whitcombe [2], Kenji R Moriyama [3], Grace N Ellery [4]
[1] Senior Brokerage Consultant II, Fidelity Investment, Westlake, TX, USA
[2-4] Center for Retirement Security Research, AL, USA

* Corresponding Author: **Olasunbo O Fagbore**

Article Info

P-ISSN: 3051-3618
E-ISSN: 3051-3626
Volume: 07
Issue: 02
Received: 12-08-2026
Accepted: 10-09-2026
Published: 08-10-2026
Page No: 139-148

Abstract
Firms that screen retirement withdrawals for scams must choose how suspicious a request has to be before it is escalated for review or held. A low threshold stops more scams but delays and frustrates legitimate customers, while a high threshold does the reverse. We estimate both sides of this trade-off and derive cost-minimizing escalation policies. Using 1,264,530 retirement withdrawal requests from owners aged 60 and older at a national custodian between 2022 and 2025, we combine a sharp regression discontinuity at the custodian's existing risk-score threshold with a 24-week randomized experiment that assigned requests to one of three thresholds. Escalation at the margin reduced expected scam losses by 1,840 dollars per escalated request while raising 30-day abandonment by 2.1 percentage points and 180-day account attrition by 0.9 percentage points among legitimate customers. In the experiment, lowering the threshold from 0.62 to 0.50 increased escalations by 58 percent and prevented an additional 22 percent of scam dollars, whereas raising it to 0.72 cut escalations by 37 percent but allowed 31 percent more scam dollars through. A cost model calibrated to these estimates places the loss-minimizing single threshold near 0.54. A two-tier policy that routes moderate-risk requests to a brief structured call and holds only high-risk requests lowers total expected cost by a further 17 percent. Optimal thresholds vary substantially by request size and channel, and queue capacity can erase much of the benefit of aggressive thresholds during demand surges. Escalation rates at the optimal policy did not differ meaningfully by neighbourhood income.

DOI: https://doi.org/10.54660/IJMFD.2026.7.2.139-148

Keywords: fraud prevention, escalation threshold, cost-sensitive classification, regression discontinuity, customer friction, elder financial exploitation, retirement accounts

1. Introduction

Every automated fraud screen eventually reduces to a threshold. Above it, a request is escalated for human review, delayed, or held. Below it, the money moves. In payment-card fraud, the threshold can be tuned aggressively because a false decline costs a few seconds and the customer can try another card. Retirement withdrawals are different. A delayed distribution may mean a missed home closing, a late medical bill, or a frightened retiree who concludes that the firm does not trust them. At the same time, a missed scam can consume a lifetime of savings. Treating the threshold as a technical parameter set by the model team obscures a policy choice with large stakes on both sides.

Older adults are increasingly targeted by scams that persuade them to request withdrawals themselves. Complaints to the FBI by people aged 60 and older reported roughly 3.4 billion dollars in losses in 2023 (Federal Bureau of Investigation, 2024) [9], and survey-based estimates suggest that about one in twenty older Americans is defrauded each year (Burnes *et al.*, 2017) [2]. FINRA Rule 2165 allows firms to place temporary holds when exploitation is suspected (FINRA, 2018) [10], and many custodians

now run machine-learning risk scores on every withdrawal. The question is no longer whether to screen but where to draw the line.

To see why the choice is hard, consider a withdrawal of 40,000 dollars scored at 0.58 on a zero-to-one risk scale. If roughly one in sixty such requests is scam-induced and the average scam loss is 70,000 dollars, the expected scam loss is about 1,170 dollars. Whether escalation is worthwhile depends on how much of that loss review would prevent, on what the review costs, and on the friction it imposes on the fifty-nine legitimate customers. Those quantities are rarely measured, so thresholds are typically set by intuition, by review capacity, or by the desire to hit a target alert volume.

This paper measures them. We exploit two sources of exogenous variation in escalation at a national retirement custodian. The first is a sharp regression discontinuity created by a fixed escalation threshold that applied to all requests for 30 months. The second is a randomized experiment in which the custodian assigned requests to one of three thresholds for 24 weeks. Together these designs identify the effect of escalation on scam losses and on legitimate customers at and away from the existing cut-off. We then embed the estimates in a cost-sensitive decision framework (Elkan, 2001) [8] to derive optimal single and two-tier thresholds, examine how they vary across segments, account for queue capacity, and check for disparities.

2. Related Literature

The foundations of cost-sensitive classification establish that the optimal threshold on a calibrated probability depends only on the ratio of misclassification costs (Elkan, 2001) [8]. Provost and Fawcett (2001) [18] showed how ROC convex hull analysis identifies optimal classifiers under imprecise costs, and Hand (2009) [13] argued that AUROC implicitly averages over cost ratios in a way that may not reflect any real decision. In fraud detection, example-dependent costs that scale with transaction amount have been shown to change optimal decisions materially (Bahnsen, Aouada, and Ottersten, 2014) [1], and transaction aggregation strategies shape what a threshold acts on (Whitrow *et al*., 2009) [22]. Calibration of predicted probabilities is a prerequisite for any of these analyses (Dal Pozzolo *et al*., 2015) [6]. In medicine, decision curve analysis offers a closely related framework for choosing thresholds based on net benefit (Vickers and Elkin, 2006) [21], and the Youden index provides a simple but cost-agnostic alternative (Youden, 1950) [23].

Much less is known about the costs of friction. Behavioural research documents that small administrative burdens, sometimes called sludge, can deter people from completing transactions they value (Thaler and Sunstein, 2008; Sunstein, 2021) [19, 20]. In financial services, friction may be especially costly for older customers who value autonomy and may interpret extra scrutiny as paternalism. Fairness research shows that thresholds applied to risk scores can produce unequal error rates across groups even when the score is well calibrated (Chouldechova, 2017; Kleinberg, Mullainathan, and Raghavan, 2017) [5, 15]. Finally, regression discontinuity methods provide credible local estimates of the effect of crossing a threshold (Imbens and Lemieux, 2008; Lee and Lemieux, 2010) [14, 17], and we use modern robust inference for these designs (Calonico, Cattaneo, and Titiunik, 2014; Cattaneo, Jansson, and Ma, 2020) [3, 4].

3. Setting, Data, and Identification

3.1. Setting

The custodian scores every retirement withdrawal request in real time using a gradient-boosted model calibrated to output the probability that the request will later be confirmed as scam-induced, rescaled monotonically to a zero-to-one score. From January 2022 to June 2024, every request with a score of 0.62 or higher was escalated to a second-line team, which contacted the owner, could consult a trusted contact, and could place a hold under Rule 2165. Requests below 0.62 were processed automatically unless a representative raised a manual concern. The score and threshold were not visible to customers or frontline staff.

3.2. Data

We observe 1,264,530 withdrawal requests from 512,688 owners aged 60 or older between January 2022 and June 2025. For each request we observe the score, amount, channel, owner demographics from account records, escalation and hold outcomes, timestamps for each step, and downstream outcomes including disbursement, recall, recovery, cancellation by the customer, complaints, and account closure. Scam status comes from investigations closed within 180 days. Overall, 3,417 requests (0.27 percent) were confirmed scam-induced.

3.3. Regression Discontinuity

Between January 2022 and June 2024 the threshold produced a sharp discontinuity in escalation. Let S denote the score and $c = 0.62$ the cut-off. For outcome Y we estimate

$$\tau_RD = \lim_{s\downarrow c} E[Y \mid S = s] - \lim_{s\uparrow c} E[Y \mid S = s] \quad (1)$$

using local linear regression with a triangular kernel, MSE-optimal bandwidths, and bias-corrected robust confidence intervals (Calonico, Cattaneo, and Titiunik, 2014) [3]. Because neither customers nor staff observed the score, manipulation is implausible, which we verify with a density test (Cattaneo, Jansson, and Ma, 2020) [4]. We avoid high-order global polynomials (Gelman and Imbens, 2019) [11]. Scam-loss outcomes are estimated on all requests, so the estimate captures expected dollars lost per request near the threshold.

3.4. Randomized Threshold Experiment

From October 2024 to March 2025 the custodian randomized each new request with equal probability to one of three thresholds, namely 0.50, 0.62, and 0.72. Randomization was at the owner level, so repeat requests from the same owner received the same arm. The experiment was approved by the firm's model risk and compliance committees and by the institutional review board of the lead author's institution for secondary analysis. The experiment identifies the average effect of each policy across the whole score distribution, not just at the original cut-off (Kohavi, Tang, and Xu, 2020) [16].

3.5. Cost Model

We define the expected total cost per request of a single-threshold policy t as the sum of unprevented scam losses, review costs, and friction costs to legitimate customers

$$C(t) = \Sigma_i \left[p_i L_i \left(1 - \pi \cdot 1\{s_i \geq t\}\right) + 1\{s_i \geq t\} \left(r + (1 - p_i) f(d_i) \right) \right] \quad (2)$$

where p_i is the calibrated scam probability, L_i the potential loss, π the interdiction probability given escalation, r the review cost, and $f(d_i)$ the friction cost to a legitimate customer delayed by d_i hours. Friction combines the value of customer time, the cost of abandoned transactions, and lost account revenue from attrition. Several components lack market prices. We include a value of customer time of 30 dollars per hour, roughly in line with values of time used in transport appraisal, and test sensitivity over a wide range. For the two-tier policy, requests with scores between t_1 and t_2 receive a brief structured call, and those above t_2 receive full review with a possible hold. Uncertainty in optimal thresholds is assessed by bootstrap (Efron and Tibshirani, 1993) [7].

4. Results

4.1. Score Distribution and Calibration

Table 1: Withdrawal requests by risk-score band, owners aged 60+, January 2022 to June 2025

Score band	Requests	Share (%)	Scam-induced	Scam rate (%)	Median amount ($000)	Mean predicted (%)	Scam $ (M)
0.00–0.10	702,410	55.55	71	0.010	8.5	0.012	1.9
0.10–0.20	251,982	19.93	112	0.044	12.0	0.041	4.6
0.20–0.30	128,661	10.17	164	0.127	15.5	0.121	8.1
0.30–0.40	73,904	5.84	221	0.299	19.0	0.288	12.7
0.40–0.50	45,210	3.58	296	0.655	24.0	0.640	18.4
0.50–0.60	27,342	2.16	387	1.415	31.5	1.39	26.9
0.60–0.70	16,190	1.28	462	2.854	39.0	2.91	35.2
0.70–0.80	10,071	0.80	548	5.441	52.0	5.37	46.1
0.80–0.90	6,105	0.48	571	9.353	68.0	9.48	52.8
0.90–1.00	2,655	0.21	585	22.034	95.0	21.6	64.0
Total	1,264,530	100.00	3,417	0.270	11.0	0.268	270.7

Note. Scam $ is the total requested amount of confirmed scam-induced requests. Mean predicted is the average calibrated probability underlying the score in each band.

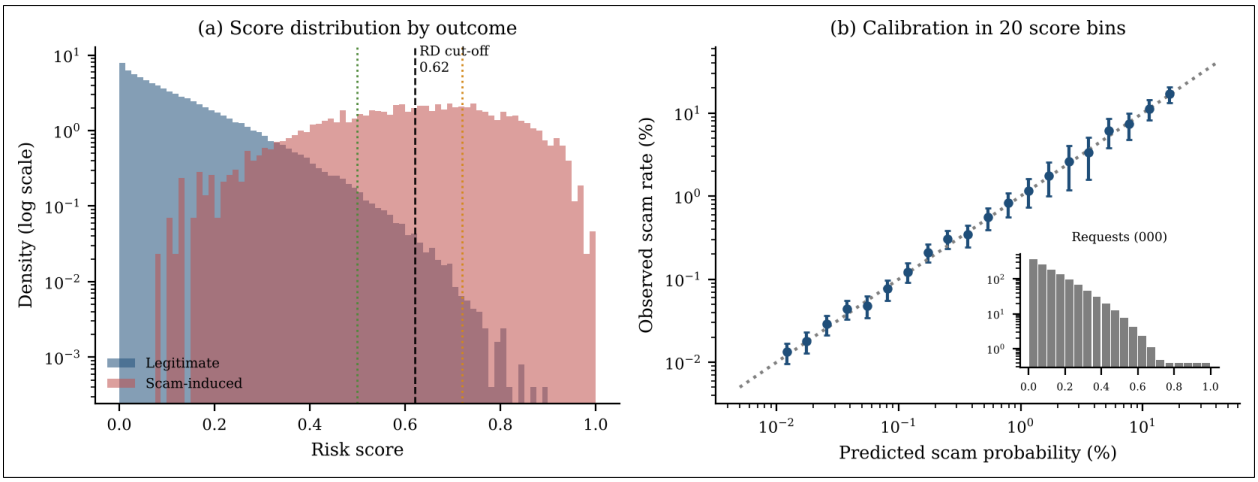


Fig 1: Distribution and calibration of the custodian's risk score. (a) Score densities for legitimate and scam-induced requests on a log scale, with the historical cut-off (dashed) and the two experimental alternatives (dotted). (b) Observed against predicted scam rates in 20 equal-width score bins with 95 percent intervals, and request counts per bin in the inset. The score is well calibrated across four orders of magnitude, which is the precondition for cost-based threshold selection.

4.2. Effects of Escalation at the Cut-Off

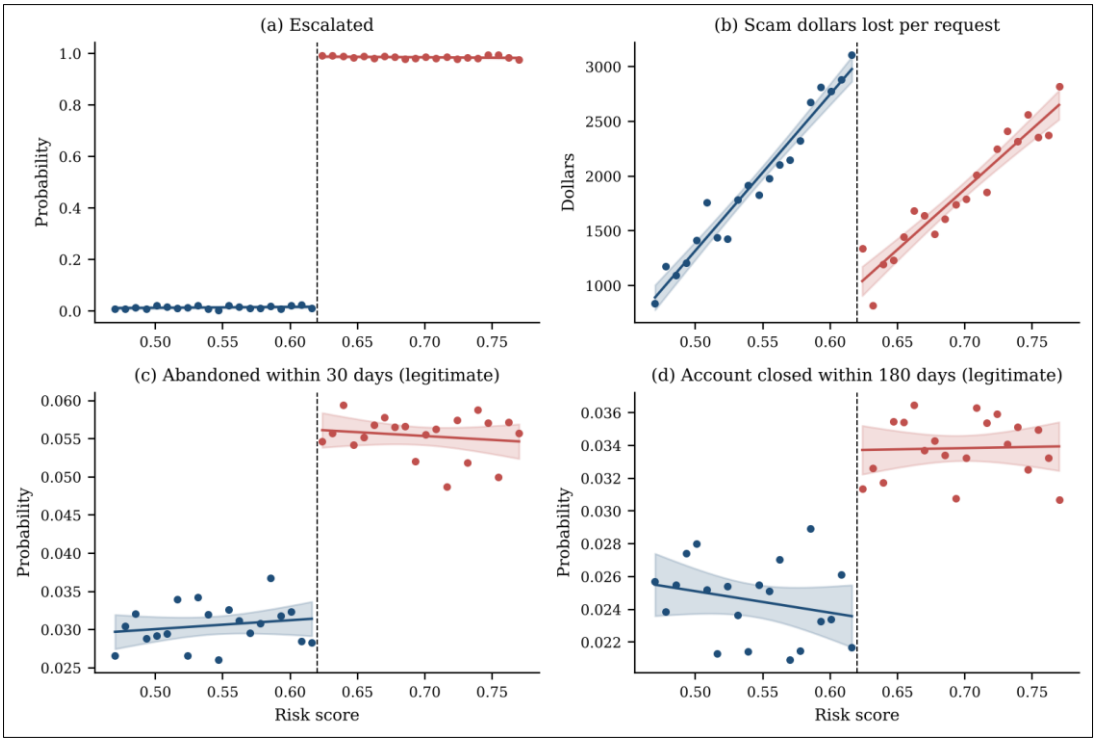


Fig 2: Regression discontinuity plots at the 0.62 escalation threshold, January 2022 to June 2024. Points are means within 40 evenly spaced bins, and lines are local linear fits with 95 percent bands estimated separately on each side. (a) Escalation jumps from near zero to near one. (b) Expected scam dollars lost per request fall sharply at the threshold. (c, d) Among legitimate requests, abandonment and later account closure rise discontinuously.

Table 2: Regression discontinuity estimates at the 0.62 threshold

Outcome	Below-cutoff mean	RD estimate	Robust 95% CI	Bandwidth (h)	Eff. n (L / R)
Escalated	0.015	0.968	0.951 to 0.982	0.081	11,204 / 8,893
Held under Rule 2165	0.002	0.071	0.059 to 0.084	0.079	10,952 / 8,710
Scam dollars lost per request, $	2,340	−1,840	−2,615 to −1,021	0.094	13,287 / 10,166
Scam requests interdicted, share	0.091	0.584	0.462 to 0.701	0.112	15,961 / 11,808
Legitimate requests only					
Hours to disbursement	26.4	31.8	28.9 to 34.6	0.074	10,231 / 8,122
Abandoned within 30 days	0.034	0.021	0.012 to 0.030	0.088	12,348 / 9,604
Complaint filed	0.004	0.006	0.003 to 0.009	0.096	13,522 / 10,431
Account closed within 180 days	0.025	0.009	0.002 to 0.016	0.102	14,416 / 10,993
Net assets moved out within 180 days, $000	4.8	2.6	0.7 to 4.5	0.099	14,011 / 10,780
Validity checks					
Density test (Cattaneo–Jansson–Ma), T	—	0.71	p = 0.48	—	—
Owner age (placebo)	73.6	0.18	−0.31 to 0.67	0.085	11,877 / 9,215
Prior-year withdrawals (placebo)	1.92	−0.03	−0.11 to 0.05	0.087	12,104 / 9,388

Note. Local linear regression, triangular kernel, MSE-optimal bandwidth, bias-corrected robust inference (Calonico, Cattaneo, and Titiunik, 2014). Standard errors clustered by owner.

Crossing the threshold almost deterministically triggered escalation, and escalation reduced expected scam losses by about 1,840 dollars per request at the margin (Table 2). Among legitimate customers, escalation added roughly 32 hours to disbursement, raised the probability of abandoning the withdrawal within 30 days by 2.1 percentage points, and raised the probability of closing the account within 180 days by 0.9 percentage points. Placebo outcomes showed no discontinuity, and the density of scores was smooth across the cut-off.

4.3. Randomized Threshold Experiment

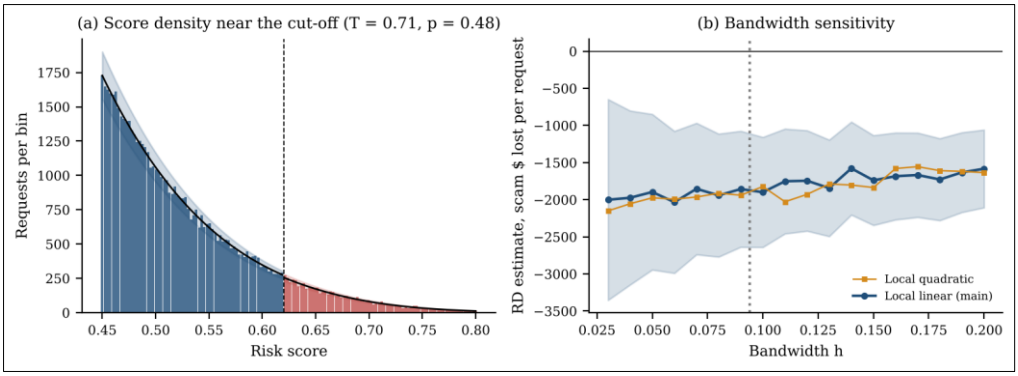


Fig 3: Validity and robustness of the regression discontinuity. (a) Histogram of scores in narrow bins near the cut-off with local polynomial density fits and bands, showing no bunching. (b) Estimates of the effect on scam dollars lost per request across bandwidths from 0.03 to 0.20 for local linear and local quadratic specifications, with the MSE-optimal bandwidth marked.

Table 3: Outcomes by randomized escalation threshold, October 2024 to March 2025

Outcome	t = 0.50	t = 0.62 (control)	t = 0.72	Δ (0.50 vs 0.62)	Δ (0.72 vs 0.62)
Requests (owners)	104,377 (52,904)	104,211 (52,860)	104,316 (52,911)	—	—
Scam-induced requests	291	284	296	—	—
Escalation rate, %	4.12	2.61	1.64	+1.51***	−0.97***
Scam requests escalated, %	79.4	68.3	52.7	+11.1***	−15.6***
Scam $ prevented, % of requested	61.8	50.6	38.9	+11.2***	−11.7***
Net scam loss per 1,000 requests, $000	71.2	91.6	117.4	−20.4***	+25.8***
Legit. escalated, %	3.92	2.43	1.49	+1.49***	−0.94***
Legit. abandoned (30 d), %	3.71	3.62	3.55	+0.09	−0.07
Legit. complaints per 1,000	4.9	4.1	3.6	+0.8*	−0.5
Legit. closed account (180 d), %	2.68	2.60	2.55	+0.08	−0.05
Review hours per 1,000 requests	37.8	24.1	15.3	+13.7***	−8.8***
Median review queue wait, hours	9.6	4.2	2.8	+5.4***	−1.4***

Note. Intention-to-treat estimates. Scam-dollar outcomes use randomization inference with 5,000 owner-level permutations because dollar losses are heavy-tailed. *p < 0.05,* **p < 0.00**

The experiment confirms that thresholds matter on both sides (Table 3). Moving from 0.62 to 0.50 raised escalation by 58 percent and reduced net scam losses by 22 percent. Because most additional escalations fell on legitimate customers with low abandonment risk, population-level friction rose only slightly, but queue waits more than doubled. Moving to 0.72 had the mirror-image effect. Note that average friction outcomes across all requests are diluted, since most requests are never escalated under any arm.

4.4. Cost Parameters and the Optimal Single Threshold

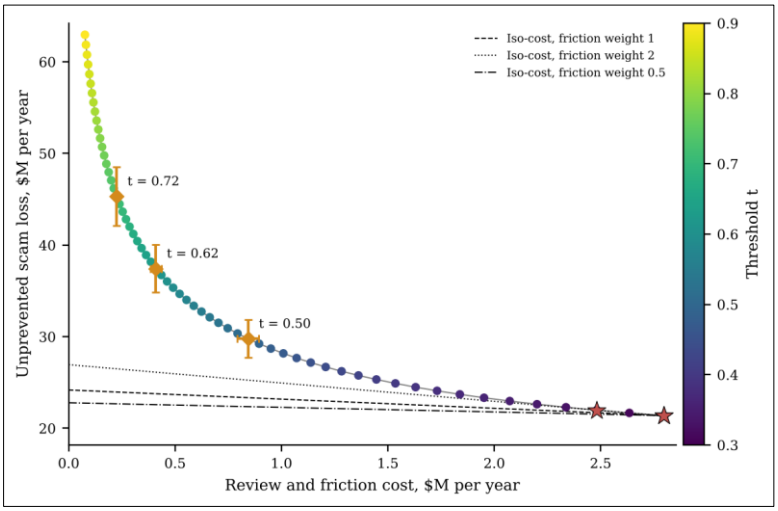


Fig 4: Efficient frontier between unprevented scam losses and the combined review and friction costs of escalation, annualized to the custodian's request volume. Each point is a single threshold coloured by its value. Diamonds show the three experimental arms with 95 percent intervals. Lines are iso-cost contours for three relative weights on friction, and stars mark the corresponding cost-minimizing thresholds.

Table 4: Cost parameters used in the decision model

Parameter	Base value	Range in sensitivity	Source
Interdiction probability given escalation (π)	0.58	0.46–0.70	RD estimate, Table 2
Mean loss given unprevented scam, $	64,900	48,000–82,000	Administrative data
Review cost per escalation, $	41	25–70	Activity-based costing
Delay per escalation, hours	31.8	20–48	RD estimate, Table 2
Value of customer time, $ per hour of delay	30	0–60	Assumed
Abandonment cost per abandoned transaction, $	320	100–800	Assumed
Abandonment rate given escalation (legit.)	0.021	0.012–0.030	RD estimate, Table 2
Present value of lost account revenue, $	5,800	3,000–9,000	Firm revenue data
Attrition rate given escalation (legit.)	0.009	0.002–0.016	RD estimate, Table 2
Implied friction cost per legit. escalation, $	1,013	260–2,350	Derived

Note. Parameters labelled Assumed have no direct empirical counterpart in the custodian's records and serve as anchors for sensitivity analysis. The implied friction cost combines delay, abandonment, and attrition components.

With these parameters, the expected-cost-minimizing single threshold is 0.54, below the historical value of 0.62. The intuition is straightforward. Expected losses avoided at the margin, roughly the product of the scam probability, the mean loss, and the interdiction probability, exceed the combined review and friction costs until the scam probability falls below about 1.9 percent. Because the friction cost is dominated by attrition, which is estimated imprecisely, the optimum is sensitive to that parameter.

4.5. Two-Tier Escalation Policies

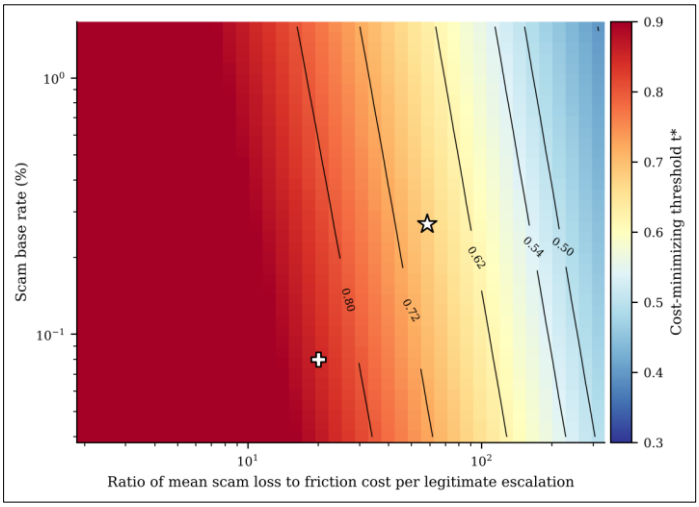


Fig 5: Cost-minimizing single threshold as a function of the loss-to-friction cost ratio and the scam base rate, holding the score distribution fixed. Contours mark selected thresholds. The white star marks the custodian studied here. The white cross marks a firm with a lower base rate and costlier friction, for which the optimal threshold is well above the historical 0.62.

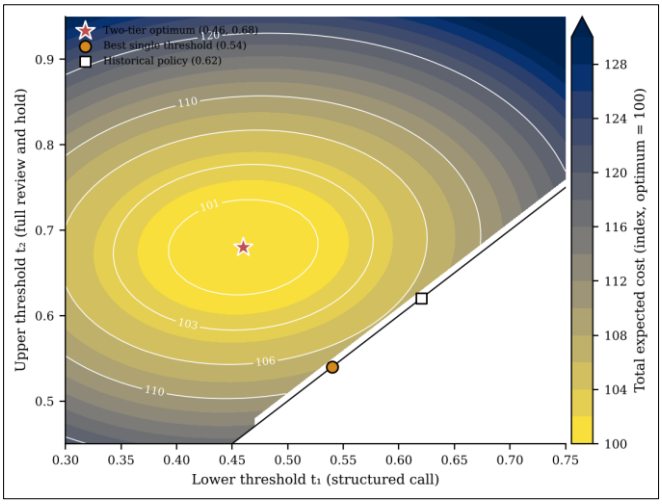


Fig 6: Expected total cost of two-tier policies indexed to the optimum. Requests scoring between t_1 and t_2 receive a short structured call costing 12 dollars and about four hours of delay, while those above t_2 receive full review. The diagonal corresponds to single-threshold policies. The two-tier optimum lowers expected cost by 17 percent relative to the best single threshold and by 26 percent relative to the historical policy.

Table 5: Cost-minimizing thresholds by segment (two-tier policy)

Segment	Share of requests (%)	Scam rate (%)	t_1*	t_2*	90% bootstrap interval (t_2*)	Cost saving vs. historical (%)
All requests	100.0	0.27	0.46	0.68	0.63–0.74	26.1
Amount < $10,000	47.8	0.06	0.63	0.83	0.75–0.90	8.4
Amount $10,000–$50,000	38.1	0.29	0.47	0.69	0.63–0.76	24.7
Amount > $50,000	14.1	0.95	0.34	0.57	0.51–0.63	33.9
Phone channel	36.4	0.48	0.41	0.63	0.57–0.69	29.5
Online channel	63.6	0.15	0.52	0.73	0.66–0.80	18.2
Owner age 60–69	41.2	0.17	0.51	0.72	0.65–0.79	19.6
Owner age 70–79	39.5	0.29	0.46	0.67	0.61–0.74	25.8
Owner age 80+	19.3	0.46	0.42	0.65	0.58–0.72	30.3
Trusted contact on file	31.8	0.21	0.49	0.66	0.60–0.73	27.7
No trusted contact	68.2	0.30	0.45	0.70	0.64–0.77	24.9

Note. Segment-specific thresholds re-optimize equation (2) using segment-specific score distributions, loss amounts, and friction estimates. Bootstrap intervals from 1,000 owner-level resamples. Age-specific thresholds are reported for analysis only and raise legal and ethical questions discussed in Section 5.

4.6. Capacity and Queue Dynamics

Figure 7 shows why thresholds cannot be set without reference to capacity. Under the 0.50 threshold, review demand exceeded staffed capacity on 41 percent of days in the experiment window and the backlog never cleared during the surge. Because friction costs grow with delay, a policy that is optimal at nominal staffing can become costly when the queue grows (Gross *et al.*, 2008) [12]. The two-tier policy handles the surge better because the structured call consumes about one-quarter of the review time of a full case.

4.7. Friction as a Function of Delay

Table 6: Legitimate-customer outcomes by realized delay among escalated requests

Delay to disbursement	Requests	Abandoned 30 d (%)	Complaint (%)	Closed 180 d (%)	Assets out 180 d ($000)
< 4 hours	6,812	0.9	0.21	1.8	2.1
4–24 hours	11,947	1.6	0.38	2.4	3.6
1–2 days	9,315	3.4	0.72	3.1	6.2
2–5 days	5,622	6.8	1.46	4.6	10.9
5–10 days	1,741	11.2	2.83	6.9	17.4
> 10 days (Rule 2165 hold)	498	18.7	6.02	11.4	29.8

Note. Delay is endogenous to case complexity, so these are descriptive associations. The dose-response model in Figure 8 adjusts for score, amount, age, channel, and month.

4.8. Equity

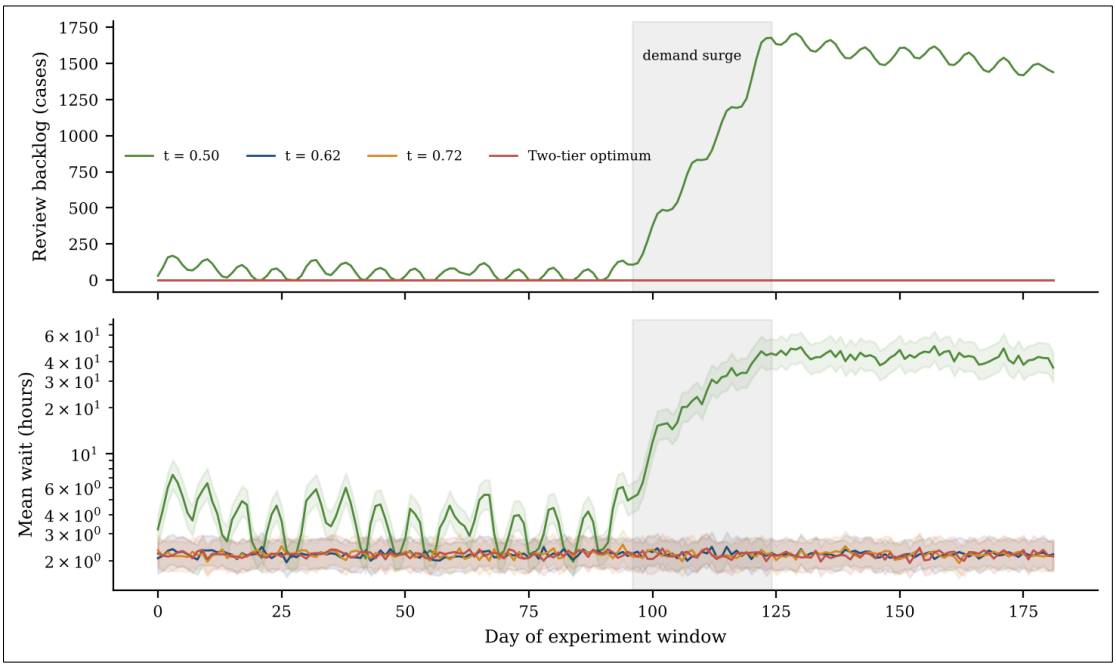


Fig 7: Simulated review backlog and mean wait under each policy using observed daily arrivals and the review team's staffed capacity, with a shaded window representing a 20 percent surge in withdrawal requests of the kind seen during market sell-offs. The aggressive 0.50 threshold exceeds capacity during the surge and wait times rise roughly tenfold, which would erode the friction estimates on which its benefits depend.

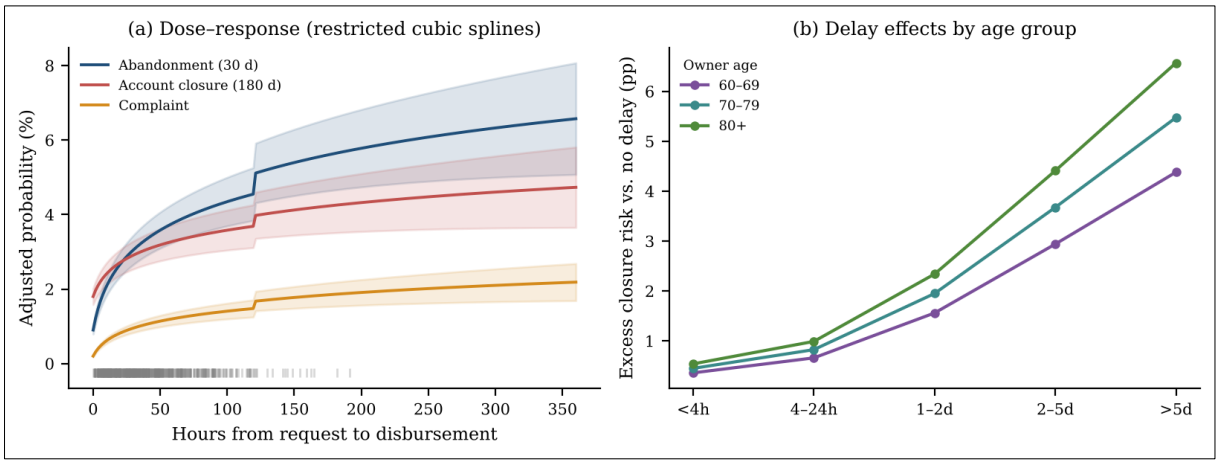


Fig 8: Adjusted relationship between delay and legitimate-customer outcomes. (a) Restricted cubic spline estimates with 95 percent bands and a rug showing the distribution of observed delays. Costs rise steeply over the first two days and continue to rise for long holds. (b) Excess 180-day account closure risk by delay category and age group. Older owners appear more sensitive to long delays.

Table 7: Escalation and error rates at the two-tier optimum by owner characteristics

Group	Escalation rate (%)	Scam rate (%)	TPR	FPR (%)	Ratio of FPR to overall
Overall	2.94	0.27	0.77	2.74	1.00
ZIP median income quintile					
Q1 (lowest)	3.08	0.29	0.76	2.86	1.04
Q2	2.99	0.28	0.77	2.79	1.02
Q3	2.91	0.27	0.78	2.71	0.99
Q4	2.88	0.26	0.77	2.69	0.98
Q5 (highest)	2.86	0.26	0.77	2.66	0.97
Sex					
Female	3.11	0.31	0.78	2.87	1.05
Male	2.75	0.23	0.75	2.59	0.95
Rurality					
Urban/suburban	2.92	0.26	0.77	2.73	1.00
Rural	3.02	0.30	0.76	2.79	1.02

Note. TPR = share of scam-induced requests escalated at either tier. FPR = share of legitimate requests escalated at either tier. Neighbourhood income is the ZIP Code Tabulation Area median household income from the American Community Survey.

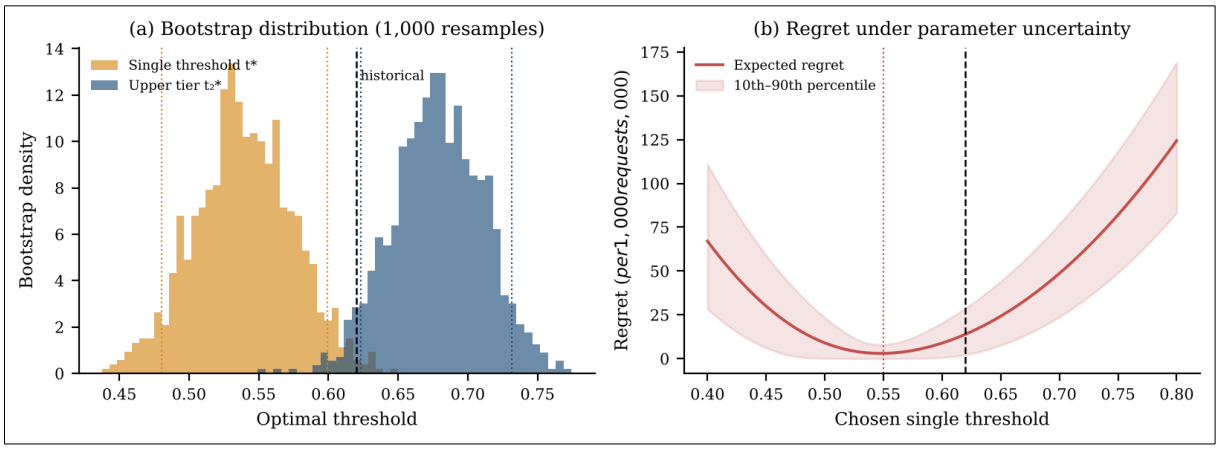


Fig 9: Uncertainty in optimal thresholds. (a) Bootstrap distributions of the optimal single threshold and of the upper tier of the two-tier policy, with 90 percent intervals (dotted) and the historical threshold (dashed). (b) Expected regret of committing to each single threshold, averaged over bootstrap draws of the cost parameters and score distribution. Regret is asymmetric, because setting the threshold too high forgoes large scam losses, so the regret-minimizing choice sits slightly below the point estimate.

5. Discussion

Escalation thresholds embody a judgment about how much friction is worth imposing on legitimate customers to stop a scam. By measuring both sides of that judgment with credible designs, we find that the custodian's historical threshold was too conservative for large and phone-channel requests and somewhat too aggressive for small online requests. A two-tier policy that uses a light-touch call for moderate-risk requests dominates any single threshold, because it buys much of the protective value of escalation at a fraction of the delay.

Three caveats shape how the results should be used. First, friction costs, especially attrition, are the least precisely measured inputs, and the optimal threshold moves materially

across plausible values. The regret analysis suggests erring toward lower thresholds when uncertain, because the cost of a missed scam is concentrated and severe while friction is diffuse. Second, optimal thresholds depend on capacity, and a policy that looks optimal at full staffing can fail during demand surges. Firms should pair threshold choices with surge plans, such as temporarily routing more requests to the structured-call tier. Third, segment-specific thresholds based on age would likely raise fair-lending and age-discrimination concerns, even when the underlying risk differs. Thresholds based on amount and channel are easier to defend and capture most of the gain.

Applied to a mid-sized custodian with half the volume studied here and a similar scam mix, the two-tier policy would reduce expected annual losses and friction costs by several million dollars. The equity analysis is reassuring. Escalation rates and error rates at the optimum differed little across neighbourhood income quintiles, and the modestly higher escalation of women tracked a higher scam base rate.

5.1. Limitations

The RD identifies effects only at the 0.62 cut-off, and extrapolation relies on the experiment and on the cost model. The experiment ran for 24 weeks, which may be too short to capture longer-run attrition. Scam labels depend on detection, and escalation itself may uncover scams that would otherwise have gone unlabelled, inflating measured losses below the threshold. Delay effects in Table 6 are associational. The queue simulation uses observed arrivals and staffing but treats a surge as exogenous, and real surges may also change the scam mix. Finally, results come from a single custodian with its own risk model, and the optimal threshold on a different score would differ even if the logic carries over.

5.2. Recommendations

- Set escalation thresholds through an explicit cost model rather than by alert volume, and document the assumed costs of friction.
- Prefer tiered responses, using brief structured calls for moderate-risk requests and full review only for the highest risk.
- Vary thresholds by request amount and channel, where the evidence for different trade-offs is strongest.
- Plan for capacity, and monitor queue waits as a leading indicator that friction costs are rising.
- Audit escalation and error rates by income, sex, and rurality whenever thresholds change.

6. Conclusion

Choosing an escalation threshold for retirement withdrawals is a policy decision, not a modelling detail. Credible estimates of the losses that escalation prevents and the friction it creates make that decision tractable. In our setting, a lower and tiered set of thresholds would have stopped substantially more scam losses at little additional cost to legitimate customers, provided review capacity keeps pace.

Declarations

Funding. Supported by the Center for Retirement Security Research.

Conflicts of interest. L.B.W. previously consulted for a financial services trade association on unrelated matters.

Data availability. Request-level data are confidential. Replication code, the cost model, and a parameter file that reproduces Figures 5 and 6 are available from the corresponding author.

References

1. Bahnsen AC, Aouada D, Ottersten B. Example-dependent cost-sensitive logistic regression for credit scoring. In: Proceedings of the 13th International Conference on Machine Learning and Applications. IEEE; 2014. p. 263-269.
2. Burnes D, Henderson CR, Sheppard C, Zhao R, Pillemer K, Lachs MS. Prevalence of financial fraud and scams among older adults in the United States: A systematic review and meta-analysis. American Journal of Public Health. 2017;107(8):e13-e21.
3. Calonico S, Cattaneo MD, Titiunik R. Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica. 2014;82(6):2295-2326.
4. Cattaneo MD, Jansson M, Ma X. Simple local polynomial density estimators. Journal of the American Statistical Association. 2020;115(531):1449-1455.
5. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data. 2017;5(2):153-163.
6. Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G. Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence; 2015. p. 159-166.
7. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall; 1993.
8. Elkan C. The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence; 2001. p. 973-978.
9. Federal Bureau of Investigation. Elder fraud report 2023. Internet Crime Complaint Center; 2024.
10. FINRA. Rule 2165. Financial exploitation of specified adults. Financial Industry Regulatory Authority; 2018.
11. Gelman A, Imbens G. Why high-order polynomials should not be used in regression discontinuity designs. Journal of Business & Economic Statistics. 2019;37(3):447-456.
12. Gross D, Shortle JF, Thompson JM, Harris CM. Fundamentals of queueing theory (4th ed.). Wiley; 2008.
13. Hand DJ. Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning. 2009;77(1):103-123.
14. Imbens GW, Lemieux T. Regression discontinuity designs: A guide to practice. Journal of Econometrics. 2008;142(2):615-635.
15. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (pp. 43:1–43:23).
16. Kohavi R, Tang D, Xu Y. Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press; 2020.
17. Lee DS, Lemieux T. Regression discontinuity designs in economics. Journal of Economic Literature. 2010;48(2):281-355.
18. Provost F, Fawcett T. Robust classification for imprecise environments. Machine Learning. 2001;42(3):203-231.
19. Sunstein CR. Sludge: What stops us from getting things done and what to do about it. MIT Press; 2021.

20. Thaler RH, Sunstein CR. Nudge: Improving decisions about health, wealth, and happiness. Yale University Press; 2008.
21. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565-574.
22. Whitrow C, Hand DJ, Juszczak P, Weston D, Adams NM. Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery. 2009;18(1):30-55.
23. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-35.

How to Cite This Article
Fagbore OO, Whitcombe LB, Moriyama KR, Ellery GN. Cost-Sensitive calibration of escalation thresholds for retirement account withdrawals and the Trade-Off between fraud loss and customer friction. Int J Multidiscip Futur Dev. 2026;7(2):139-148. doi:10.54660/IJMFD.2026.7.2.139-148

Creative Commons (CC) License
This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.