World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

# STRUCTURED BEHAVIORAL INTERVIEWS AS SECOND-LINE CONTROL AGAINST AUTHORIZED-PAYMENT SCAMS IN RETIREMENT WITHDRAWALS, WITH EVIDENCE FROM A STAGGERED BRANCH ROLLOUT

Adeola Ibitogbe [1,*], Olasunbo O. Fagbore [2], Robert K. Ansah [3], Hannah J. Lindqvist [3] and Marcus D. Pell [3]

[1] The Annie E. Casey Foundation Baltimore, USA.
[2] Senior Brokerage Consultant II, Fidelity Investments, Westlake, TX, USA.
[3] Eastbrook Financial Company, Lincoln, United Kingdom.
* Corresponding Author

World Journal of Advanced Research and Reviews, 2025, 28(02), 2680–2693

Publication history: Received on 11 October 2025; revised on 16 November 2025; accepted on 18 November 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.28.2.3890

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

## ABSTRACT

Identity verification confirms that the person requesting a retirement withdrawal is the account owner, but it cannot tell whether the owner is acting under a scammer's direction. We evaluate a 12-item structured behavioral interview (SBI) that replaced a knowledge-based authentication and callback procedure for retirement withdrawals escalated for second-line review at a national brokerage and IRA custodian. The SBI was introduced to 41 service units in five cohorts between March 2023 and March 2024, which permits a staggered difference-in-differences design. Across 31,208 escalated requests from owners aged 60 and older between January 2022 and December 2024, of which 2,104 were confirmed scam-induced, adoption of the SBI raised the share of scam-induced requests interdicted before payout by 18.4 percentage points (95% CI 12.9 to 23.9) from a pre-period mean of 41.2 percent, and lowered net loss per scam-induced request by about 21,700 dollars. Effects on false holds of legitimate requests, abandonment, and complaints were small and statistically indistinguishable from zero, although average handle time rose by 6.4 minutes. A composite interview score discriminated scam-induced from legitimate requests with an AUROC of 0.86, compared with 0.58 for authentication outcomes. Items probing secrecy instructions, unsolicited first contact, and narrative coherence carried the most information, and inter-rater agreement on recorded calls was substantial. A survey of 2,940 customers found that most considered the questions appropriate, including a majority of those whose withdrawals were delayed. Structured interviewing appears to be a scalable, low-friction complement to identity verification for authorized-payment scams.

**Keywords.** Elder Financial Exploitation, Authorized Push Payment Fraud, Structured Interview, Difference-In-Differences, Retirement Accounts, Cognitive Interviewing

# 1 INTRODUCTION

Financial institutions have spent two decades hardening identity verification. Knowledge-based authentication, one-time passcodes, device fingerprinting, and voice biometrics now make it difficult for an impostor to withdraw money from someone else's account. Scammers have adapted by changing who presses the button. In romance, impersonation, and so-called phantom hacker schemes, the scammer persuades the account owner to request the transfer personally. In the United Kingdom this category, known as authorized push payment fraud, became large enough that the Payment Systems Regulator introduced mandatory reimbursement in 2024 (Payment Systems Regulator, 2023). In the United States, complaints from people aged 60 and older to the FBI reported losses of roughly 3.4 billion dollars in 2023 (Federal Bureau of Investigation, 2024), much of it voluntarily transferred by victims who passed every identity check.

Retirement accounts are a prime target because they hold large balances and can be liquidated quickly. When a withdrawal looks unusual, firms typically escalate it for a callback. The traditional callback, however, is designed to confirm identity rather than intent. A representative asks for the owner's date of birth and the last four digits of a linked account, confirms the amount, and releases the funds. A coached victim passes easily. Consider an exchange in which a caller explains, calmly and correctly, that she is moving her savings to a new account because a federal agent warned her that her identity had been stolen. A standard callback has no question that would surface the federal agent at all.

This paper evaluates an alternative. A structured behavioral interview (SBI) replaces identity questions with a short, scripted sequence of open and closed questions designed to surface the hallmarks of scam coaching, such as secrecy instructions, unsolicited first contact, urgency, and pressure to use unusual payment rails. The approach draws on research showing that cognitive and information-gathering interview techniques outperform intuitive judgment for detecting deception (Fisher and Geiselman, 1992; Vrij, Fisher, and Blank, 2017), adapted to a setting in which the interviewee is not lying but has been deceived.

We exploit the staggered rollout of the SBI across 41 service units of a national custodian to estimate its causal effect on scam interdiction, losses, and customer friction. We complement the quasi-experimental estimates with item-level psychometric analysis, an inter-rater reliability study using recorded calls, an analysis of interviewer heterogeneity, and a customer experience survey. Together these components answer three questions. Does the SBI stop more scams? Which questions do the work? And what does it cost legitimate customers?

# 2 BACKGROUND

## 2.1 Why scam victims pass identity checks

Scam victimization in later life reflects a combination of persuasion techniques and individual vulnerability. Scammers rely on authority, scarcity, reciprocity, and commitment cues (Cialdini, 2001), and often build rapport over weeks before any request for money (Langenderfer and Shimp, 2001). Victims of investment fraud display distinctive mindsets and behaviours, including openness to unsolicited offers (DeLiema, Shadel, and Pak, 2020). Population estimates suggest that around 5.6 percent of older Americans experience fraud each year (Burnes et al., 2017). Psychological vulnerability, social isolation, and declining decision-making capacity raise risk (Lichtenberg, Stickney, and Paulson, 2013; Hanoch and Wood, 2021). None of these factors interferes with the ability to answer authentication questions correctly.

## 2.2 Interviewing for deception and influence

Laboratory and field research shows that people are poor intuitive lie detectors and tend to default to believing what they are told (Levine, 2014). Information-gathering interviews that elicit detailed narratives, ask unexpected questions, and test for verifiable details perform better (Vrij, Fisher, and Blank, 2017). The cognitive interview, developed for eliciting accurate memory from cooperative witnesses, emphasizes open-ended prompts and rapport (Fisher and Geiselman, 1992). Structured screening instruments are already used to assess financial decision-making capacity in older adults, notably the Lichtenberg Financial Decision Rating Scale (Lichtenberg et al., 2015). The SBI evaluated here borrows from both traditions. It treats the customer as a cooperative witness whose account of the transaction may reveal the hidden hand of a third party.

## 2.3 Regulatory context

FINRA Rule 2165 permits firms to place temporary holds on disbursements when exploitation of a specified adult is reasonably suspected, and Rule 4512 requires reasonable efforts to obtain a trusted contact (FINRA, 2018a, 2018b). The Senior Safe Act of 2018 protects trained employees who report suspected exploitation in good faith. Suspicious activity reports on elder financial exploitation have risen sharply (Consumer Financial Protection Bureau, 2019). These rules create the authority to intervene but say little about how frontline staff should decide when to do so.

## 3 INTERVENTION, DATA, AND DESIGN

### 3.1 The structured behavioral interview

The SBI consists of 12 items delivered in a fixed order after identity verification (Table 3). Eight items are closed questions with yes, no, or unsure responses. Two items invite a short narrative, which the representative rates for coherence and independence on a three-point anchored scale. One item asks whether anyone is present or listening, and one asks permission to contact the trusted contact. Items were developed by a working group of fraud investigators, compliance officers, and two of the authors, piloted with scripts voiced by trained actors, and refined over three iterations. Representatives received four hours of training that included role play and a certification call. A score from 0 to 18 was computed by summing weighted item responses, and the protocol recommended a hold and referral when the score reached 6 or more or when any of items 3, 5, or 11 was endorsed.

### 3.2 Comparator procedure

Before adoption, escalated requests were handled through a standard callback consisting of knowledge-based authentication, confirmation of amount and destination, and a scripted fraud warning. Representatives could place a hold at their discretion.

### 3.3 Rollout and data

Operational capacity for training determined the rollout schedule. The 41 service units (9 contact-centre pods and 32 regional branch clusters) adopted the SBI in five cohorts in March 2023, June 2023, September 2023, December 2023, and March 2024. Cohort assignment was made by the operations team on the basis of trainer availability and was not related to prior fraud rates, which we verify below. Our sample comprises every retirement withdrawal request from an owner aged 60 or older that the custodian's automated monitoring escalated for second-line review between January 2022 and December 2024. Labels for scam-induced requests come from fraud investigations closed within 180 days. The final sample contains 31,208 escalated requests, 2,104 of them confirmed scam-induced.

### 3.4 Outcomes

Interdiction rate. Among scam-induced escalated requests, the share for which funds were not disbursed or were recovered within 10 business days.

- Net loss. Dollars disbursed and not recovered within 90 days, per scam-induced escalated request.
- False hold rate. Among legitimate escalated requests, the share held for more than 24 hours.
- Abandonment. Share of legitimate escalated requests withdrawn by the customer within 30 days of escalation.
- Handle time. Representative minutes on the escalation call.
- Complaints. Formal complaints mentioning the withdrawal process per 1,000 escalated requests.

### 3.5 Estimation

Because adoption was staggered, two-way fixed-effects estimators can be biased by heterogeneous treatment effects (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020). Our primary estimator is the group-time average treatment effect of Callaway and Sant'Anna (2021), using not-yet-treated units as controls and aggregating to an overall effect and to event time. For outcome $Y$ in unit $u$ adopting in period $g$, the group-time effect is

$$ATT(g, t) = E[Y_t - Y_{g-1} \mid G = g] - E[Y_t - Y_{g-1} \mid D_t = 0, G \neq g] \quad (1)$$

and the event-study coefficients average ATT(g, g+e) across cohorts. Request-level covariates (age, amount, scam typology prior, risk score decile) enter through doubly robust adjustment. Standard errors are clustered by service unit, and we report wild cluster bootstrap p values given the modest number of clusters (Cameron, Gelbach, and Miller, 2008; Bertrand, Duflo, and Mullainathan, 2004). Robustness checks include the interaction-weighted estimator of Sun and Abraham (2021), a stacked design, synthetic control for the first cohort (Abadie, Diamond, and Hainmueller, 2010), and placebo outcomes, following current recommendations (Roth et al., 2023).

### 3.6 Supplementary studies

Item analysis used the post-adoption sample of 17,546 SBI-administered requests. We estimated a two-parameter logistic item response model treating scam status as the criterion and reported item discrimination and difficulty (Embretson and Reise, 2000). For reliability, two independent quality analysts re-scored 180 randomly selected recorded SBI calls, and agreement was summarized with Cohen's κ (Cohen, 1960) and interpreted with conventional benchmarks (Landis and Koch, 1977). The customer survey was fielded by email within 14 days of the escalation call to a stratified random sample of owners and obtained 2,940 responses (response rate 31.6 percent).

# 4 RESULTS

## 4.1 Sample

**Table 1: Escalated retirement withdrawal requests before and after SBI adoption**

| Characteristic | Pre-adoption (n = 13,662) | Post-adoption (n = 17,546) | Std. diff. |
|---|---|---|---|
| Owner age, mean (SD), years | 72.8 (7.1) | 73.1 (7.0) | 0.04 |
| Female, % | 54.6 | 55.2 | 0.01 |
| Request amount, median (IQR), $000 | 38.0 (15.0–92.0) | 40.5 (16.0–96.0) | 0.03 |
| Model risk score, mean (SD) | 0.31 (0.22) | 0.32 (0.22) | 0.05 |
| Channel, phone % | 71.4 | 68.9 | 0.05 |
| Trusted contact on file, % | 29.7 | 36.8 | 0.15 |
| Confirmed scam-induced, n (%) | 921 (6.7) | 1,183 (6.7) | 0.00 |
| Scam typology among confirmed, % | | | |
| Romance / relationship | 27.1 | 24.5 | 0.06 |
| Government / bank impersonation | 23.9 | 25.8 | 0.04 |
| Tech support / phantom hacker | 17.3 | 21.4 | 0.10 |
| Investment / cryptocurrency | 17.6 | 17.9 | 0.01 |
| Family emergency | 8.8 | 6.9 | 0.07 |
| Other | 5.3 | 3.5 | 0.09 |

*Note.* Pre and post refer to unit-months before and after each unit's adoption date. Std. diff. = absolute standardized difference.

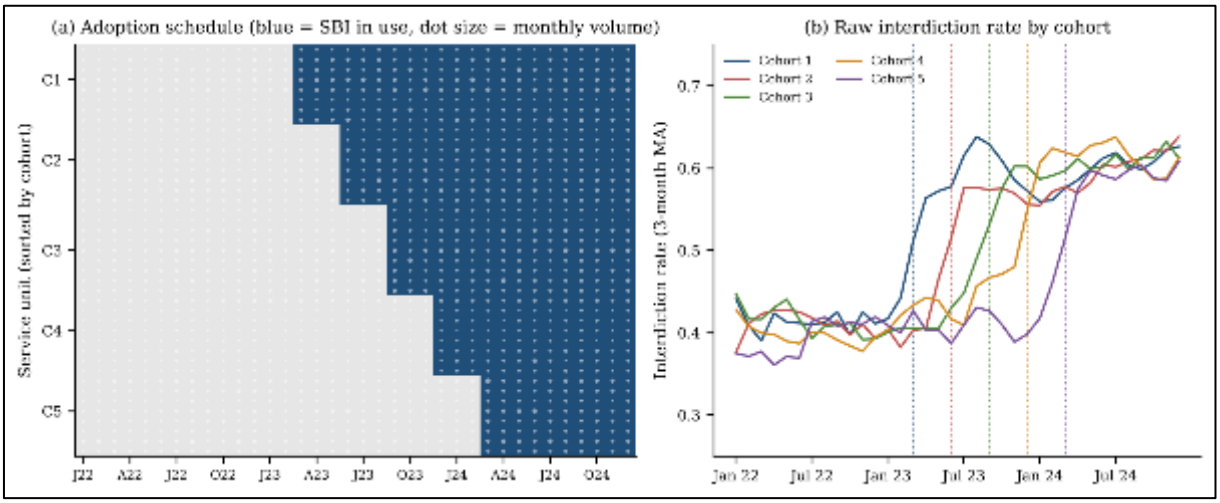



**Figure 1: Staggered adoption of the structured behavioral interview. (a) Treatment status of each of 41 service units by month, grouped into five adoption cohorts, with dot size proportional to escalated volume. (b) Three-month moving average of the interdiction rate among scam-induced escalated requests by cohort, with dotted lines at each cohort's adoption month.**

### 4.2 Main effects

**Table 2: Effect of SBI adoption on primary and secondary outcomes**

| Outcome | Pre-period mean | ATT (SE) | 95% CI | Wild bootstrap p | Relative change |
|---|---|---|---|---|---|
| Scam-induced requests (n = 2,104) | | | | | |
| Interdiction rate, % | 41.2 | 18.4 (2.8) | 12.9 to 23.9 | <0.001 | +44.7% |
| Net loss per request, $000 | 47.9 | −21.7 (5.1) | −31.7 to −11.7 | <0.001 | −45.3% |
| Days from escalation to referral | 6.8 | −3.1 (0.9) | −4.9 to −1.3 | 0.002 | −45.6% |
| Adult protective services referral, % | 22.4 | 9.6 (3.0) | 3.7 to 15.5 | 0.004 | +42.9% |
| Legitimate requests (n = 29,104) | | | | | |
| False hold rate (> 24 h), % | 4.3 | 1.1 (0.8) | −0.5 to 2.7 | 0.186 | +25.6% |
| Abandonment within 30 days, % | 3.6 | 0.8 (0.7) | −0.6 to 2.2 | 0.271 | +22.2% |
| Handle time, minutes | 11.9 | 6.4 (0.6) | 5.2 to 7.6 | <0.001 | +53.8% |
| Complaints per 1,000 requests | 5.2 | 0.9 (1.1) | −1.3 to 3.1 | 0.428 | +17.3% |
| Account closure within 180 days, % | 2.9 | 0.2 (0.5) | −0.8 to 1.2 | 0.694 | +6.9% |

*Note.* Callaway and Sant'Anna (2021) estimator with not-yet-treated controls and doubly robust covariate adjustment. Standard errors clustered by service unit (41 clusters). Wild cluster bootstrap with 9,999 Rademacher draws. ATT = average treatment effect on the treated.

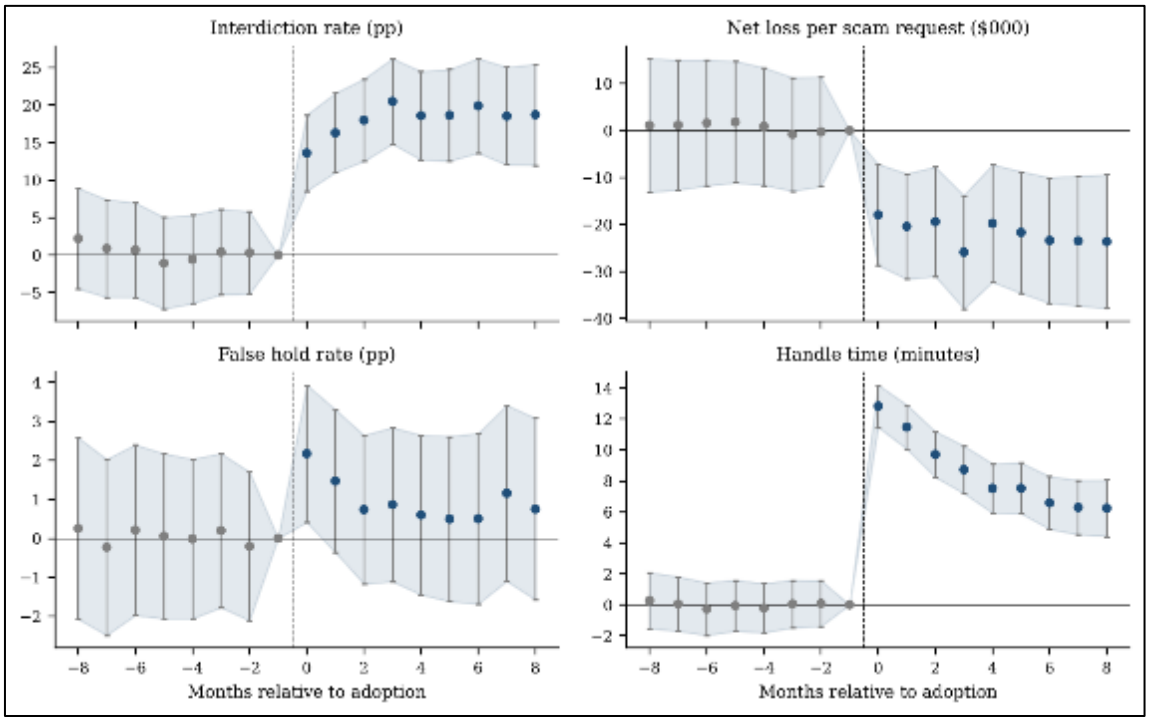



**Figure 2. Event-study estimates from the Callaway and Sant'Anna estimator for four outcomes, with 95 percent simultaneous confidence bands from the multiplier bootstrap. Month −1 is the reference period. Pre-adoption coefficients are small and jointly insignificant (joint p values 0.62, 0.48, 0.71, and 0.55), supporting parallel trends. Handle-time effects decay as representatives gain experience, while interdiction effects grow over the first six months.**

Adoption of the SBI increased the interdiction rate by 18.4 percentage points, a relative increase of about 45 percent (Table 2). Net loss per scam-induced request fell by roughly 21,700 dollars. Referrals to adult protective services rose, consistent with interviews producing documentable concerns. On the legitimate side, handle time rose by 6.4 minutes, while false holds, abandonment, complaints, and account closures moved by small amounts that were not statistically distinguishable from zero.

The event-study estimates in Figure 2 show no evidence of pre-trends, with flat and jointly insignificant coefficients in the eight months before adoption. The interdiction effect appears in the first month and grows over the following six months, consistent with representatives becoming more skilled at the narrative items. The handle-time cost was largest immediately after adoption and fell by about half within a year.

### 4.3 Which questions do the work

**Table 3: Item-level performance of the SBI among 17,546 post-adoption escalated requests**

| Item | Endorsed, scam (%) | Endorsed, legit. (%) | LR+ | LR− | IRT a | IRT b |
|---|---|---|---|---|---|---|
| Q1 Anyone asked you to do this or told you what to say? | 58.0 | 4.1 | 14.1 | 0.44 | 1.42 | 0.41 |
| Q2 First contact was unsolicited call, text, or pop-up? | 49.0 | 2.8 | 17.5 | 0.52 | 1.61 | 0.63 |
| Q3 Told to keep this confidential? | 37.0 | 0.6 | 61.7 | 0.63 | 2.08 | 1.05 |
| Q4 Destination opened recently or at another's direction? | 52.0 | 7.1 | 7.3 | 0.52 | 1.17 | 0.52 |
| Q5 Told your money is unsafe and must be moved? | 34.0 | 0.9 | 37.8 | 0.67 | 1.94 | 1.12 |
| Q6 Promised return, prize, or refund needing payment first? | 21.0 | 1.2 | 17.5 | 0.80 | 1.33 | 1.48 |
| Q7 Someone met only online or by phone involved? | 31.0 | 1.8 | 17.2 | 0.70 | 1.49 | 1.21 |
| Q8 Asked to convert to crypto, gift cards, or cash? | 26.0 | 1.5 | 17.3 | 0.75 | 1.71 | 1.36 |
| Q9 Deadline or penalty if not sent today? | 44.0 | 3.4 | 12.9 | 0.58 | 1.38 | 0.79 |
| Q10 Purpose narrative (coherence rating) | 63.0 | 8.8 | 7.2 | 0.41 | 1.86 | 0.29 |
| Q11 Anyone with you or listening now? | 19.0 | 1.1 | 17.3 | 0.82 | 1.52 | 1.54 |
| Q12 Comfortable if we contact your trusted contact? | 46.0 | 9.4 | 4.9 | 0.60 | 0.97 | 0.70 |

*Note.* Endorsement for narrative items (Q10) means a rating of low coherence or apparent third-party scripting, and for Q12 means the owner declined contact. LR = likelihood ratio. IRT parameters from a two-parameter logistic model in which a is discrimination and b is location on the latent scam-indication scale.

Secrecy instructions (Q3), being told that money is unsafe (Q5), and the narrative coherence rating (Q10) had the highest discrimination. Secrecy and safety items were rarely endorsed by legitimate customers, so a single endorsement was highly informative, with positive likelihood ratios above 35. The narrative item was endorsed more often by both groups but discriminated well across the middle of the scale. The trusted-contact item (Q12) contributed least, mainly because many legitimate owners also declined.

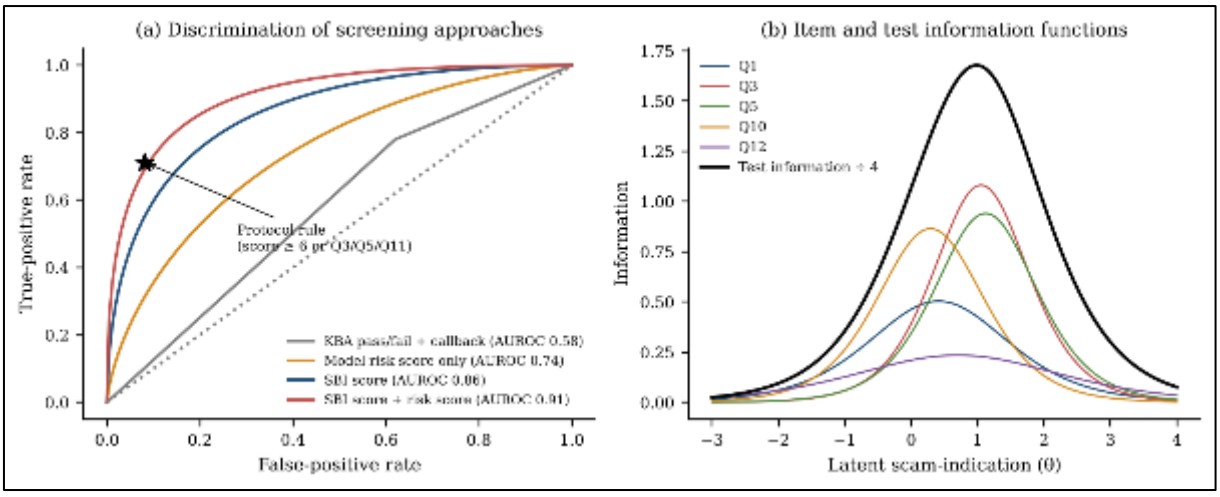



**Figure 3: (a) ROC curves comparing authentication outcomes, the automated risk score, the SBI composite score, and their combination on post-adoption requests. The star marks the operating point of the protocol's hold rule. (b) Item information functions from the two-parameter IRT model for five items and the total test information (scaled), showing that the instrument is most precise in the moderate-to-high range where hold decisions are made.**

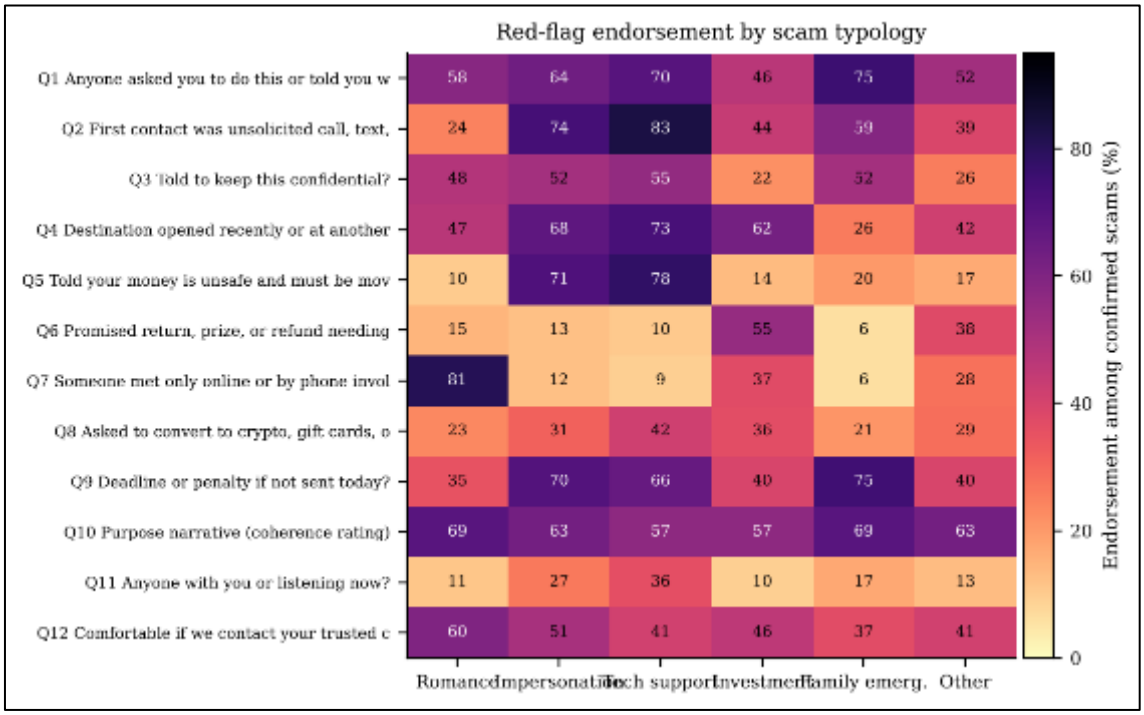



**Figure 4: Endorsement of each SBI item among confirmed scam-induced requests by scam typology (n = 1,183). Impersonation and technology-support scams were marked by secrecy and safety instructions, romance scams by an online-only relationship, and investment scams by promised returns. No single item covered every typology, which motivates a multi-item instrument.**

### 4.4 Heterogeneity

**Table 4: Heterogeneity in the effect of SBI adoption on the interdiction rate (percentage points)**

| Subgroup | n (scam-induced) | Pre-period mean (%) | ATT (SE) | 95% CI | Interaction p |
|---|---|---|---|---|---|
| All scam-induced requests | 2,104 | 42.3 | 18.4 (2.8) | 12.9 to 23.9 | — |
| Romance / relationship | 541 | 40.9 | 11.2 (5.1) | 1.2 to 21.2 | 0.012 |
| Impersonation | 524 | 44.4 | 24.6 (4.7) | 15.4 to 33.8 | |
| Tech support / phantom hacker | 412 | 38.8 | 27.9 (5.6) | 16.9 to 38.9 | |
| Investment / crypto | 374 | 34.7 | 9.8 (5.9) | -1.8 to 21.4 | |
| Family emergency | 162 | 46.3 | 21.3 (8.8) | 4.1 to 38.5 | |
| Age 60–69 | 588 | 43.7 | 15.1 (4.6) | 6.1 to 24.1 | 0.274 |
| Age 70–79 | 902 | 34.6 | 18.9 (3.9) | 11.3 to 26.5 | |
| Age 80+ | 614 | 46.0 | 22.4 (4.8) | 13.0 to 31.8 | |
| Phone channel | 1,497 | 36.6 | 20.8 (3.1) | 14.7 to 26.9 | 0.181 |
| Digital / branch channel | 607 | 33.7 | 12.9 (5.4) | 2.3 to 23.5 | |
| Risk score tercile 1 | 312 | 39.9 | 13.5 (6.2) | 1.3 to 25.7 | 0.416 |
| Risk score tercile 2 | 701 | 42.3 | 17.2 (4.4) | 8.6 to 25.8 | |
| Risk score tercile 3 | 1,091 | 39.5 | 20.6 (3.6) | 13.5 to 27.7 | |
| Interviewer tenure < 2 yr | 846 | 39.7 | 14.7 (4.1) | 6.7 to 22.7 | 0.198 |
| Interviewer tenure ≥ 2 yr | 1,258 | 47.1 | 21.5 (3.5) | 14.6 to 28.4 | |

*Note.* Interaction p values test equality of effects within each grouping and are reported on the first row of each group. Typology was assigned after investigation and is therefore conditioned on confirmation.

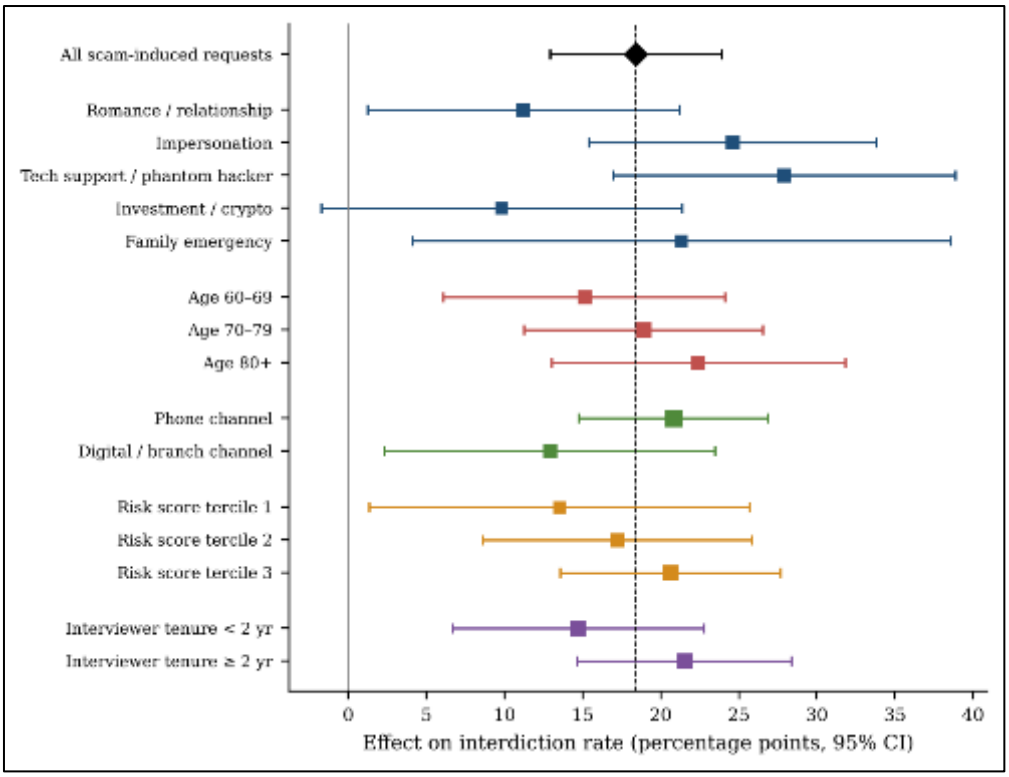



**Figure 5: Forest plot of heterogeneous effects on the interdiction rate. Marker size is proportional to the square root of the number of scam-induced requests. The dashed line marks the pooled effect. Effects were largest for impersonation and technology-support scams, whose scripts depend on secrecy and safety instructions that the SBI probes directly, and smallest for romance and investment scams, where victims are often emotionally committed to the transaction.**

### 4.5 Reliability and fidelity

**Table 5: Inter-rater agreement and delivery fidelity for 180 double-coded SBI calls**

| Item | Agreement (%) | Cohen's κ | 95% CI | Asked verbatim (%) | Asked at all (%) |
|---|---|---|---|---|---|
| Q1 Anyone asked you to… | 97.4 | 0.84 | 0.77–0.90 | 96 | 100 |
| Q2 First contact was unsolicited… | 97.8 | 0.88 | 0.81–0.94 | 94 | 99 |
| Q3 Told to keep this… | 98.1 | 0.91 | 0.84–0.97 | 91 | 99 |
| Q4 Destination opened recently or… | 96.9 | 0.79 | 0.72–0.85 | 89 | 98 |
| Q5 Told your money is… | 98.0 | 0.90 | 0.83–0.96 | 87 | 98 |
| Q6 Promised return, prize, or… | 97.6 | 0.86 | 0.79–0.92 | 84 | 97 |
| Q7 Someone met only online… | 97.2 | 0.82 | 0.75–0.88 | 82 | 96 |
| Q8 Asked to convert to… | 98.3 | 0.93 | 0.86–0.99 | 80 | 96 |
| Q9 Deadline or penalty if… | 97.5 | 0.85 | 0.78–0.91 | 78 | 95 |
| Q10 Purpose narrative (coherence rating)… | 95.4 | 0.64 | 0.57–0.70 | 57 | 95 |
| Q11 Anyone with you or… | 97.9 | 0.89 | 0.82–0.95 | 73 | 94 |
| Q12 Comfortable if we contact… | 98.2 | 0.92 | 0.85–0.98 | 71 | 88 |
| Total score (ICC, absolute agreement) | — | 0.91 | 0.88–0.93 | — | — |

*Note.* κ benchmarks follow Landis and Koch (1977). Verbatim delivery means the representative read the scripted wording without material change. ICC = intraclass correlation coefficient.

Agreement was almost perfect for closed items and substantial for the narrative coherence rating (κ = 0.64), which requires more judgment. Verbatim delivery declined for later items, most notably the narrative item, where representatives tended to paraphrase. Paraphrasing was associated with lower discrimination of that item in exploratory analyses, suggesting that scripting matters.

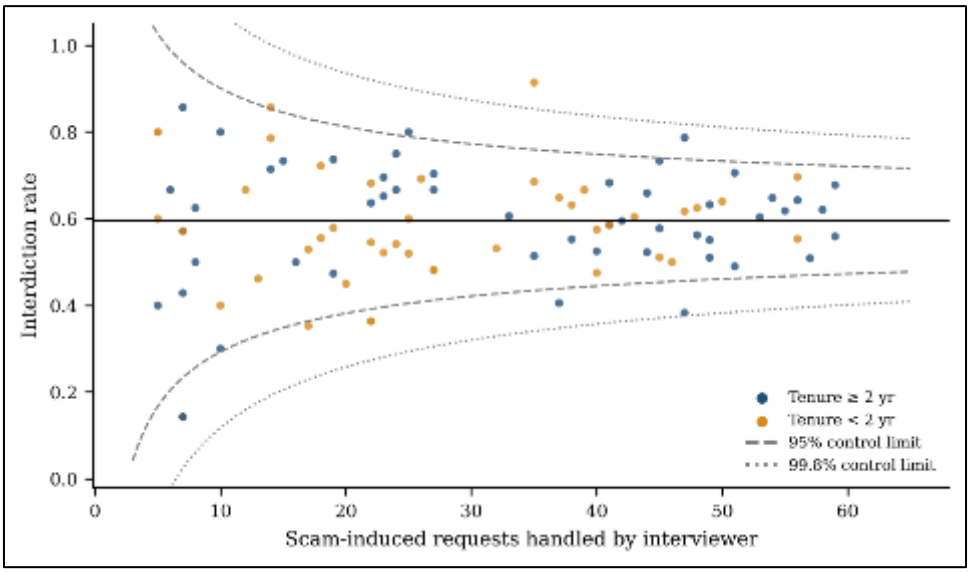



**Figure 6: Funnel plot of interdiction rates for 96 interviewers after adoption. Control limits are drawn around the pooled rate of 59.7 percent. Most interviewers fall within the 95 percent limits, implying that the protocol produced consistent performance; four fall below the 99.8 percent limit and were flagged for coaching.**

### 4.6 Customer experience

**Table 6: Customer survey responses by escalation outcome (n = 2,940)**

| Statement (% agree or strongly agree) | Released same day (n = 2,214) | Delayed, legitimate (n = 451) | Delayed, scam confirmed (n = 275) | $\chi^2$ p |
|---|---|---|---|---|
| The questions were appropriate for my request | 81.4 | 68.7 | 79.6 | <0.001 |
| The representative was respectful | 92.3 | 86.0 | 90.2 | 0.002 |
| The process took too long | 14.8 | 47.2 | 38.9 | <0.001 |
| I felt my independence was questioned | 8.9 | 23.5 | 17.1 | <0.001 |
| I trust the firm more after this call | 61.7 | 44.6 | 72.4 | <0.001 |
| Firms should ask these questions of all customers my age | 58.2 | 41.9 | 76.0 | <0.001 |
| Overall satisfaction (mean, 1–10) | 8.4 | 6.9 | 7.8 | <0.001 |

*Note.* Survey weights adjust for stratified sampling and nonresponse by age, sex, and outcome. Customers whose withdrawals were interdicted as scam-induced were surveyed at least 30 days after resolution.

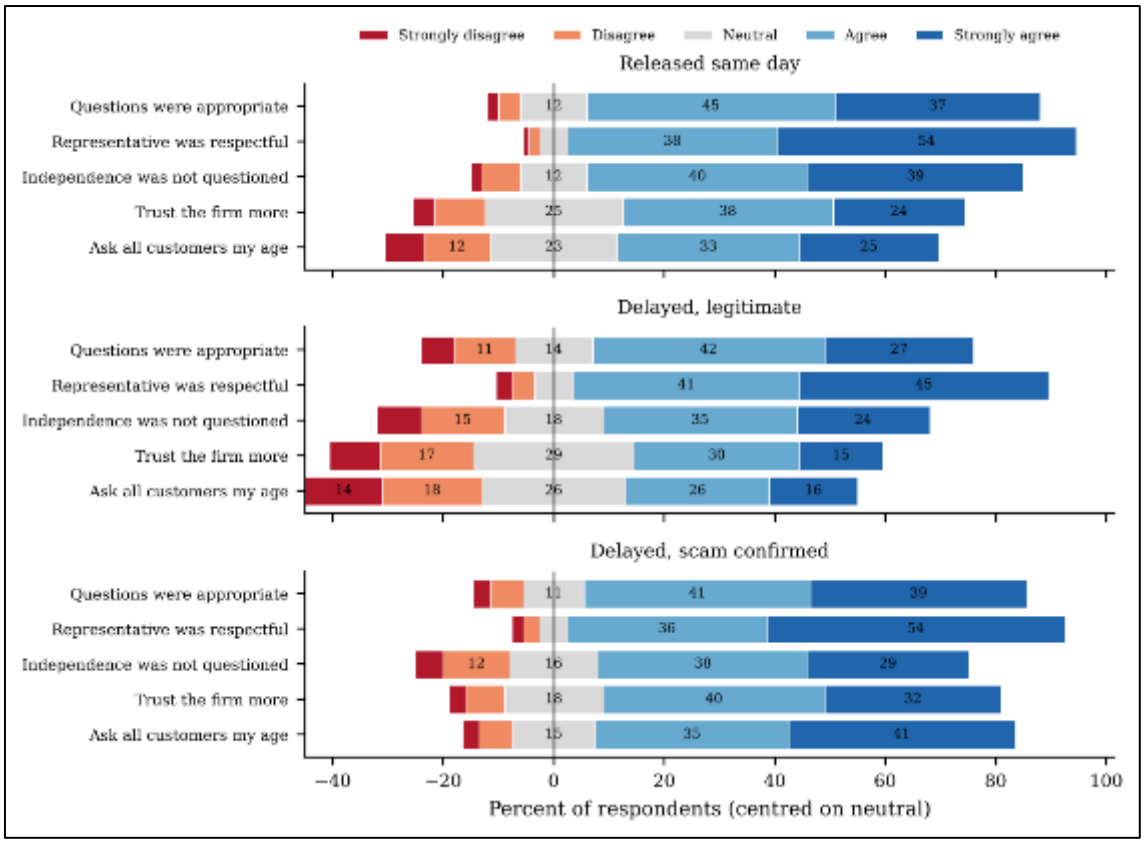



**Figure 7: Diverging stacked bar charts of customer agreement with five statements about the SBI call, by escalation outcome, following the presentation recommended by Robbins and Heiberger (2011). Legitimate customers whose withdrawals were delayed were the least positive, yet a majority still judged the questions appropriate.**

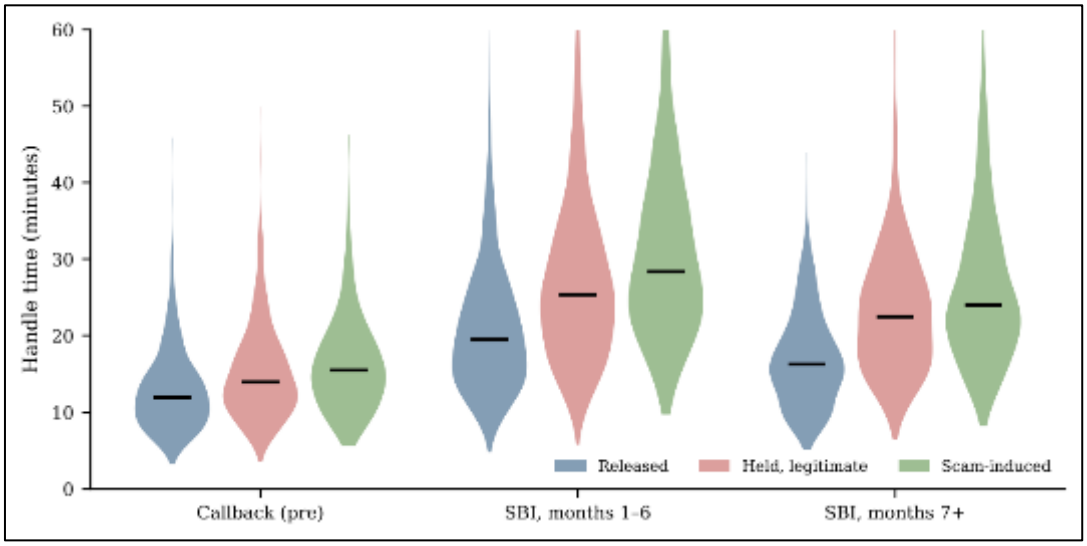



**Figure 8: Distribution of escalation call handle times by procedure period and call outcome. Horizontal lines mark medians. The SBI lengthened calls most for held and scam-induced cases, where longer conversations are the intended mechanism, and the increase for released calls narrowed after the first six months.**

### 4.7 Robustness

**Table 7: Robustness of the main interdiction estimate**

| Specification | Estimate (pp) | SE | 95% CI | Notes |
|---|---|---|---|---|
| Callaway–Sant'Anna, not-yet-treated (main) | 18.4 | 2.8 | 12.9 to 23.9 | Doubly robust |
| Callaway–Sant'Anna, never-treated proxy* | 17.6 | 3.3 | 11.1 to 24.1 | Last cohort pre-period |
| Sun–Abraham interaction-weighted | 18.9 | 3.0 | 13.0 to 24.8 | Last cohort as control |
| Stacked DiD (±8 months) | 17.1 | 2.6 | 12.0 to 22.2 | Cohort-specific FE |
| Two-way fixed effects (naive) | 14.2 | 2.4 | 9.5 to 18.9 | Biased toward zero |
| Synthetic control, cohort 1 only | 19.7 | — | Placebo rank 1/33 | Permutation p = 0.03 |
| Excluding first 2 months post | 19.8 | 3.0 | 13.9 to 25.7 | Learning period |
| Excluding branch clusters | 20.6 | 3.4 | 13.9 to 27.3 | Contact centres only |
| Placebo outcome, RMD requests | 0.6 | 1.9 | −3.1 to 4.3 | Should be null |
| Placebo adoption date (−6 months) | −0.9 | 2.7 | −6.2 to 4.4 | Should be null |

*Note.* *No unit was never treated. This row uses the final cohort's pre-adoption months as a never-treated proxy. The synthetic control placebo distribution comes from adoption dates assigned to 32 units not yet treated at the time.

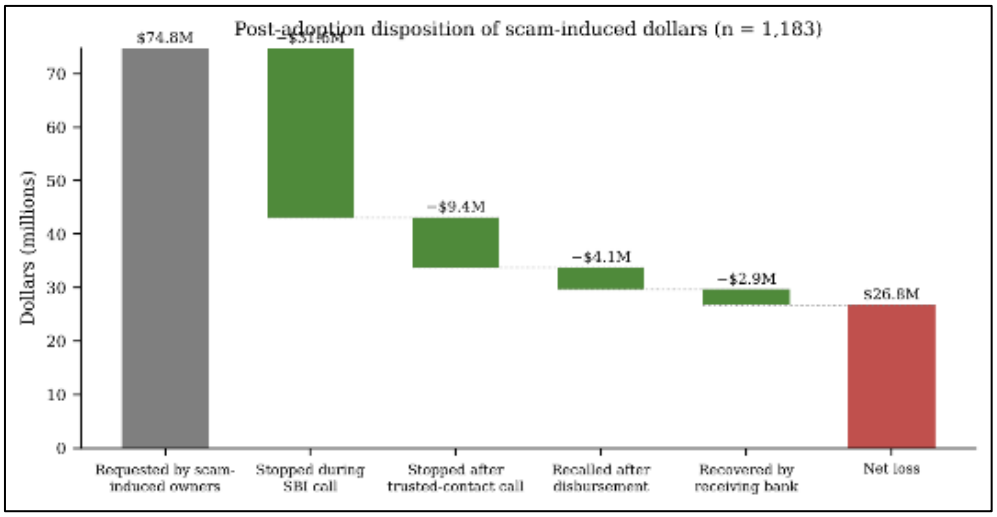



**Figure 9: Waterfall of requested dollars among scam-induced escalated requests after SBI adoption. Most prevented losses were stopped during the interview itself, and a further share was stopped after outreach to a trusted contact prompted by the interview. Under a counterfactual with the pre-adoption interdiction rate, net loss would have been roughly 41 million dollars.**

## 5 DISCUSSION

Identity verification answers the question of who is calling. The SBI answers a different question, namely whose idea the withdrawal is. Our estimates suggest that asking that question in a structured way raised interdiction of scam-induced retirement withdrawals by nearly half, with modest costs to legitimate customers that consisted mostly of a few extra minutes on the phone. The mechanism appears to be informational. Items that probe the scammer's script, such as secrecy and safety instructions, carried the most information and drove the largest effects in the scam types whose scripts rely on them.

Effects were weaker for romance and investment scams. Victims in these cases often believe in the relationship or the opportunity and answer truthfully that nobody told them what to say. For these typologies, the interview may need to be paired with longer holds, trusted-contact engagement, or third-party verification of the counterparty. The larger effects among owners aged 80 and older are encouraging because this group experiences the highest median losses.

To give a sense of scale, a industry-wide deployment that achieved the same proportional improvement among retirement withdrawals escalated by large custodians would prevent losses on the order of hundreds of millions of dollars per year, although such extrapolation assumes similar escalation practices and scam mixes and should be read as illustrative. The costs are mainly staff time. At the custodian studied here, the additional handle time translated to roughly 1.4 full-time-equivalent representatives per 10,000 escalations per year, far less than the losses prevented.

### *Limitations*

The rollout was not randomized, and although cohort timing was driven by trainer availability and pre-trends are flat, we cannot rule out unobserved changes coinciding with adoption. All units were eventually treated, so late-period estimates rely on not-yet-treated comparisons. Scam labels depend on detection, and better interviews may themselves uncover scams that would otherwise have been coded as legitimate, which would bias the interdiction rate in ways that are hard to sign. The customer survey had a modest response rate, and respondents may differ from nonrespondents. Finally, results come from one custodian, and the generalizability to banks or to digital-only channels is uncertain. Several respondents in free-text comments noted that they would have preferred a written explanation, a point we could only explore through vignettes in cognitive pretesting rather than in the field.

### *Practical recommendations*

- Ask about the scammer's script, not just the customer's identity. Secrecy, safety, and unsolicited-contact items were the most informative.
- Keep the wording scripted. Paraphrasing of the narrative item reduced its discriminating power.
- Pair the interview with trusted-contact outreach, which accounted for a meaningful share of prevented losses.
- Monitor interviewer performance with funnel plots or similar statistical process control to target coaching.

## 6 CONCLUSION

A 12-question structured interview, delivered in under seven extra minutes, substantially increased the share of scam-induced retirement withdrawals stopped before the money left. As scams increasingly work through the account owner rather than around them, controls that engage the owner as a witness to their own situation deserve a central place alongside identity verification.

## STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

### *Acknowledgments*

The author acknowledge Institute for elderly and aged research for the permission to access their database.

### *Disclosure of conflict of interest*

The authors declare no conflicts of interest regarding this manuscript.

### *Data availability*

Administrative data are confidential. The full SBI instrument and training script are available from the corresponding author for non-commercial research, together with a teaching dataset that reproduces the structure of the analysis files.

## REFERENCES

[1] Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association, 105*(490), 493–505.

[2] Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics, 119*(1), 249–275.

[3] Burnes, D., Henderson, C. R., Jr., Sheppard, C., Zhao, R., Pillemer, K., & Lachs, M. S. (2017). Prevalence of financial fraud and scams among older adults in the United States: A systematic review and meta-analysis. *American Journal of Public Health, 107*(8), e13–e21.

[4] Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230.

[5] Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics, 90*(3), 414–427.

[6] Cialdini, R. B. (2001). *Influence: Science and practice* (4th ed.). Allyn & Bacon.

[7] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

[8] Consumer Financial Protection Bureau. (2019). *Suspicious activity reports on elder financial exploitation: Issues and trends*.

[9] de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review, 110*(9), 2964–2996.

[10] DeLiema, M., Shadel, D., & Pak, K. (2020). Profiling victims of investment fraud: Mindsets and risky behaviors. *Journal of Consumer Research, 46*(5), 904–914.

[11] Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

[12] Federal Bureau of Investigation. (2024). *Elder fraud report 2023*. Internet Crime Complaint Center.

[13] FINRA. (2018a). *Rule 2165. Financial exploitation of specified adults*. Financial Industry Regulatory Authority.

[14] FINRA. (2018b). *Rule 4512. Customer account information*. Financial Industry Regulatory Authority.

[15] Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Charles C Thomas.

[16] Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277.

[17] Hanoch, Y., & Wood, S. (2021). The scams among us: Who falls prey and why. *Current Directions in Psychological Science, 30*(3), 260–266.

[18] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

[19] Langenderfer, J., & Shimp, T. A. (2001). Consumer vulnerability to scams, swindles, and fraud: A new theory of visceral influences on persuasion. *Psychology & Marketing, 18*(7), 763–783.

[20] Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378–392.

[21] Lichtenberg, P. A., Stickney, L., & Paulson, D. (2013). Is psychological vulnerability related to the experience of fraud in older adults? *Clinical Gerontologist, 36*(2), 132–146.

[22] Lichtenberg, P. A., Stoltman, J., Ficker, L. J., Iris, M., & Mast, B. (2015). A person-centered approach to financial capacity assessment: Preliminary development of a new rating scale. *Clinical Gerontologist, 38*(1), 49–67.

[23] Payment Systems Regulator. (2023). *Fighting authorised push payment fraud: A new reimbursement requirement* (Policy Statement PS23/3).

[24] Robbins, N. B., & Heiberger, R. M. (2011). Plotting Likert and other rating scales. In *Proceedings of the 2011 Joint Statistical Meeting* (pp. 1058–1066). American Statistical Association.

[25] Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics, 235*(2), 2218–2244.

[26] Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2), 175–199.

[27] Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology, 22*(1), 1–21.

## APPENDIX A. ABBREVIATED INTERVIEW EXCERPT

The following excerpt illustrates delivery of items Q3 and Q10. It is a composite constructed for training and does not reproduce any customer call. Representative: "Has anyone asked you to keep this withdrawal private, including from your family or from us?" Customer: "Well, the agent said not to tell the bank because there may be an employee involved." Representative: "Thank you. Could you tell me, in your own words, what the funds are for?" The representative would rate the customer's answer to the second question as low in coherence if it relied on phrases supplied by a third party, such as a case number the customer cannot explain.