Upgrading Financial Markets

Francis Annan*†
UC Berkeley and NBER

Seth Garz‡
Gates Foundation

Version: October 9, 2026

Abstract

Expanding access to financial services may not be enough to develop markets in low-income settings: providers also need to be held accountable to consumers. We test this in a field experiment across 195 rural Sierra Leonean communities, upgrading Village Savings and Loan Associations (VSLAs) into mobile money and bank agents, cross-randomized with a monthly consumer feedback mechanism. Upgrading expands access, but feedback drives large operational improvements (+20.5 pp, 85%), particularly in privacy and consumer protection, raising usage and revenues. Upgraded VSLAs also show higher savings digitization and higher incomes, evidence that disciplining agent actions benefits not only consumers but also the principal that finances and stands behind its agent. Interpreted through a framework of delegated intermediation, we show accountability, not infrastructure alone, aligns provider incentives with consumer welfare and is central to designing scalable financial intermediaries.

JEL Codes: G21, O17, D22, D82, C93

Keywords: Financial Intermediation, Formal–Informal Upgrade, Enforcement (Barray), Firm Behavior

* Field and partnership support from Innovations for Poverty Action (IPA)-Sierra Leone, Catholic Agency for Overseas Development (CAFOD), Orange Money, and GT Bank is acknowledged. Thanks to Lucas Li, Ei Thandar Myint, Chiman Cheung and Kenshi Kawade for excellent research assistance. We are grateful to ReFinD Research Initiative for funding. Institutional Review Board (IRB) approvals for research data collection were obtained from IPA (#16387), the Government of Sierra Leone (#009/11/2023) and UC Berkeley (#2024-06-17541). The project and outcome variables were pre-registered in the AEA RCT Registry, ID-0011825. For helpful comments, we thank Paul Niehaus, Cátia Batista, Pedro C. Vicente, Xavier Gine, Peter Quartey, Travis Lybbert, Dean Karlan, NCLUSION.com, and several other colleagues.

† 207 Giannini Hall #3310, Berkeley, CA 94720-3310, USA and NBER. fannan@berkeley.edu.

‡ Data and Technology Adoption (DATA) Strategy. Seth.Garz@gatesfoundation.org.

1. Introduction

Access to finance is critical for economic development, yet in many rural areas of low-income countries formal financial markets are absent (Morduch and Karlan (2009)). Households and small enterprises face severe constraints in saving, transferring funds, managing risk, and financing productive investments. Informal institutions, such as Village Savings and Loan Associations (VSLAs), typically women-led groups that pool members' contributions and extend small loans to one another, partially mitigate these frictions but provide limited services and operate at small scale. As a result, rural communities exhibit substantial latent demand for financial services alongside limited supply.

A growing literature studies the expansion of formal finance into underserved areas, including mobile money (Suri et al. (2023); Batista and Vicente (2025)) and rural banking (Burgess and Pande (2005); Barboni, Field, and Pande (2025)), showing that connecting households to formal finance can generate meaningful benefits and reduce poverty. This work pays limited attention, however, to the incentive and information structures that let consumers hold providers accountable and sustain market growth once access is established. VSLAs, though widespread in rural economies, have rarely been used as vehicles for introducing formal digital financial services (DFS).[1] Consequently, we know relatively little about how newly created financial markets function. We also know little about how market design at the point of entry shapes service quality, trust, sustained usage, and provider behavior.

This is important for at least two reasons. First, adoption often lags behind usage: across Sub-Saharan Africa, account ownership rose from 34% of adults in 2014 to 58% in 2024, yet account-holder inactivity remains substantial, recently estimated at roughly one in eight (Demirguc-Kunt et al. (2022); Klapper et al. (2025)). Merely providing access does not guarantee that financial markets function effectively (Karlan et al. (2014); Prina (2015)) or that households derive the full range of potential benefits. Second, technology adoption and market integration are prone to path dependence and lock-in (David (1991); Katz and Shapiro (1986); Krugman (1994)), which can leave initial inefficiencies uncorrected. The challenge therefore extends beyond market creation to ensuring that newly connected markets remain functional, accountable, and responsive—dimensions largely ignored in prior

[1] CARE, a major international humanitarian and development organization, reports that its own VSLA network alone comprises 536,977 groups and 29,889,968 members (72% female) across 67 countries as of fiscal year 2025 (CARE International (2025), https://www.care.org/wp-content/uploads/2025/10/VSLA-Annual-Report-2025_EN.pdf). Other major international NGOs operate VSLA-type savings-group networks of comparable scale across Sub-Saharan Africa and beyond (Catholic Relief Services (2025); World Vision International (n.d.)). Any intervention that interlocks with or upgrades existing VSLAs can draw on a very large, already-functioning base. This is exactly the kind of scale a commercial DFS provider needs to make such a partnership worthwhile, making it a commercially viable and highly scalable approach to expanding financial services to the poor.

work.

In this paper, we study whether and how financial markets can be upgraded: not only by expanding access, but also by increasing usage, service quality, and provider responsiveness through consumer accountability. We investigate this in rural Sierra Leone through a large-scale, multi-year, community-level randomized controlled trial that builds on a 2019–2020 pilot, testing a scalable approach that upgrades existing VSLAs to operate as formal DFS agents and examining whether consumer evaluation and feedback mechanisms further improve market performance by shaping provider behavior and incentives.

We focus on two research questions: first, what are the broader impacts of introducing DFS into virgin rural markets, where trusted informal finance networks such as VSLAs supply formal financial services and may generate general equilibrium effects on households, businesses, and communities; and second, once markets are created, can consumer evaluation mechanisms improve service quality, usage, and trust—by shaping provider effort and provider–customer relationships—and through which channels do these changes translate into improved economic outcomes?

Our randomized controlled trial covers 195 rural communities and is implemented in partnership with the country's largest financial service providers. The first treatment promotes market entry by encouraging VSLAs to enroll as DFS agents, offering mobile money services through Orange Money and basic banking services through GT Bank.[2] In each treated community, one VSLA is onboarded and represented by two financial champions, one male and one female, who operate on its behalf. This design leverages the liquidity, social capital, and local reputation of VSLAs, thereby reducing entry frictions and facilitating trust. The second treatment augments agent entry with a consumer evaluation mechanism. DFS users provide monthly evaluations of local agents along multiple dimensions: fees, privacy, transparency, reliability, and overall satisfaction. Evaluations are digitized and shared with agents and communities through community gathering, creating reputational incentives in markets with limited competition. This mechanism shifts trust from informal, analog reputation to structured feedback. The experimental design allows us to measure both the extensive margin (market entry and account ownership) and the intensive margin (usage, provider effort, and business outcomes). This lets us isolate the causal role of accountability in driving market upgrading. Randomizing both agent entry and feedback lets us separately identify the effects of market creation and market discipline.

Theoretically, when competition is weak or absent and consumers lack information, providers may exert low effort, charge high or opaque fees, or deliver unreliable service

[2]Appendix H.4 reproduces excerpts of the standard agent agreement onboarded VSLAs sign with Orange Money.

(Blackmon, Mazer, and Warren (2021); Garz et al. (2021); Annan (2025)); consumer evaluation mechanisms that give consumers voice can mitigate these distortions by strengthening reputational incentives to align provider behavior with consumer welfare. This is a standard principal-agent problem in which disciplining agent effort is not zero-sum: each DFS agent is fielded by, and financially backed by, its own VSLA, so better service benefits consumers while also benefiting the VSLA itself, since the agent's performance feeds back into the group's own financial standing.

We collect rich data to measure impacts on financial access and usage of DFS, provider service quality and trust, household shock mitigation, and spillovers to VSLAs and communities. The analysis proceeds in five steps. First, effects of provider agent entry on market creation. Second, effects of consumer feedback on provider behavior and service quality. Third, impacts on community and household financial usage, trust, and economic outcomes. Fourth, spillovers within communities, VSLAs, and commercial providers. Fifth, effects on household-level shock mitigation and financial resilience.

We document five main results. First, connecting communities to digital financial services significantly expands financial access, through the entry of mobile money and bank agents. Provider presence increases (+0.4 agents = +75%), and formal account ownership rises (+5.2 pp = +23%). Usage effects under provider entry alone are narrower: households increase deposits into digital accounts and digital savings, but the broader adoption-usage gap (whether an account holder actually transacts at all) does not close, peer-to-peer transfers and cash withdrawals remain unchanged, and the intensive margin – transaction value and frequency among those who do transact – shows no detectable change either. These patterns indicate that facilitating provider entry activates latent demand and brings households into the formal financial system. On its own, though, it generates only a partial shift toward active usage and market development.

Second, introducing structured consumer feedback substantially amplifies both access and usage, closing much of this remaining gap. Formal account ownership increases by 9.1 pp (+39%), and mobile-money usage specifically rises by 10.4 pp (+29%). This additional usage gain comes through the two channels provider entry alone left unchanged: cash withdrawals and peer-to-peer transfers, and it shows up on the intensive margin too: among households that do transact, transaction values (in SLL) increase by over 200% relative to control. In particular, feedback induces providers to improve service quality and consumer protection, raising business revenues and consumers' financial resilience.

Third, provider entry alone raises household trust in agents of mobile-money (Momo) by 6.7–6.9 pp (8.5–8.7%), with no change in trust toward family, friends, other community members, or the banking agent in either arm. Feedback's distinctive contribution is instead

to provider conduct: Upgrade+Feedback agents are 13.2 pp less likely than Upgrade-alone agents to allow a third person present during a transaction and 0.249 SD lower on a privacy-practice index (both significant at 5%), and 20.5 pp more likely to report having made changes in response to feedback, an 85% increase over the Upgrade arm's 24% base rate. Active consumer voice thus sharpens provider conduct even where it leaves households' own stated trust in the DFS channel unchanged.

Fourth, we document significant spillover effects. Upgraded VSLAs exhibit greater digitization of member contributions, lower default rates, and higher group-level income. This indicates that organizational upgrading strengthens financial discipline and collective performance. Disciplining agent effort through consumer feedback is therefore not zero-sum: it benefits the consumers the agent serves, and it benefits the VSLA that finances and stands behind that agent, a win-win consistent with resolving the underlying principal-agent problem through consumer voice. This default reduction is a matter of members' contributions to the group being delayed or unmade, not loan repayment. It is concentrated among previously low-risk groups, VSLAs with few members in default or delayed contribution, consistent with digitization resolving mechanical rather than strategic non-compliance in member contributions.

Fifth, structured consumer feedback specifically, rather than the underlying upgrade alone, improves household-level shock mitigation: across alternative measures, households in Upgrade+Feedback communities are 3.8–4.6% of the control-arm mean less likely to report a shock they could not remedy financially, and the two arms are statistically distinguishable. Because Upgrade alone holds the market's infrastructure fixed while Upgrade+Feedback adds feedback on top of it, this difference isolates the market-discipline channel, consistent with our Second and Third results. We find no impact on overall poverty, though our design targeted market development rather than other structural household welfare outcomes directly.

Together, these results indicate that changes in provider behavior and stronger provider-consumer relationships drive the improvements documented above: expanding access is a necessary first step, but interventions that combine access with structured feedback and active consumer engagement are far more likely to achieve sustained usage, improved service quality, and broader economic benefits across the market. Consumers get better service, the VSLAs that capitalize and stand behind their agents perform better, and the commercial providers that supply them see more transaction activity. Successful market development therefore depends not only on increasing the supply of financial services (Wieser et al. (2019)), but also on ensuring that providers respond to consumer experiences and that consumers actively participate in shaping the market.

Our study contributes to three existing literatures. First, it extends research on financial market expansions in developing countries. We provide new evidence on creating DFS markets in unserved rural areas (Burgess and Pande (2005); Wieser et al. (2019); Batista and Vicente (2025)). Most prior studies focus on incumbent markets with existing providers and experienced consumers (Annan (2025); Garz et al. (2021); Deserranno, León-Ciliotta, and Witoelar (2019)). We instead study virgin markets, where DFS is essentially nonexistent. We add to this literature by showing how formal financial services can leverage informal institutions, such as VSLAs, as entry points to expand access, build trust, and stimulate adoption and economic engagement. We are unaware of any study that upgrades existing VSLAs to provide local economy-wide DFS: Francois and Squires (2021) instead link "anonymous" shopkeepers via mobile money accounts in the Democratic Republic of Congo, and Banerjee et al. (2024) document negative effects of microfinance on informal social networks, whereas we find complementary effects on informal group-level discipline and performance.

Second, we shed light on a novel mechanism for scaling and disciplining new markets: upgrading VSLAs into DFS agents creates a scalable entry model that exploits their local reputation, liquidity, and ability to crowd in consumers, especially women, complemented by consumer evaluation and feedback mechanisms that discipline providers in low-competition markets. This builds on a growing literature on review-based accountability mechanisms (Pallais (2014); Wang (2022); Fradkin-Holtz (2022); Mbiti and Serra (2022); Cilliers, Mbiti, and Zeitlin (2021)); ours is the first to test such mechanisms in newly created DFS markets, combining market creation and early market discipline, a combination prior research largely overlooks.

Third, our results provide one potential explanation, and thus solution, to the observed differences between adoption and usage of financial services (Karlan et al. (2014); Prina (2015)). Technology-driven financial markets exhibit path dependence and lock-in (David (1991); Katz and Shapiro (1986); Krugman (1994)), whereby inefficiencies in early-stage markets can persist. Our study demonstrates that initial design choices, such as leveraging trusted informal financial networks, structuring provider incentives, and integrating consumer feedback, have meaningful effects on provider behavior. These, in turn, drive increases in service quality and consumer trust, which raise adoption and usage outcomes jointly.

The rest of the paper is structured as follows. Section 2 describes the study's background and institutional setting. Section 3 details our experimental design. Section 4 develops our conceptual framework. Section 5 describes our data. Section 6 presents our main results. Section 7 interprets these results and prospects for scale. Section 8 concludes.

2. Background

We carry out our field experiment in rural Sierra Leone, a context characterized by significant financial under-depth. Since the end of the civil war, Sierra Leone has struggled to expand formal financial services to its rural interior, where over 60% of the population resides. During our study period, the mobile money sector was the fastest-growing segment of the financial system, yet retail agent networks remained concentrated in urban hubs due to high liquidity costs and fragmented demand. Commercial providers, including our partners Orange Money and GT Bank, have been hesitant to scale their networks to rural villages. This hesitation stems from a lack of empirical evidence to justify the high fixed costs of expansion and a lack of specific, tested models capable of overcoming the unique liquidity and trust barriers prevalent in the "last mile" of the financial market. Digital financial services (DFS) are now a central pillar of the government's National Financial Inclusion Strategy, aimed at reducing the reliance on physical cash in a country where the vast majority of the population remains unbanked.

Our project partner is a unique consortium led by the Catholic Agency for Overseas Development (CAFOD), Orange Money, and GT Bank. CAFOD is a leading development finance NGO actively engaged in implementing and scaling financial inclusion programs. Their work focuses on promoting the access and availability of needs-based financial services tailored to the specific constraints of rural communities. Orange Money serves as the country's largest mobile money operator, while GT Bank provides the regulatory and banking infrastructure for branchless intermediation. We coordinate with Innovations for Poverty Action (IPA) to manage the experimental implementation across 195 communities.

The primary unit of financial organization in these communities is the Village Savings and Loans Association (VSLA). VSLAs are ubiquitous across rural sub-Saharan Africa and serve as the main institutionalized setting for borrowing and saving in the absence of commercial banks (Francois and Squires (2021)). A typical VSLA in our sample consists of approximately 25 members, of whom roughly 66% are women. These groups are highly localized; our study focuses on 195 "virgin" communities that are geographically distinct to limit cross-community spillovers, have mobile network connectivity, and lack any pre-existing retail DFS agents. These communities have populations ranging from 115 to just over 5,700 people (mean of under 750).

The internal financial structure of a VSLA is capital intensive relative to the local economy. VSLAs in our sample were formed between 2010 and 2022 and hold financial assets between 1,000 and 46,486 SLL (mean of roughly 8,300 SLL), with average weekly savings of about 164 SLL per member. Because these groups possess established village-level reputa-

tions and significant internal capital, they are uniquely positioned to overcome the liquidity and trust challenges that commonly constrain individual retail DFS agents (Karlan et al. (2016)). Appendix H.1 shows the terrain that motivates this last-mile access problem, along with the VSLA money box whose multi-key physical security underpins these groups' local reputation for safeguarding member funds. Moreover, they have the ability to crowd-in historically excluded populations, particularly women. This setting provides a clean testing ground for the introduction of both infrastructure (via the VSLA-to-agent upgrade) and accountability (via the feedback mechanism), as there is no pre-existing competition from other formal financial intermediaries.[3]

Absent a local DFS agent, households rely on one of two informal workarounds to transact digitally. The first is to collate cash, a personal phone, and personal passwords and PINs, and hand them to an okada driver who travels to the urban area to carry out the transaction on their behalf. This channel is expensive, since drivers charge a hefty fee on top of the standard DFS tariff, and it carries risks of scams and loss of privacy: customers cannot check their own balances, cannot use their phone for the day, and, given low digital literacy, have limited ability to verify what the driver actually did on their behalf. The second is to wait for mobile agents who periodically visit market days – the periodic markets held in larger chiefdom communities, with each chiefdom holding its own market day. This channel still involves meaningful travel cost to reach the market, uncertain agent presence once there, and the same scam-driven trust barriers that affect the okada channel. Both workarounds leave households paying a premium, in cash, time, or risk, for DFS access that a local agent would otherwise provide directly.

Prior to our intervention, these groups operated as purely informal entities with no external monitoring. Decisions regarding credit allocation and the physical security of funds are managed by a group-appointed provider. While members receive informal feedback during weekly meetings, there is no formal mechanism for verifying provider conduct or ensuring adherence to consumer protection protocols, such as fee transparency or transaction privacy. This lack of verifiability creates a classic principal-agent problem, where providers may shirk on effort to maintain high-quality service, thereby suppressing the intensive margin of financial usage.

[3]At baseline, only 6% of households report DFS agent access inside their own community (Table C.3, column 2); this does not mean residents have no DFS access at all. 32% of households report access to a DFS agent outside the community (Table C.3, column 3), typically in the nearest town, and residents without local access commonly reach these agents by collating their cash, phone, and mobile-money passwords and PINs with a third-party motor-bike ("okada") rider, who then transacts on their behalf in town.

3. Experiment

The experiment was designed to measure and separate the impacts of infrastructure expansion and active accountability. Our intervention leverages the existing organizational set up and capital of VSLAs to introduce formal financial access to "virgin" rural markets. As one trader noted during our pilot phase: "Since mobile money was introduced... it has eased movement to the far away bank." This qualitative evidence underscores the high latent demand for localized financial intermediation.

3.A. Treatments

We randomly assigned the 195 study communities into a control group and two distinct treatment arms. Randomization was conducted at the community level to ensure that the accountability mechanism operated over a defined social network. To minimize the risk of geographical spillovers while allowing for the measurement of within-village behavioral changes, we selected communities that are geographically distinct and significantly distanced from one another.

Treatment I: Upgrade (Supply-side). In 72 communities, we encouraged existing VSLAs to enroll as formal retail agents for Orange Money and GT Bank. Take-up was universal; all VSLAs offered the intervention accepted. To ensure operational viability and mitigate liquidity constraints, eligibility was restricted to VSLAs with at least 1,500 SLL in financial assets—a threshold met by the vast majority of associations in our sample. Each participating VSLA appointed two "Financial Champions" (FCs)—one male and one female—who served as the primary agents for the broader community. To standardize the retail infrastructure, each VSLA-agent was supplied with a solar panel, two mobile handsets, a loudspeaker, an umbrella, a branded bracelet, and a service table. Additionally, each agent received from their VSLA an initial liquidity injection of at least 1,500 SLL to serve as working capital, split as 1,000 SLL for mobile-money float and 500 SLL for agency-banking float. The VSLA maintained autonomy over the internal wage structure for these FCs. This treatment arm represents a pure supply-side expansion of the physical and digital financial infrastructure.

Treatment II: Upgrade + Feedback (Supply + Accountability). In an additional 77 communities, we augmented the infrastructure upgrade with a monthly, anonymized feedback mechanism designed to "digitize trust." IPA enumerators conducted brief monthly phone audits with DFS users to solicit binary evaluations across five service dimensions. Respondents were read the following script:

> "Hi—We noticed that you transact DFS at Agent [X]. I am going to ask you to review Agent [X] and you will receive 10 SLL worth of airtime credits for

> participating. Reviews enable others to make informed decisions and help build the new DF marketplace. All review information will be aggregated, anonymized, and shared at the community level. How will you rate Agent [X] based on: (i) fees charged, (ii) privacy protection, (iii) tariff non-posting, (iv) agent absence or failed transactions, and (v) overall satisfaction? [-1 = negative, +1 = positive]."

The transmission of this feedback utilized a traditional institutional channel: the community court *barray*. In Sierra Leone, the *barray* functions as a community-based public forum for discussion, dispute resolution, and collective decision-making, akin to an informal court system. As part of CAFOD's monthly engagement activities, these digitized user reviews were aggregated and presented during *barray* gatherings. By injecting objective performance data into this high-stakes public forum, the treatment provides a formal verification of conduct in a market characterized by information asymmetries and limited competition. Appendix H.2 shows a fielded DFS agent and a *barray* gathering at which this feedback was presented.

3.B. Randomization and Implementation

Randomization of treatments was stratified by community population (low vs high) and the baseline density of village informal businesses (low vs high), resulting in four distinct strata. Misfits were resolved through random assignment within strata. We ensure balance on baseline characteristics (Bruhn and McKenzie (2009)), using baseline information gathered on the communities and VSLAs by our partner CAFOD.

To manage the logistics of the large-scale rollout, the onboarding process was deployed over a three-year horizon, with roughly 25 VSLAs assigned to each treatment arm per year and the remainder of that year's enrolled communities retained as pure controls, receiving neither infrastructure upgrades nor consumer feedback. Pooled across all three enrollment cohorts, this yields 72 communities in Treatment I, 77 in Treatment II, and 46 controls (N=195). Appendix A reports the full data collection timetable—customer-survey rounds and endline surveys, with sample sizes and start and end dates—for each of the three enrollment cohorts.

4. Conceptual Framework

We develop a simple framework of delegated intermediation to clarify how financial infrastructure and active accountability interact to shape market development. The model contrasts two interventions: those that expand access (upgrading VSLAs to agents) and those that introduce accountability (the feedback mechanism). In the model, the provider acts

as a delegated agent whose effort in maintaining service quality and consumer protection is costly to monitor.

Formally, a provider (the VSLA agent) chooses a level of operational effort $e \in [0, 1]$, representing adherence to costly consumer protection protocols (e.g., transaction privacy and fee transparency). This effort is observable to consumers and to the researcher (through agent self-reports and consumer-perceived privacy practices) but non-contractible: no formal contract can specify, verify, or enforce a target privacy practice, so absent an external accountability mechanism the provider cannot credibly commit to it. Exerting effort incurs a convex cost $C(e) = \frac{1}{2}\phi e^2$. Consumers choose a level of usage $u \geq 0$, which depends on their trust in the provider $T(e)$, where $T'(e) > 0$. We take $T(e)$ to be trust tied specifically to the provider's observed conduct—the kind formed through a consumer's own recent transactions with, or ratings of, that agent—rather than a diffuse, general attitude toward the provider that need not track e at all; Section 6.D tests this distinction directly. We define usage as:

$$u(e) = A[u_0 + \tau T(e)] \tag{1}$$

where $A \in \{0, 1\}$ is an indicator for infrastructure access, u_0 is baseline usage, and $\tau > 0$ is the sensitivity of usage to provider conduct. The provider earns revenue $R(u) = \gamma u$.

Two features of this market motivate the model having only two levers, A and θ, rather than a price or entry margin. First, each community is served by at most one agent, so there is no local price competition to discipline behavior. Second, transaction fees are set by the mobile-money and banking providers under fixed schedules, not chosen by the local agent, so price is not a decision variable either. Overcharging is accordingly not a price set above a competitive or regulated benchmark, but a form of non-compliance – charging above the mandated fee, or levying informal add-on charges – and we bundle it, along with PIN and third-person-transaction privacy practices, into the single effort variable e. With one agent per community and fees fixed exogenously, e is the provider's only behavioral margin relevant to consumer welfare, and A and θ – infrastructure and accountability – are the only two levers this paper's design varies.

4.A. Accountability and Incentive Alignment

Without accountability, effort e is observable but non-verifiable (non-contractible), leading to a moral hazard problem where providers shirk on privacy protection to save costs. We introduce accountability through a feedback parameter $\theta \in [0, 1]$, representing the probability

that misconduct is verified and penalized. The provider maximizes:

$$U_p = \gamma A[u_0 + \tau T(e)] - \frac{1}{2}\phi e^2 + \theta V(e) \tag{2}$$

where $V(e)$ is the reputational value or future rents associated with positive evaluations. The first-order condition, $\gamma A\tau T'(e) + \theta V'(e) = \phi e$, generates three stylized predictions, which we state formally as Propositions 1–3 below; proofs are collected in Appendix B.

Proposition 1 (Access Effect)**.** *Infrastructure alone ($A = 1$) is a necessary condition for usage ($u > 0$). However, if $\theta = 0$, the provider's effort is driven solely by the baseline commission $\gamma\tau T'(e)$. If the cost of effort ϕ is high, the provider chooses a minimal effort level e^*, limiting trust and intensive-margin usage.*

Proof. See Proofs 1A–1B, Appendix B.

Proposition 2 (Behavioral Shift and "Step-Function" Jump)**.** *Introducing feedback ($\theta > 0$) increases the marginal return to effort. Since $\frac{\partial^2 U_p}{\partial e \partial \theta} = V'(e) > 0$, accountability unambiguously raises effort ($\frac{\partial e^*}{\partial \theta} > 0$). Because this is a discrete shift in the incentive environment rather than a gradual acquisition of technical skill, the model implies an immediate "step-function" jump in e as soon as θ becomes positive, rather than the gradual increase in e that a standard learning curve would predict as effort costs ϕ decline with experience.*

Proof. See Proofs 2A–2B, Appendix B, which nests this prediction against a learning benchmark in a single model and derives a testable discontinuity at the feedback launch date rather than relying on timing alone. This framework explains the immediate +20.5 pp surge in operational improvements and the linear response to constant monitoring pressure observed in our field experiment, and is tested directly against the round-by-round consumer-review data in Appendix D.

Proposition 3 (Complementary Effects)**.** *Let $V_p(A, \theta) \equiv U_p(e^*(A, \theta); A, \theta)$ denote the provider's indirect (optimized) utility. Then A and θ are supermodular in V_p ($\frac{\partial^2 V_p}{\partial A \partial \theta} = V'(e^*)\frac{\partial e^*}{\partial A} > 0$), so the marginal impact of accountability is amplified by the presence of infrastructure. Higher effort $e^*(\theta)$ bolsters trust, e.g., by reducing the leakage of sensitive personal or household financial information within the village, which in turn increases usage u. In the provider's revenue function, u and e act as complements.*

Proof. See Proofs 3A–3B, Appendix B; extending this result from provider profit to household welfare (Proof 3B) requires assuming welfare itself is increasing and supermodular in usage and effort, which we state explicitly there. This generates a virtuous cycle of increased digitization of contributions, higher business revenues, and improved incomes via lower default rates within the VSLAs.

5. Data

We combine three data sources to track the evolution of the market: high-frequency behavioral audits, detailed household and firm-level surveys, and administrative records from our commercial and NGO partners. The project's full multi-year enrollment and data-collection timeline and activities, spanning the year-0 pilot, each cohort's staggered onboarding, the three rounds of customer-review audits, and the endline surveys, are reported in Table A.1.

5.A. Survey and Behavioral Data

Provider Behavior and Consumer Feedback. To measure service quality and consumer protection, we utilize the high-frequency phone audits conducted by IPA. These audits provide monthly, user-level evaluations of agent conduct. By aggregating these binary ratings across dimensions such as price transparency, privacy protection, and reliability, we construct a time-varying index of provider behavior (e) that allows us to examine the month-to-month dynamics of the feedback mechanism.

Household and Provider Surveys. At the conclusion of the study, we conducted comprehensive surveys across three levels of the market: (i) community households, (ii) the Financial Champions (agents or providers), and (iii) the VSLA leadership (Secretaries). These surveys allow us to track the extensive margin (market entry, provider operational presence, and account ownership) and the intensive margin (provider effort, demand, and prices) of the financial market, including privacy practices, consumer protection, and trust outcomes. To capture broader downstream impacts, we collect data on enterprise development, including business revenues, VSLA member contributions, and contribution delays or defaults.

Poverty and Shocks. Given the sensitivity of welfare to financial shocks in rural Sierra Leone, we examine impacts on household expenses and poverty levels. We utilize the "Simple Poverty Scorecard" for Sierra Leone (Schreiner (2018)), a rigorously validated tool that uses nine low-cost indicators to estimate the likelihood that a household's consumption is below a given poverty line. This allows for a transparent and standardized assessment of the intervention's impact on absolute poverty. We additionally gather data on households' shock experience and mitigation strategies to examine the intervention's potential welfare effects.

5.B. Administrative Data

We leverage two types of administrative data. First, we utilize baseline records from CAFOD for stratification and randomization balance checks across the 195 communities. These records include community population, baseline business density, including VSLA asset holdings and group-level characteristics. Second, we integrate administrative transaction data

from Orange Money and GT Bank. These records provide measures of agent activity, including total transaction volumes, sales revenue, and commissions earned, which we use to verify the self-reported usage data from the household surveys.

5.C. Balance and Attrition

Table 1 reports summary statistics for the VSLA-level randomization frame (195 VSLA groups across the three enrollment cohorts), predating any endline attrition.

Table 1 gives a sense of the communities in this study: they are small and, by most available markers, poor. The typical community has fewer than 750 residents (ranging from 115 to just over 5,700), and the VSLAs themselves are modest, tightly-clustered groups of about 25 members holding, on average, under 6,000 Sierra Leonean Leones in current cash and roughly 8,300 Leones in total cash assets—small sums even by the standards of rural West Africa. Average weekly savings per member (about 164 Leones) are similarly modest. The table's bottom panel sharpens this picture with two additional endline figures, drawn from the untreated control arm alone and therefore uncontaminated by our intervention: 62.9% of control-arm households have an estimated likelihood of falling below the national poverty line, confirming that these are, in fact, poor communities by a direct, survey-based welfare measure and not only by the indirect VSLA-level proxies above; and only 6% of control-arm households report having a DFS/mobile-money agent physically present inside their own community, underscoring how thin formal financial-service coverage was in this setting before our intervention. This is the population—small, cash-constrained, informally organized, and poorly served by formal financial infrastructure—into which our intervention introduces formal digital financial infrastructure.

Balance. Table C.1 reports pre-intervention randomization balance on the same VSLA-level randomization frame. For each of eight baseline covariates—VSLA year of formation, female and total membership, current cash, outstanding loans, total cash assets, average weekly savings, and community population—we regress the covariate on arm indicators with randomization-strata and cohort fixed effects. No individual covariate is significant at conventional levels. A joint test of the null that covariates predict treatment arm—a multinomial probit of arm on seven of the eight covariates (total cash assets is excluded from the joint test only, since it is an accounting identity of current cash plus outstanding loans and therefore exactly collinear with them), with strata and cohort fixed effects—fails to reject at any conventional level ($\chi^2(14) = 16.95$, $p = 0.259$, $N = 156$). A companion linear-probability test collapsing both treatment arms into a single pooled indicator gives the same conclusion ($F(7, 143) = 0.79$, $p = 0.599$, $N = 156$). Randomization succeeded in balancing observable baseline characteristics across arms.

Attrition. Table C.2 compares this baseline roster of 195 VSLAs against the endline analytic samples, separately for households, agents, and VSLA secretaries. A VSLA counts as matched for a given respondent type if at least one endline record of that type can be linked to it; twelve dual-VSLA communities that cannot be unambiguously disambiguated by name are excluded, leaving 171 VSLAs for the household/secretary panels and 136 (the two treatment arms only, where agents were fielded) for the agent panel. Differential attrition relative to control is modest, in the 6–10 pp range across arms, and is only marginally significant for the household/any-respondent measure (joint $F = 3.92$, $p = 0.022$); it is not statistically distinguishable from zero for VSLA secretaries or agents considered separately, nor—under a more permissive matching threshold that resolves likely transcription-variant community-name spellings—for any respondent type. We return to whether this residual attrition could account for our results in Section 7.B.

6. Results

6.A. Estimation Strategy

We present and discuss the treatment effects estimated using the following specification:

$$Y_{ivd} = \alpha + \beta_1 \text{Upgrade}_{iv} + \beta_2 (\text{Upgrade+Feedback})_{iv} + \delta_d + \gamma \mathbf{X}_{iv} + \epsilon_{ivd} \tag{3}$$

where Y_{ivd} is the outcome of interest for subject i (provider, VSLA, or household) in community v within randomization strata d. Upgrade_{iv} is an indicator for communities receiving the infrastructure expansion only, and $(\text{Upgrade+Feedback})_{iv}$ is an indicator for the combined treatment. $\mathbf{X}_{iv}$ is a vector of baseline community and VSLA business characteristics, and δ_d are randomization strata dummies to account for design effects and improve precision. Standard errors are clustered at the community level, our unit of randomization. To deal with the potential issue of multiple testing, we follow Kling, Liebman, and Katz (2007) (KLK), and construct an index pooling all outcomes in a related outcomes family. For a given set of K outcomes, the summary index is defined as the unweighted average of the z-scores: $I_{KLK} = \frac{1}{K} \sum_{k=1}^{K} (y_k - \mu_k)/\sigma_k$, where y_k represents the observed outcome, and μ_k and σ_k are the mean and standard deviation of that outcome for the control group.

In practice, $\mathbf{X}_{iv}$ includes cohort, survey year, gender, and baseline household assets (current cash, loans, and total cash assets), and we additionally include chiefdom fixed effects alongside the randomization-strata dummies δ_d; standard errors are clustered on community throughout. Household-level tables run on $N \approx 2{,}613$–$2{,}619$; the agent dataset has no pure control group, so agent-level effects (Table C.3, Panels B–C; Table 2, Panel A) are reported

relative to the Upgrade-alone arm rather than to control. Beyond the summary index above, each table also reports a Romano–Wolf (RW) step-down p-value (Romano and Wolf (2005)), which controls the family-wise error rate for multiple hypothesis testing (MHT) within each outcome family. These families do not always align one-to-one with a single table. We note where a family spans more than one (e.g., Table C.3 Panels B/C combine with Table 2 Panels A and B into a single 8-outcome family). The summary index and the RW correction serve complementary, not substitutable, roles. Aggregating several outcomes into a single index reduces the number of hypotheses we test in the first place. RW step-down correction instead controls the family-wise error rate across whatever family of hypotheses we ultimately report, including both the index and its components. We therefore report both throughout, rather than treating the index as a substitute for a formal MHT correction.

6.B. Mapping Theory to Parameters

Our coefficients of interest map directly to the stylized predictions in Section 4: Control corresponds to $(A = 0, \theta = 0)$, Upgrade alone to $(A = 1, \theta = 0)$, and Upgrade+Feedback to $(A = 1, \theta = \bar{\theta} > 0)$, so that β_1 identifies the pure Access Effect and $\beta_2 - \beta_1$ identifies the incremental effect of raising θ from 0 to $\bar{\theta}$ holding infrastructure access $A = 1$ fixed.

Access Effect (β_1): Captures the impact of infrastructure expansion alone. Proposition 1 predicts $\beta_1 > 0$ for extensive-margin usage (u_0), but limited gains in provider behavior and consumer protection (e.g., privacy improvements, e) in the absence of accountability.

Accountability Effect (β_2): Measures the impact of the combined intervention. Based on Proposition 2 (the Behavioral Shift prediction), we expect $\beta_2 > 0$ for provider effort as measured by operational and privacy protection improvements, reflecting the immediate "step-function" jump.

Complementary Effects: Proposition 3 predicts that infrastructure and accountability are complements, so that the marginal effect of feedback should be larger once access is already in place, i.e., $\beta_2 > \beta_1$. To formally test for this supermodularity, we perform a Wald test of the linear hypothesis $H_0 : \beta_2 = \beta_1$. Rejecting the null in favor of $\beta_2 > \beta_1$ provides empirical evidence that the feedback mechanism generates incremental value-added, confirming that accountability and infrastructure are complements in driving usage and subsequent downstream effects on business revenues. In practice, this complementarity shows up in our results along two distinct margins, both reported throughout Section 6: a direct test of the sharp ranking $\beta_2 > \beta_1$, which rejects in some outcome families but not most individual columns; and a more pervasive pattern in which β_2 survives Romano–Wolf multiple-hypothesis-testing correction while β_1 does not, even where the two point estimates are not statistically distinguishable from one another. We read the latter as the paper's

primary empirical signature of complementarity: feedback adds power and precision on top of upgrading even in the many cases where its point estimate is not sharply distinguishable from upgrading alone.

6.C. Upgrading vs Feedback and Provider Behavior

Figures 1 and 2 show agent (provider) density, providers' self-reported receipt of feedback, and providers' reported behavioral changes in response to feedback, across experimental arms. The *raw group means* version (Figure 1) plots unconditional treatment-arm means, with a confidence interval on every bar, including the reference/control bar. The *regression-based treatment effects* version (Figure 2) instead plots covariate-adjusted predictive margins from the specification used in Table C.3, again with a confidence interval on every bar.

Table C.3 presents our first main result. Panel A shows that upgrading substantially expands provider access: household-reported agent density rises significantly under both arms (Upgrade: +0.403 agents, or +74.6% relative to a control mean of 0.54 agents; Upgrade+Feedback: +0.508 agents, or +94.1%; both $p < 0.001$). This is driven by a near-quadrupling of agents accessible without leaving the community (+18.8 pp and +19.4 pp, respectively, off a 6% control base — one of the largest relative changes anywhere in the paper), alongside a corresponding 31–38% decline in reliance on agents outside the community. A summary KLK index of provider access rises by 0.354 SD under Upgrade alone and 0.406 SD under Upgrade+Feedback; a direct test that the two arms differ ($\beta_1 = \beta_2$) does not reject ($p = 0.424$), so on the pure access margin, feedback does not measurably add to what upgrading alone achieves.

Panels B–C show where feedback does add something upgrading alone cannot. There is no pure control group for agents: every agent surveyed is in a treated community, so we compare Upgrade to Upgrade+Feedback agents directly. Agents in Upgrade+Feedback communities are 21.1 pp more likely to report having received consumer feedback (against a 38% base rate in the Upgrade arm), and 20.5 pp more likely to report making behavioral changes in response (against a 24% base rate). Both are economically large shifts relative to their own comparison arm. This is the first evidence in our results that the feedback mechanism is not merely observed by providers but changes what they report doing. That is consistent with the Behavioral Shift prediction of Proposition 2 (Section 4.A).

6.D. Effort: Privacy Protection Practices and Consumer Trust

Table 2 tests whether providers' self-reported behavioral change (Table C.3, Panel C) corresponds to a measurable shift in privacy practice, and whether consumers notice. Panel

A shows agents in Upgrade+Feedback communities are 13.2 pp less likely to allow a third person to be present during a transaction[4] (significant at 5%, relative to Upgrade-alone agents) In the same direction, and also significant, they score 0.249 SD lower on a summary privacy-practice index. These practice items were fielded to, and compared directly across, both arms.

Panel B turns to consumers. Households in Upgrade+Feedback communities report 6.7–6.9 pp higher trust specifically in the mobile-money (Momo) agent under both arms (both $p < 0.05$). There is also a further, feedback-specific fall in trust in the local VSLA itself (-2.5 pp, $p < 0.05$). This is consistent with feedback sharpening rather than diffusely raising trust: it moves in the specific relationship it targets (the Momo agent), and it does not spill over to trust in family, friends, neighbors, or the banking agent, where no arm shows a significant effect. This specificity is itself informative. It is difficult to reconcile with a general Hawthorne-type "someone is watching us" story, and easier to reconcile with a channel that operates through the actual content of the feedback consumers give.

Panel A is, in effect, a direct measurement of the model's effort margin e. Section 4 defines e as the provider's costly privacy practice: observable to consumers and to us as researchers, but non-contractible, since no formal agreement can specify or enforce a target level of third-person exclusion or PIN discipline. This non-contractibility is exactly why the feedback mechanism θ, rather than a standard contract, is needed to move it. That Upgrade+Feedback agents shift privacy practice on precisely this non-contractible margin is the Behavioral Shift of Proposition 2.

As Section 4 notes, we take $T(e)$ in Equation (1) to be trust tied specifically to the provider's observed conduct, and the data bear this distinction out. Household-reported *general* trust in the Momo agent (Panel B) rises by a statistically indistinguishable margin under both arms (a direct test of $\beta_1 = \beta_2$ does not reject, $p = 0.953$). On its own, this is necessary but not sufficient evidence for $T(e)$: some baseline confidence in the agent appears to accompany upgrading regardless of accountability, consistent with u_0 in Equation (1).

The conduct-linked counterpart of $T(e)$ is instead visible in the round-by-round consumer-review data underlying Table C.3. The Core Review Index[5] and Overall Satisfaction are

[4]This figure, PIN Req., the KLK Privacy Index (Column 3 of this table), and Received Feedback and Made Changes in Table C.3 (Columns 5–6) all use the community-averaged definition of these agent-level outcomes: communities double-surveyed at endline in later cohorts are averaged across their agent records rather than keeping only the latest submission. The Lee bounds, trim percentages, and Imbens–Manski confidence sets for these same columns in Appendix Tables E.1 and E.2 are likewise constructed from the averaged definition.

[5]A six-item average of consumer-reported ratings: tariff disclosure, fee rating, privacy rating, no third person present, successful-transaction rating, and overall DFS-agent satisfaction; see Table D.1's note, Appendix D, for the exact construction.

ratings tied to consumers' actual, recent experience with that specific agent, and they jump between the first and second monthly review round before plateauing. This is the step-function shape Proposition 2 predicts, and that general trust does not display, though not every individual privacy item follows it as cleanly (Table D.1, Appendix D). Usage itself (see Table 4 below) tracks this same arm-specific pattern, rising significantly only where effort rose: the effort-usage complementarity of Proposition 3.

6.E. Community Financial Lives: Adoption vs Usage Outcomes

Tables 3 and 4 turn from provider behavior to household financial lives. They separate the adoption margin (does a household own a formal account) from the usage margin (does it actually transact). Table 3 shows adoption rises under both arms, more under Upgrade+Feedback. Formal DFS account ownership rises by 5.2 pp under Upgrade alone and 9.1 pp under Upgrade+Feedback (both against a 23% control mean). This rise is driven almost entirely by mobile-money accounts specifically. Upgrade+Feedback also raises the number of Momo accounts held (+0.179, $p < 0.01$) and a summary adoption KLK index (+0.202 SD, $p < 0.01$). A direct test of $\beta_1 = \beta_2$ does not reject for any single adoption outcome. But Upgrade+Feedback survives Romano–Wolf multiple-testing correction on every column at $p < 0.01$, while Upgrade alone does not always. Feedback adds statistical power even where the two point estimates are not distinguishable from one another.

Table 4 shows the same pattern, more sharply, on usage. Households in Upgrade+Feedback communities are 6.7 pp more likely to have used any formal financial service, and 8.2 pp more likely to have used DFS specifically, rising to 10.4 pp for mobile money usage specifically ($p < 0.01$). Upgrade alone shows the same sign throughout but never reaches significance after multiple-testing correction. The final column isolates the gap between adoption and usage directly (owning an account but not using it). Upgrade+Feedback narrows this gap by 6.2 pp ($p < 0.10$). This is the clearest evidence in our results that feedback's marginal contribution operates specifically on getting adopters to actually transact, not merely on getting more households to sign up. This adoption/usage split maps closely onto the model's two margins. Adoption tracks the extensive-margin baseline usage u_0 that Proposition 1 attributes to access alone. The adoption-usage gap, by contrast, tracks the trust-driven intensive margin $\tau T(e)$ in Equation (1), which only rises once accountability moves effort e (Proposition 2). This is the same conduct-linked trust channel that Section 6.D's round-by-round agent-rating data, rather than the general household trust measure, tracks directly.

Table 5 reports this intensive-margin usage gain at the household level, as a household-level average treatment effect with household covariates included as controls. A natural concern with any household-level average is whether it reflects a genuine per-household be-

havior change, or simply a compositional difference in which households get surveyed, or how many, across arms, rather than a real treatment effect. Appendix Tables C.4 and C.5 address this directly by aggregating the same five outcomes to the community level, first as totals and then as averages. Both reproduce the identical qualitative pattern as Table 5: Upgrade+Feedback is significant on P2P Value, P2P Attempts, and the KLK Index, and nothing else is distinguishable from zero. This pattern is robust even after explicitly controlling for the number of surveyed households per community, which, as those tables' note documents, differs meaningfully between Control communities (16.8 households on average, up to 35) and the two treatment arms (a tight ∼14.9 each, not distinguishable from one another). The same pattern holds whether communities are summed, averaged, or explicitly adjusted for this household-count imbalance. That is evidence that Table 5's household-level result is a genuine treatment effect, rather than an artifact of how many, or which, households each arm happens to contribute.

Beyond ruling out that artifact, the intensive margin itself is substantively important. It is the direct empirical counterpart to the model's trust-driven margin $\tau T(e)$ in Equation (1). It also shows that feedback's usage effect does not stop at inducing a first transaction, but continues to raise transaction value and frequency among households already engaged. Peer-to-peer transfer value and frequency are the two outcomes that survive Romano–Wolf correction most robustly in Table 5: they rise by 222% and 303% relative to control, respectively. Non-cash transfer activity overall (Sent+Recv., Columns 6–7) also rises, by 126% under Upgrade+Feedback ($p < 0.10$); this is the measure used in the cost-effectiveness exercise in Section 7.E. Peer-to-peer activity is one of the fee-bearing transaction types agents and VSLAs earn revenue on, discussed below, so this same intensive-margin channel also feeds directly into the VSLA business-income spillovers documented in Section 6.F below. Appendix Tables C.4 and C.5 report no Romano–Wolf p-values of their own, because they re-test these same five outcomes, already corrected for multiple comparisons once in Table 5, at a different (community) level of aggregation.

Table 6 decomposes this usage gain into four specific transaction channels: cash-in deposits, cash-out withdrawals, peer-to-peer transfers, and savings instrument choice. The pattern is not uniform across channels. Savings mobilized through a DFS instrument rise significantly under both arms (Upgrade: +7.8 pp; Upgrade+Feedback: +6.6 pp), and both survive Romano–Wolf correction under the five-outcome family (MHT $p = 0.001$ and $p = 0.005$, respectively). Savings held through informal instruments, by contrast, fall in both arms, though not significantly. Cash-out withdrawals rise significantly only under Upgrade+Feedback (+9.2 pp, $p < 0.05$), while cash-in deposits are significant only for Upgrade alone (+4.2 pp, $p < 0.10$). This pattern is more consistent with noisy estimation around

a common underlying effect than with a channel-specific mechanism. Peer-to-peer transfers are the one channel where the two arms are directly distinguishable from one another ($p = 0.010$): Upgrade+Feedback raises P2P usage by 4.6 pp, against a negligible and insignificant 1.1 pp under Upgrade alone. Taken together, Table 6 shows that the aggregate usage gain in Table 4 is not driven by any single transaction type. It instead reflects broad-based increases in digital engagement across deposits, withdrawals, and transfers, with peer-to-peer activity the clearest single channel where feedback adds something upgrading alone does not.

This channel pattern lines up with how our partner DFS provider actually prices its services. Orange Money Sierra Leone's published tariff schedule charges no fee on cash-in (deposit) transactions, while both cash-out withdrawals and peer-to-peer transfers carry a fee (Orange Sierra Leone 2026, p. 28).[6] An agent's conduct, such as charging the correct tariff, not overcharging, or being transparent about pricing, has no scope to matter on a transaction that carries no fee in the first place. So any accountability-driven change in trust or usage should concentrate in the fee-bearing channels rather than cash-in. This is exactly the pattern Table 6 shows. The two channels where our treatment effects are most robust, and where Upgrade+Feedback distinguishes itself from Upgrade alone, are cash-out and peer-to-peer. These are precisely the fee-bearing transaction types, while cash-in, the one channel that is free regardless of agent conduct, shows the weakest and least differentiated treatment effect. This is also consistent with the VSLA-income results in Section 6.F. Revenue to the agent, and in turn to the VSLA that finances and stands behind that agent, is earned on cash-out and P2P activity rather than cash-in. The usage shift our intervention produces along exactly those channels is therefore a plausible channel through which upgrading and feedback raise VSLA incomes.

6.F. Downstream Spillover Effects: VSLA Business and Household Businesses

These usage gains are the empirical counterpart to the model's provider-side revenue function $R(u) = \gamma u$. Because u and e enter as complements in the provider's problem (Proposition 3), higher usage driven by accountability should feed back into higher reported VSLA business activity, not only higher household-level usage. Table 7 asks whether upgrading and feedback also changed the VSLAs' own financial functioning as organizations, downstream of individual adoption and usage. Reported VSLA business income rises by 159.7 SLL under Upgrade+Feedback (against a control mean of 18.5 SLL, $p < 0.10$), and the share of member contributions made digitally rises by 6.9 pp (against an essentially zero control mean, $p < 0.01$). This is the clearest evidence that upgrading changes not just how households transact individually, but how the VSLA itself operates.

[6] https://www.orange.sl/en/orange-money-tariffs.html

Although the income estimate does not survive the five-outcome Romano–Wolf correction reported below ($p = 0.409$) and should be read as suggestive rather than definitive — unsurprising given it is identified from only 172 VSLA-secretary observations, an order of magnitude smaller than our household-level samples (around 2,600–2,900 observations) — the relative magnitudes across arms are informative in their own right. Upgrade+Feedback's effect on reported business income (159.7 SLL) is more than three times the size of Upgrade alone (51.6 SLL, itself not distinguishable from zero), though we cannot formally reject equality between the two coefficients at conventional levels ($p = 0.239$). Read as a comparison of magnitudes rather than of statistical significance, this pattern is consistent with the accountability channel introduced by feedback, rather than digitization alone, driving whatever VSLA-level income effect is present.

On member-contribution discipline, meaning whether members made their required contributions to the group on time rather than loan repayment per se, the overall default-rate effect (Column 5) is negative for Upgrade alone ($\beta_1 = -0.55$) and small and positive for Upgrade+Feedback ($\beta_2 = 1.20$); neither is significant. This null average, however, masks substantial heterogeneity. Splitting VSLAs by pre-experiment default risk (fewer than four members currently in default or delayed contribution, versus four or more) shows that Upgrade+Feedback significantly reduces the number of members in default among previously low-risk VSLAs (-0.72, $p < 0.05$), with no significant effect among previously high-risk VSLAs. The default-risk-split columns are not part of the five-outcome Romano–Wolf family reported below (business income, current members, contribution amount, percent digital contribution, and members defaulted). None of those five outcomes survives that correction, so we read the business-income, digitization, and default-risk results as suggestive.

One way to read this concentration of the default-reduction effect among previously low-risk VSLAs is through the lens of what a "default" actually captures. A member's missed or delayed contribution can be mechanical: she has the funds but fails to contribute on time because she travelled outside the village, because of bad weather, or for some other logistical reason unrelated to intent. Or it can be strategic, where a member withholds a contribution to free-ride on the group despite having the funds available. Digitizing VSLA contributions plausibly resolves the mechanical component specifically. A digitized record and reminder system directly addresses missed-but-intended contributions, the kind of lapse a previously low-risk VSLA is disproportionately likely to be experiencing. A persistently high-risk VSLA's defaults, by contrast, more plausibly reflect deeper strategic non-compliance: free-riding that better record-keeping alone cannot resolve, since it requires social or contractual enforcement rather than better information. This is consistent with digitization's measured effect landing entirely on previously low-risk groups, and being essentially null

among previously high-risk ones. Taken together, these results suggest that digitizing VSLAs and other savings groups is unlikely to resolve strategic moral-hazard constraints. But it does have the potential to be transformative in resolving the mechanical challenges to group financial discipline. This is precisely the pattern implied by the low-risk-concentrated result above, and a distinction that, to our knowledge, has not been drawn out in the prior literature on digitizing informal savings groups.

6.G. Effects on Shock Mitigation

Our theoretical framework (Section 4) predicts that higher provider effort and usage should eventually translate into household-level welfare gains, through lower default rates and higher group-level income (Proposition 3, the Complementary Effects prediction: Proof 3B formalizes this as an Efficiency Effect operating through the default-rate channel $\delta(e)$, alongside a Usage Effect). Section 6.E documents the paper's most robust household-level usage gains under Upgrade+Feedback: higher DFS adoption broadly (Table 4, Column 2), higher mobile-money (Momo) usage specifically (Table 4, Column 3), and higher peer-to-peer financial transactions (Table 6, Column 3; MHT $p = 0.001$). A natural follow-on question is whether these usage gains were large enough for households to draw on when something went wrong: did higher financial-service usage translate into fewer, or less severe, negative income and wealth shocks? Table 8 reports treatment effects on four household-level shock measures, all measured over the 90 days before the endline survey: any negative income or wealth shock (Column 1); the subset of shocks the household reports being unable to mitigate, i.e. where its own coping resources or financial resilience proved insufficient and resource constraints bound (Column 2, "Unmitigated Shock"); a shock index averaging the two (Column 3, "Shock Index (Mean)"); and a shock index equal to 1 if either measure is 1 (Column 4, "Shock Index (Any)").

Two things stand out. First, the point estimates are directionally consistent across all four ways of measuring a shock. Upgrade+Feedback reduces shock exposure by an estimated 3.8–4.6% of the control-arm mean on every column (Column 1: $-0.036/0.777 \approx -4.6\%$; Column 2: $-0.015/0.399 \approx -3.8\%$; Column 3: $-0.025/0.588 \approx -4.3\%$; Column 4: $-0.029/0.781 \approx -3.8\%$), while Upgrade alone is small, inconsistent in sign across columns, and never distinguishable from zero. Second, the two arms differ from one another at conventional levels on three of the four measures, and at the 10% level on the fourth (Column 1: $p = 0.087$; Column 2: $p = 0.069$; Column 3: $p = 0.035$; Column 4: $p = 0.037$). This pattern is robust to exactly how the shock variable is constructed.

Neither individual coefficient, however, survives Romano–Wolf correction against the 14-outcome family described in the table note (all MHT $p \geq 0.53$). The arm-difference itself,

once corrected the same way, falls to $p = 0.045$ for the shock index and $p = 0.070$–0.242 for the other three columns; that correction re-expresses $\beta_2 - \beta_1$ as its own testable coefficient and steps it down jointly with the same 14-outcome family. Because Upgrade alone holds the market's underlying infrastructure fixed while Upgrade+Feedback adds structured consumer feedback on top of that same infrastructure, the arm difference $\beta_2 - \beta_1$ isolates the market-discipline channel specifically, net of the infrastructure upgrade itself. Read this way, we take the results above as suggestive evidence that market discipline, not infrastructure alone, is what moves shock exposure. This is directionally consistent with, but not on its own sufficient to establish, the usage-driven mechanism of Section 6.E. These estimates provide an objective proxy for the financial resilience and insurance value that market discipline delivers by aligning provider behavior with consumer welfare. We do not yet have the statistical power, in this single endline round, to confirm that value at the level of individual outcomes. We do not find evidence for an impact on overall poverty levels.

Upgrading alone produces gains in agent density and household DFS adoption, with Upgrade + Feedback consistently adding further gains on top of Upgrading alone. This shows up most clearly in providers' receipt of and response to feedback (Table C.3, Panels B–C), consumer trust in the mobile-money agent (Table 2, Panel B), DFS adoption and usage (Tables 3–6), and VSLA-business spillovers (Table 7). Across nearly every outcome family, β_2 (Upgrade+Feedback) survives Romano–Wolf multiple-hypothesis-testing correction more consistently than β_1 (Upgrade alone), even where the two arms are not statistically distinguishable from one another by a direct test. This is consistent with feedback adding power on top of upgrading alone.

7. Interpreting the Results and Scalability

Our working interpretation of the results in Section 6 is that accountability—building provider–customer relationships that align provider incentives with consumer welfare—rather than infrastructure alone, drives the gains we document. Upgrading alone expands access but generates limited operational or usage gains on its own; adding structured consumer feedback consistently produces larger, more robust improvements in provider behavior, trust, usage, and downstream business outcomes (Section 6). We probe this interpretation from several directions below: whether it instead reflects gradual provider learning rather than accountability (Section 7.A), whether it is an artifact of differential survey attrition (Section 7.B) or of spillovers across nearby markets (Section 7.C), and whether it could instead reflect Hawthorne or demand effects rather than genuine changes in provider behavior (Section 7.D), before turning to this intervention's cost-effectiveness and prospects for scale (Section

7.E).

7.A. Learning vs. Accountability

This is a direct empirical test of Proposition 2's predicted step-function shape (Appendix B, Proof 2B) against the smooth, decelerating path a learning story would predict. One might argue that the operational improvements reflect provider learning, i.e., "I didn't know I shouldn't ask for PINs" as opposed to accountability, i.e., "It's easier to ask for the PIN, but now I'll get a bad rating". However, three pieces of evidence suggest an incentive-based mechanism. First, the behaviors we track—such as respecting transaction privacy and ensuring fee transparency—are fundamental norms of financial intermediation rather than complex technical tasks requiring specialized training. Second, the immediacy and magnitude of the shift (+20.5 pp) suggests that the feedback mechanism did not merely provide new information, but rather increased the shadow of future consequences, inducing higher levels of effort that were previously latent. Finally, the month-to-month dynamics of the feedback loop, focusing on the Core Review Index and Overall Satisfaction outcomes, are linear, rather than following a concave learning curve: the round-2-to-3 change is statistically indistinguishable from, or smaller than, the round-1-to-2 change, rather than the smaller-but-still-positive gain a decelerating learning curve predicts (Core Review Index, round-to-round difference −0.052, SE 0.019, significant at the 1 percent level; Table D.1, Column 6); this suggests that agents are responding to constant monitoring pressure and maintaining effort in anticipation of the next evaluation. Appendix D presents the full round-by-round consumer-review data and the formal accountability-versus-learning test underlying this claim (Table D.1), including targeted follow-up analyses of within-arm targeting and open-ended complaint text.

7.B. Survey Attrition

Section 5.C established that residual differential attrition, after correcting a community-name linkage problem in the baseline-to-endline match, remains modestly higher in the treatment arms than in control (roughly 6–10 pp). We ask directly whether this residual attrition, on its own, could be driving the positive treatment effects reported in Tables 2–7, using Lee (2009) trimming bounds, which construct the most pessimistic and most optimistic treatment effects consistent with the observed attrition rates without requiring any assumption about *why* some VSLAs are missing from the endline sample. Across roughly eighty coefficient-level bounds computed this way, essentially every starred coefficient in Tables 2–7 has a Lee-bounds interval that excludes zero and shares its sign with the main estimate; residual attrition is therefore not large enough on its own to manufacture nearly any of our positive,

significant treatment effects at the point-identified level. A small number of already-weak (10%-significance, Romano–Wolf-fragile) coefficients do not survive this check, and a stricter Imbens and Manski (2004) confidence-set construction—which additionally accounts for sampling noise in the bounds themselves—further narrows the set of individual coefficients we can consider fully immune to attrition, concentrated in Table 2's consumer-trust items and a handful of Table 4's usage columns; our core adoption (Table 3), usage-channel (Table 6), and count-based default-risk (Table 7) results are robust to both checks in full. We view this exercise as a complementary, assumption-light check layered on top of our covariate-adjusted specifications, not a substitute for them. Appendix E presents the full battery of Lee (2009) bounds and Imbens and Manski (2004) confidence sets, computed coefficient-by-coefficient across Tables 2–7.

7.C. Cross-Market Spillovers

We also consider whether our results are robust to spillovers from treated markets onto nearby control markets—for instance, a control-arm household traveling to a neighboring treated community to use its new DFS agent. If this were happening at meaningful scale, control-arm outcomes should vary systematically with how much treatment intensity surrounds them, biasing our simple treatment-versus-control comparisons toward zero. We test this by grouping markets at the *chiefdom* level (the administrative tier directly beneath district; our three districts contain eight chiefdoms between them, each with its own realized treated share), constructing for each community i in chiefdom c a leave-one-out treated-market share among other communities in its own chiefdom, and interacting this with i's own treatment status:

$$y_i = \beta_0 + \beta_1 \text{Treated}_i + \beta_2 \,\text{ChiefdomTreatedShare}_i + \beta_3 \,\text{Treated}_i \times \text{ChiefdomTreatedShare}_i + \text{strata}_i + \text{cohort}_i + \varepsilon_i, \quad (4)$$

with randomization-strata and cohort fixed effects and standard errors clustered at the community level. Since $\text{Treated}_i = 0$ for control units, β_2 in Equation 4 is exactly the marginal association between a control unit's own outcome and how treatment-saturated its chiefdom is—the placebo test of interest.

Run across all 37 household- and secretary-level outcomes in Tables 2–7, and the four shock-mitigation measures in Table 8 (agent-level outcomes have no control arm and are not amenable to this check), both without and with an added district fixed effect, the pattern is reassuring: 36 of 37 show no significant relationship between chiefdom-level treatment saturation and control-group outcomes at conventional levels, in either specification—this

includes all four shock-mitigation measures, none of which shows a significant $\hat{\beta}_2$ at conventional levels in either specification (the closest, Any Shock, is only marginal without the district fixed effect, $p = 0.079$, and loses even that once the district fixed effect is added, $p = 0.175$). One outcome is the exception—the Adopt-but-Don't-Use gap (Table 4)—which shows a positive, significant coefficient both without a district fixed effect (0.719, $p = 0.009$) and, if anything more strongly, with one added (0.761, $p = 0.003$); because chiefdoms nest inside districts, surviving the addition of a district fixed effect is a genuine test, not a near-tautology, so this one result is a candidate for genuine chiefdom-level spillover rather than an artifact of coarse geography. It should be read in context: this outcome is already one of our less robust columns on its own terms (Section 7.B), its estimation sample is restricted to DFS-owning households and is accordingly smaller than the other columns, and a single significant result out of 37 checked outcomes is within the range expected from chance alone. It does not overturn any headline conclusion in this paper, but is the one outcome where this check cannot rule spillover out. Appendix F presents the full chiefdom-level placebo regressions, run across all 37 household- and secretary-level outcomes in Tables 2–7, and Table 8, both with and without a district fixed effect.

7.D. Hawthorne and Demand Effects

Beyond survey attrition (Section 7.B) and cross-market spillovers (Section 7.C), a third standard threat to interpreting these treatment effects as the intervention's true causal impact is that respondents or providers may have changed their behavior simply because they knew they were being studied or knew which arm they were in—either performing for the evaluator (a Hawthorne effect, most concerning in the treatment arms) or compensating out of a sense of unfair comparison (a John Henry effect, most concerning in the control arm). Several design features limit both concerns. The treatment is a physical, structural change—installing or formalizing a DFS/mobile-money agent relationship for a VSLA—not a motivational message a respondent could simply decide to perform without the underlying infrastructure actually being present, and all three arms were surveyed with the identical endline instrument by the same enumerator pool. Enumerators were also kept blind to treatment status during data collection, closing off the channel by which an enumerator might otherwise cue an "expected" response. If simple study awareness were driving our results, the Upgrade-alone arm—which received identical survey attention to Upgrade+Feedback— should show similarly inflated effects across the board; instead Upgrade-alone is frequently insignificant while Upgrade+Feedback is not (Tables 3–6), tracking the added feedback mechanism rather than mere awareness of being studied. On John Henry effects specifically, a Control VSLA cannot simply try harder to become a formal DFS agent site absent the equipment and agent

relationship the study did not provide it, and consistent with this, the endline data show no instance of a Control-arm VSLA having become a mobile money or banking retail agent by endline—unsurprising given the structural nature of the intervention, but a useful direct check against this concern nonetheless. Randomization at the VSLA/community level, rather than within a shared community, reinforces this: Control and Treatment VSLAs sit in physically different villages, so a Control respondent has no obvious treated peer group to measure themselves against day to day, reducing the salience a competitive response would need. These are plausibility arguments rather than a statistical test, and are strongest for the harder, transaction-based outcomes (Tables 3–7) and weakest for the softer, attitudinal trust measures (Table 2); Appendix G lays out this reasoning in full, including where its limits are.

7.E. Cost-Effectiveness and Scalability

Beyond statistical significance, two further considerations bear directly on this intervention's policy relevance: its cost-effectiveness relative to alternative approaches to expanding rural financial access, and its prospects for scale.

Cost-Effectiveness. The relevant comparison weighs the intervention's implementation costs—agent onboarding and training, and, for the Upgrade+Feedback arm, the recurring costs of collecting and disseminating monthly consumer evaluations—against the effect sizes documented above. Following a conservative approach used in Annan (2025), we focus on two outcomes: (i) mobile-money usage, one of our most robust market-side outcomes (Table 4, Column 3), and (ii) VSLA business income, one of our most robust downstream spillover outcomes (Table 7, Column 1). This is a deliberately narrow accounting that does not consider the intervention's other significant treatment effects (formal account ownership, provider trust and consumer protection, digitization of contributions, default risk, and shock mitigation) or its positive externalities to non-sample community members.

We first compute the total cost of the Upgrade intervention to be \$16,087, drawn from the IPA–Sierra Leone subaward budget to CAFOD (the Community Engagement Meetings line item funding onboarding/enrollment), and the total cost of Upgrade+Feedback to be \$30,086 (adding the recurring cost of monthly consumer-feedback calls: approximately 10 enumerators per round at \$400 each over 3 rounds, plus the 10 SLL airtime credit paid to each of the 4,192 respondents across the three rounds, valued here as the opportunity cost of participating in a brief, under-10-minute phone call). Spread across the relevant respondent population (Table A.1), this implies a cost of \$85 (Upgrade) and \$159 (Upgrade+Feedback) per agent, \$89 and \$166 per VSLA, and \$5.58 and \$10.43 per household. Full back-of-the-envelope detail, including per-respondent-type cost-effectiveness ratios and an aggregate

return by treatment arm, is reported in Appendix H.5.

For provider behavior, Upgrade+Feedback's $159-per-agent cost buys a 21.1 pp increase in agents reporting having received consumer feedback (Table C.3, Column 5; significant at the 5 percent level) – $7.54 per pp. Agents are separately 20.5 pp more likely to report having made changes in response to that feedback (Table C.3, Column 6). For VSLA business income (Table 7, Column 1) – not statistically distinguishable from zero for Upgrade alone, and significant only at the 10 percent level before Romano–Wolf correction for Upgrade+Feedback – the annualized treatment effect implies a 144% return under Upgrade and a 238% return under Upgrade+Feedback. For household usage, valuing the increase in non-cash transaction activity directly in dollars – the average of total non-cash digital transactions sent and received over the past 90 days (Table 5, Column 7; not statistically distinguishable from zero under Upgrade alone, but significant at the 10 percent level under Upgrade+Feedback – Column 7 is a derived/composite measure and is not part of that table's Romano–Wolf family) implies a 132.1% annualized return under Upgrade and 85.1% under Upgrade+Feedback.

Aggregating the dollar-valued VSLA income and household usage-value effects into a single annualized return – a deliberately conservative accounting that excludes the agent-level feedback-behavior effect, which has no direct dollar value, as well as the intervention's other significant effects on formal account ownership, consumer trust and protection, and shock mitigation, and any positive externalities to non-sample community members – yields an overall return of 276.1% under Upgrade and 323.4% under Upgrade+Feedback. As the significance notes above make clear, most of the point estimates underlying these back-of-envelope returns are only marginally significant, or not significant at all, once adjusted for multiple comparisons; Appendix H.5 reports each effect's standard error, raw significance stars, and (where available) its Romano–Wolf p-value, and these figures should be read as illustrative back-of-envelope calculations rather than as uniformly robust effect sizes.

These rough, illustrative estimates are not easily benchmarked against the existing literature, because that literature has generally not been able to complete the same exercise. Karlan et al. (2016), evaluating savings-group programs in Ghana, Malawi, and Uganda at a broadly similar per-household cost ($7.01 in Malawi and $10.85 in Uganda, versus $5.58–$10.43 here; cost data were not available for Ghana), find no significant effect on household income or food security, and do not report a formal benefit-cost ratio. Batista and Vicente (2025), the closest existing field experiment on digital-financial-services expansion itself, report a per-participant cost of about $3 (76 Meticais in free trial mobile-money credit) but likewise run no cost-effectiveness or benefit-cost calculation. For both the closest savings-group comparison and the closest digital-financial-services-expansion comparison, then, a

benefit-cost analysis for this class of intervention is something the existing literature has not yet produced, not a benchmark our results fall short of. Our setting differs in generating a measurable, dollar-valued usage effect to set against program cost in the first place, even though the resulting back-of-envelope return figures should still be read as illustrative rather than as a settled welfare comparison. Our findings suggest that upgrading existing informal financial institutions with structured consumer feedback could be a cost-effective way to improve local financial markets.

Scalability. This intervention's design—leveraging existing VSLA infrastructure and local reputation rather than building new distribution channels from scratch—was intended from the outset to be low-cost and readily replicable. Consistent with the descriptive cost-effectiveness case made above, our implementation partner, CAFOD, judged the net gains from the Upgrade+Feedback arm sufficiently compelling to begin scaling the intervention beyond this study's original footprint: as of March 2026, CAFOD is expanding the Upgrade+Feedback model to additional villages across Sierra Leone, independent of this evaluation. We view this real-world take-up as itself informative evidence of the intervention's practical scalability, complementing the experimental estimates above. More broadly, the VSLA chassis this intervention upgrades is not a niche institution: CARE's own VSLA network alone spans 536,977 groups and 29,889,968 members across 67 countries as of fiscal year 2025 (CARE International (2025)). Catholic Relief Services' own savings-led microfinance program, built primarily around Savings and Internal Lending Communities (SILC) groups, similarly reports 1.4 million program participants across 64 projects in 36 countries in fiscal year 2025 (Catholic Relief Services (2025)). World Vision's Savings for Transformation program reports over 79,000 savings groups and 1.6 million members across multiple countries in Africa, the Americas, Asia Pacific, and the Middle East/Europe (World Vision International, n.d.). An upgrading model that works through existing VSLAs or VSLA-type networks, rather than building new distribution infrastructure, therefore has a very large existing base to scale onto beyond Sierra Leone.

8. Conclusion

This paper studies how upgrading financial markets, through both market creation and market discipline, shapes the development of a newly created market. Using a randomized supply-side intervention cross-randomized with structured consumer feedback in rural Sierra Leone, we show that feedback drives large gains in agent privacy and consumer protection (+20.5 pp, 85%), raising both usage and provider revenues. While Wieser et al. (2019) find that expanding mobile-money agent access alone generates little change in rural market

activity, we show that pairing access with structured consumer feedback is what makes a newly created market function, and that doing so is win-win-win: consumers get better service, the VSLAs that capitalize and stand behind their agents perform better, and the commercial providers that supply them see more transaction activity.

One limitation is worth noting upfront: our results come from a single post-intervention endline, as with many field experiments, so we cannot speak to whether these market-creation and market-discipline gains persist, deepen, or fade over a longer horizon. While we expect even larger effects based on the nature of our intervention, future work that follows these communities further out could track a broader set of structural household welfare outcomes, beyond the market design and development outcomes we study here.

These results offer market design lessons for designing scalable, inclusive financial intermediaries. Because each DFS agent is financially backed by the VSLA that fields it, accountability can be built into existing community structures rather than imposed from outside: pooling monitoring at the VSLA level, rather than asking a provider to monitor each low-volume rural agent individually, is what makes a feedback mechanism worth operating in a thin market. This is a transferable design principle for intermediaries expanding into underserved markets at low fixed cost.

Beyond market-level effects, structured feedback also improves household welfare through shock mitigation: households in Upgrade+Feedback communities are 3.8–4.6% of the control-arm mean less likely to report a shock they could not financially remedy. This is consistent with prior evidence that digital financial access strengthens households' ability to cope with shocks (Jack and Suri 2014; Suri and Jack 2016; Annan 2025). We find no effect on aggregate poverty, which this design was not powered to resolve, but the shock-mitigation result bears directly on policy debates about scaling financial intermediaries for the poor: accountability mechanisms are not just a service-quality add-on, but a lever on welfare itself.

More broadly, our results speak to the underappreciated role of quality in tech-enabled services adoption. Account ownership and active use remain persistently disconnected, a central challenge for inclusive finance worldwide (Demirguc-Kunt et al. 2022; Klapper et al. 2025; Karlan et al. 2014; Prina 2015). This gap is often attributed to path dependence in how early-stage markets settle (David 1991; Katz and Shapiro 1986; Krugman 1994). We show this gap is not fixed at entry: design choices that build in trusted institutions, provider incentives, and consumer voice causally shape service quality, and quality is what converts access into adoption. To our knowledge, this is among the first studies to experimentally separate the market-creation effects of financial infrastructure from the effects of pairing that infrastructure with structured consumer accountability, and to show that the two are complements, not substitutes, in whether new markets deliver meaningful gains in efficiency,

usage, and welfare.

References

[1] Annan, Francis. 2025. "Misconduct and Reputation under Imperfect Information." *Journal of Political Economy*, 133(5): 1460–1496.

[2] Banerjee, Abhijit V., Emily Breza, Arun G. Chandrasekhar, Esther Duflo, Matthew O. Jackson, and Cynthia Kinnan. 2024. "Changes in Social Network Structure in Response to Exposure to Formal Credit Markets." *Review of Economic Studies*, 91(3): 1331–1372.

[3] Barboni, Giorgia, Erica Field, and Rohini Pande. 2025. "Rural Banks Can Reduce Poverty: Experimental Evidence from 870 Indian Villages." Yale Economic Growth Center Working Paper.

[4] Batista, Cátia, and Pedro C. Vicente. 2025. "Is Mobile Money Changing Rural Africa? Evidence from a Field Experiment." *Review of Economics and Statistics*, 107(3).

[5] Blackmon, William, Rafe Mazer, and Shana Warren. 2021. "IPA Consumer Protection Research Initiative: RFP Overview." Innovations for Poverty Action.

[6] Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200–232.

[7] Burgess, Robin, and Rohini Pande. 2005. "Do Rural Banks Matter? Evidence from the Indian Social Banking Experiment." *American Economic Review*, 95(3): 780–795.

[8] CARE International. 2025. "VSLA Global Annual Report 2025." https://www.care.org/wp-content/uploads/2025/10/VSLA-Annual-Report-2025_EN.pdf, accessed October 2026.

[9] Catholic Relief Services. 2025. "CRS Facts and Figures, Fiscal Year 2025." https://www.crs.org/sites/default/files/documents/2026-06/CRS-Facts-Figures-FY25-EN.pdf, accessed October 2026.

[10] Cilliers, Jacobus, Isaac M. Mbiti, and Andrew Zeitlin. 2021. "Can Public Rankings Improve School Performance? Evidence from a Nationwide Reform in Tanzania." *Journal of Human Resources*, 56(3).

[11] David, Paul A. 1991. "Computer and Dynamo: The Modern Productivity Paradox in a Not-Too-Distant Mirror." In *Technology and Productivity: The Challenge for Economic Policy*, Paris: OECD.

[12] Demirguc-Kunt, Asli, Saniya Ansar, Leora Klapper, and Dorothe Singer. 2022. "The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19." Washington, DC: World Bank.

[13] Deserranno, Erika, Gianmarco León-Ciliotta, and Firman Witoelar. 2019. "The Supply and Demand Effects of Financial Incentives: Experimental Evidence from Branchless

Banking in Indonesia." Working paper.

[14] Fradkin, Andrey, and David Holtz. 2022. "Do Incentives to Review Help the Market? Evidence from a Field Experiment on Airbnb." *Marketing Science*, forthcoming.

[15] Francois, Patrick, and Munir Squires. 2021. "Linking Mobile Money Networks to 'e-ROSCAs': An Experimental Study." *Science Advances*, 7(1): eabc5831.

[16] Garz, Seth, Xavier Giné, Dean Karlan, Rafe Mazer, Caitlin Sanford, and Jonathan Zinman. 2021. "Consumer Protection for Financial Inclusion in Low- and Middle-Income Countries: Bridging Regulator and Academic Perspectives." *Annual Review of Financial Economics*, 13: 219–246.

[17] Imbens, Guido W., and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica*, 72(6): 1845–1857.

[18] Jack, William, and Tavneet Suri. 2014. "Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution." *American Economic Review*, 104(1): 183–223.

[19] Karlan, Dean, Aishwarya Lakshmi Ratan, and Jonathan Zinman. 2014. "Savings by and for the Poor: A Research Review and Agenda." *Review of Income and Wealth*, 60(1): 36–78.

[20] Karlan, Dean, Beniamino Savonitto, Bram Thuysbaert, and Christopher Udry. 2016. "Impact of Savings Groups on the Lives of the Poor." *Proceedings of the National Academy of Sciences*, 114(12): 3079–3084.

[21] Katz, Michael L., and Carl Shapiro. 1986. "Technology Adoption in the Presence of Network Externalities." *Journal of Political Economy*, 94(4): 822–841.

[22] Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

[23] Klapper, Leora, Dorothe Singer, Laura Starita, and Alexandra Norris. 2025. "The Global Findex Database 2025: Connectivity and Financial Inclusion in the Digital Economy." Washington, DC: World Bank.

[24] Krugman, Paul. 1994. *Peddling Prosperity: Economic Sense and Nonsense in an Age of Diminished Expectations*. New York: W.W. Norton.

[25] Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, 76(3): 1071–1102.

[26] Mbiti, Isaac, and Danila Serra. 2022. "Health Workers' Behavior, Patient Reporting, and Reputational Concerns: Lab-in-the-Field Experimental Evidence from Kenya." *Experimental Economics*.

[27] Morduch, Jonathan, and Dean Karlan. 2009. "Access to Finance." In *Handbook of Development Economics*, Vol. 5, edited by Dani Rodrik and Mark Rosenzweig. Amsterdam: North-Holland.

[28] Orange Sierra Leone. 2026. "Orange Money Tariffs." https://www.orange.sl/en/orange-money-tariffs.html, accessed October 2026.

[29] Pallais, Amanda. 2014. "Inefficient Hiring in Entry-Level Labor Markets." *American Economic Review*, 104(11): 3565–3599.

[30] Prina, Silvia. 2015. "Banking the Poor via Savings Accounts: Evidence from a Field Experiment." *Journal of Development Economics*, 115: 16–31.

[31] Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica*, 73(4): 1237–1282.

[32] Schreiner, Mark. 2018. *Simple Poverty Scorecard® Poverty-Assessment Tool: Sierra Leone.* Microfinance Risk Management, L.L.C.

[33] Suri, Tavneet, and William Jack. 2016. "The Long-Run Poverty and Gender Impacts of Mobile Money." *Science*, 354(6317): 1288–1292.

[34] Suri, Tavneet, et al. 2023. "Mobile Money." *VoxDevLit*, 2(1).

[35] Wang, Shing-Yi. 2022. "Improving Management Through Worker Evaluations: Evidence from Auto Manufacturing." *Quarterly Journal of Economics*, 137(4): 2459–2497.

[36] Wieser, Christina, Miriam Bruhn, Johannes Kinzinger, Christian Ruckteschler, and Soren Heitmann. 2019. "The Impact of Mobile Money on Poor Rural Households: Experimental Evidence from Uganda." World Bank Policy Research Working Paper No. 8913.

[37] World Vision International. N.d. "Savings for Transformation." https://www.wvi.org/livelihoods/savings-transformation, accessed October 2026.

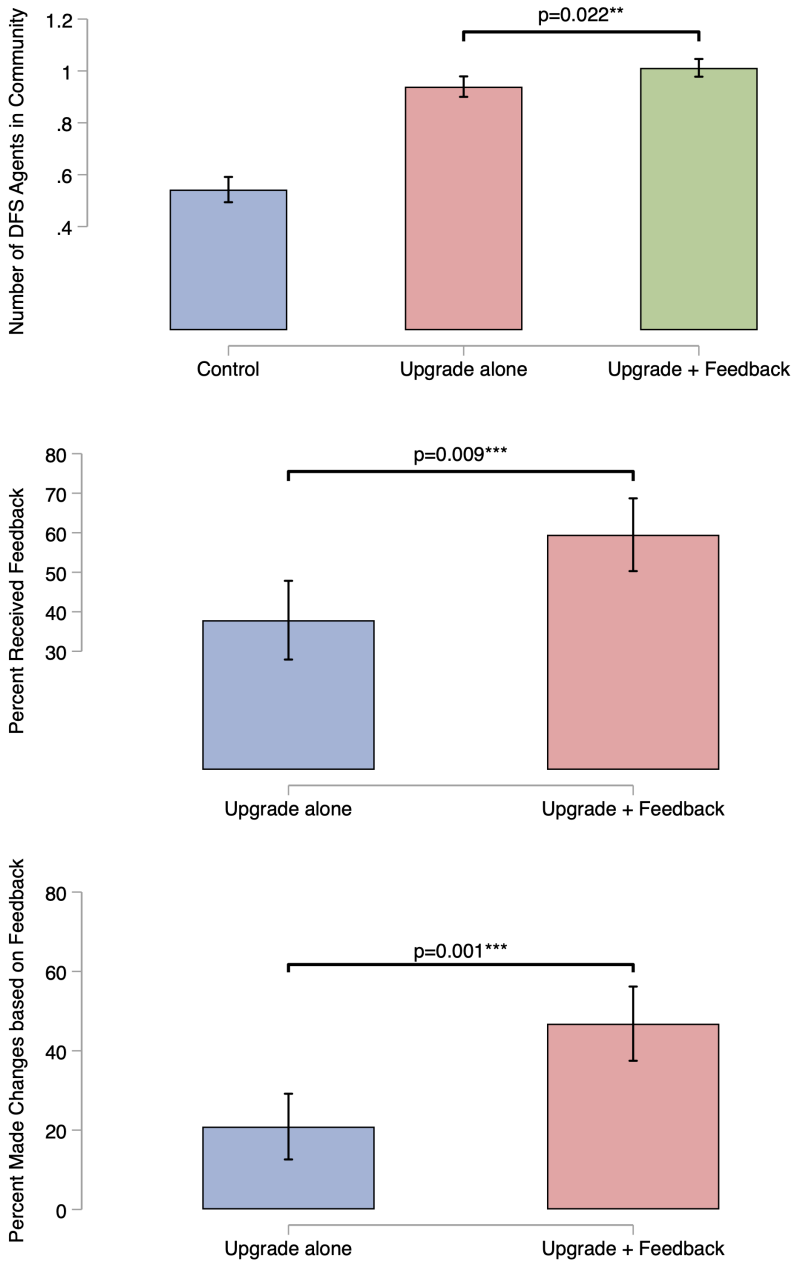
Main Results for Text

Table 1: Summary Statistics

	Mean (1)	SD (2)	Min (3)	Max (4)	N (5)
VSLA Year of formation (yrs)	2020.4	2.6	2010	2022	195
Female members	16.4	3.5	0	29	195
Total members	25.0	0.5	20	30	195
Current cash (SLL)	5557.2	4352.3	800	24300	190
Loan (SLL)	3091.4	4303.0	0	43428	173
Total cash assets (SLL)	8327.9	6224.8	1000	46486	195
Avg. weekly savings	163.5	201.7	5	700	181
Community population	739.3	517.1	115	5722	192
Endline (Household-Level, Control Arm Only)					
Poverty likelihood (%, Schreiner 2018)	62.9	25.9	0.1	96.4	629
DFS/mobile-money agent present in community	0.06	0.23	0	1	649

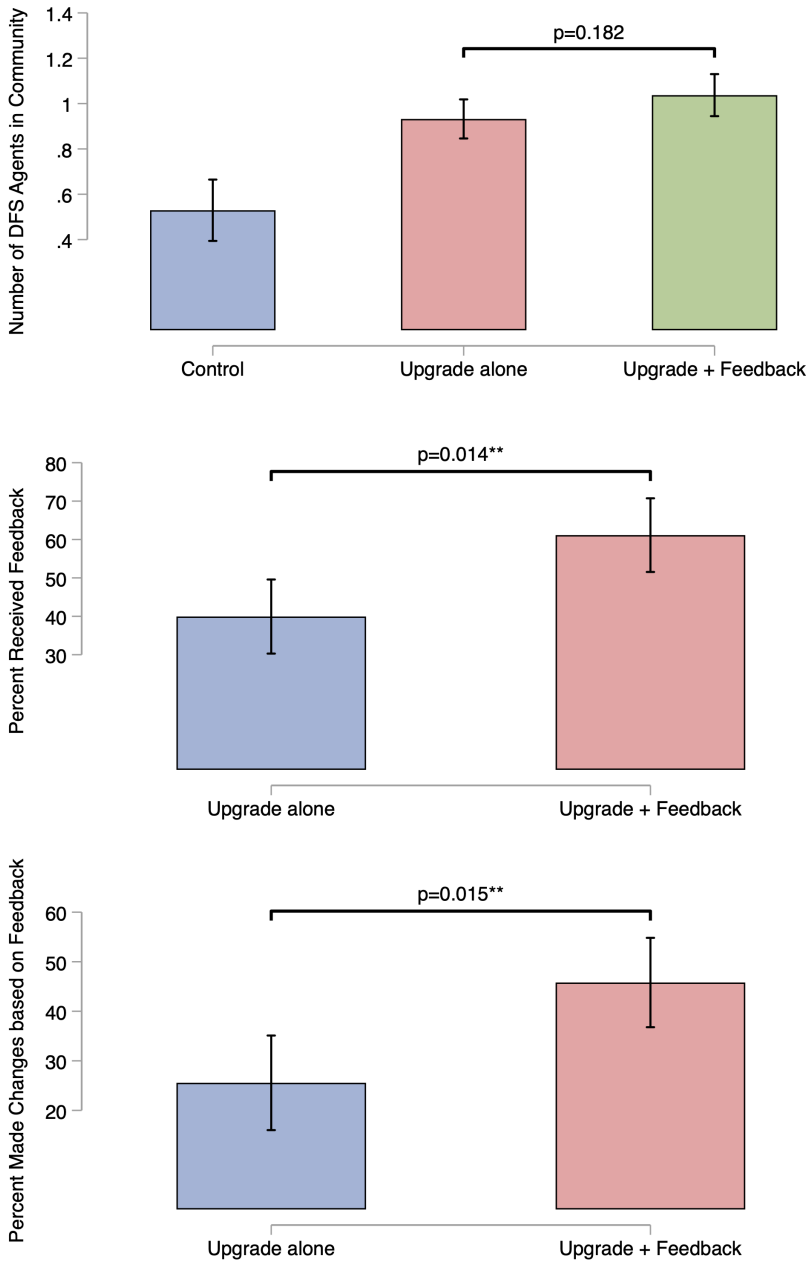
NOTE: The top panel is the full VSLA-level randomization frame, pooled across all three treatment arms and all three cohorts, prior to any endline attrition; this pre-intervention (baseline) data was collected by the implementation partner, CAFOD. See Table C.1 for the same covariates broken out by treatment arm. The bottom panel differs in unit, timing, and sample from the rows above: it reports household-level *endline* survey measures, restricted to the untreated control arm only, so as to characterize these communities' poverty and financial-service access uncontaminated by our intervention. Poverty likelihood is the household-level probability (percent) that per-capita consumption falls below the national poverty line, computed from the 2018 Simple Poverty Scorecard for Sierra Leone (Schreiner 2018; see Section 5.A). DFS/mobile-money agent presence is an indicator for having such an agent physically located inside the household's own community (see Table C.3, column 2). Exchange rate: 1 US$ = 20.97 Sierra Leonean Leone (SLL).

Figure 1: Supply-Side Impact: Raw Group Means



NOTE: Endline changes in agent (provider) density (Panel A: # DFS agents in the community), providers' self-reported receipt of monthly feedback (Panel B), and providers' reported behavioral changes in response to feedback (Panel C), across experimental arms. In Panel A, observations are at the household level; in Panels B and C they are at the agent level. Error bars are 90% confidence intervals from each arm's own sample mean. The bracketed p-value is a simple (unadjusted) two-sample t-test comparing the Upgrade alone and Upgrade + Feedback means directly. Sample sizes: Panel A $N = 647$ Control, 1,056 Upgrade alone, 1,098 Upgrade + Feedback; Panels B–C $N = 66$–67 Upgrade alone, 79 Upgrade + Feedback.

Figure 2: Supply-Side Impact: Regression-Based Treatment Effects



NOTE: Endline changes in agent (provider) density (Panel A: # DFS agents in the community), providers' self-reported receipt of monthly feedback (Panel B), and providers' reported behavioral changes in response to feedback (Panel C), across experimental arms. In Panel A, observations are at the household level; in Panels B and C they are at the agent level. Error bars are 90% confidence intervals from each arm's covariate-adjusted predictive margin. The bracketed p-value is a Wald test of equality between the Upgrade alone and Upgrade + Feedback coefficients in the same covariate-adjusted specification (household/agent controls, strata fixed effects, clustered SEs) used for Table C.3. Sample sizes: Panel A $N = 647$ Control, 1,056 Upgrade alone, 1,098 Upgrade + Feedback; Panels B–C $N = 66$–67 Upgrade alone, 79 Upgrade + Feedback.

Table 2: Changes in Provider Effort and Consumer Trust Effects

	A. Agent Privacy Practices			B. Consumer Trust In:					
	3rd Person Present (1)	PIN Req. (2)	KLK Privacy (3)	Family/ Friends (4)	Comm. People (5)	Local VSLA (6)	Momo Agent (7)	Banking Agent (8)	KLK Index (9)
Upgrade (β_1)	–	–	–	-0.028*	-0.009	-0.020	0.069**	0.009	-0.000
				(0.016)	(0.032)	(0.012)	(0.032)	(0.047)	(0.062)
Upgr.+Feedback (β_2)	-0.132**	-0.160	-0.249**	-0.018	0.004	-0.025**	0.067**	-0.012	0.007
	(0.059)	(0.141)	(0.114)	(0.017)	(0.033)	(0.012)	(0.030)	(0.048)	(0.060)
Test: $\beta_1 = \beta_2$ (p)	–	–	–	0.381	0.536	0.724	0.951	0.608	0.892
MHT (RW p), β_1	–	–	–	–	0.849	0.190	0.002	0.849	–
MHT (RW p), β_2	0.131	0.663	–	–	0.899	0.062	0.003	0.899	–
Obs.	136	136	136	2,619	2,619	2,619	2,619	2,619	2,619
Mean (ref. grp.)	0.24	2.12	0.00	0.96	0.86	0.97	0.79	0.61	0.00
SD (ref. grp.)	0.43	0.95	0.77	0.20	0.35	0.17	0.41	0.49	0.58
Data source	Agent	Agent	Agent	HH	HH	HH	HH	HH	HH

NOTE: In Panel A, observations are at the agent level, while in Panel B they are at the household level. Dependent variables are end-line survey-based measures. Robust standard errors clustered at community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects. In the agent data, there is no pure control group; comparisons in Panel A are between the Upgrade and Upgrade + Feedback groups only, so no $\beta_1 = \beta_2$ test is reported there; their Romano–Wolf p-values come from the same combined 8-outcome family described in Table C.3's note, spanning feedback, privacy, and consumer-protection/trust outcomes together. The KLK index columns (both panels) and the Family/Friends trust column are not part of that family, so no MHT p-value exists for them ("–").

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 3: Treatment Effects on Adoption and Formal Financial Accounts Ownership

	Owns DFS Account (1)	DFS: Momo Indicator (2)	# Momo Accounts (3)	# Bank Accounts (4)	KLK Index (5)
Upgrade (β_1)	0.052*	0.044	0.083*	0.043***	0.111*
	(0.030)	(0.030)	(0.050)	(0.011)	(0.063)
Upgrade+Feedback (β_2)	0.091***	0.092***	0.179***	0.037***	0.202***
	(0.030)	(0.030)	(0.058)	(0.010)	(0.064)
Test: $\beta_1 = \beta_2$ (p)	0.192	0.121	0.121	0.631	0.161
MHT (RW p), β_1	0.045	0.101	0.101	0.006	–
MHT (RW p), β_2	0.002	0.002	0.002	0.003	–
Observations	2,619	2,619	2,613	2,618	2,612
Mean, dep. var. (control)	0.23	0.22	0.51	0.01	-0.19
SD, dep. var. (control)	0.42	0.42	0.73	0.10	0.88
Data source	HH	HH	HH	HH	HH

NOTE: Observations are at the household level. Dependent variables are end-line survey-based measures. Robust standard errors clustered at community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects. All continuous variables are winsorized at the 1% level. DFS stands for Digital Financial Services. The KLK Index column is not included in the Romano–Wolf family, so it has no MHT p-value ("–"); every other Upgrade+Feedback effect here survives Romano–Wolf correction at $p < 0.01$.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 4: Treatment Effects on Usage and Adoption-Usage Differences (Extensive Margin)

	Used FS	Used DFS	DFS: Momo	DFS: GT Bank	KLK Index	Adopt but Don't Use
	(1)	(2)	(3)	(4)	(5)	(6)
Upgrade (β_1)	0.028	0.025	0.043	0.014	0.073	-0.018
	(0.041)	(0.043)	(0.044)	(0.009)	(0.082)	(0.037)
Upgrade+Feedback (β_2)	0.067*	0.082**	0.104***	0.013*	0.190**	-0.062*
	(0.037)	(0.039)	(0.039)	(0.008)	(0.074)	(0.034)
Test: $\beta_1 = \beta_2$ (p)	0.211	0.134	0.118	0.945	0.117	0.119
MHT (RW p), β_1	0.470	0.588	0.288	0.182	–	0.644
MHT (RW p), β_2	0.005	0.005	0.002	0.090	–	0.090
Observations	2,619	2,619	2,619	2,619	2,619	833
Mean, dep. var. (control)	0.66	0.42	0.35	0.005	-0.17	0.13
SD, dep. var. (control)	0.47	0.49	0.48	0.068	0.91	0.33
Data source	HH	HH	HH	HH	HH	HH

NOTE: Observations are at the household level. Dependent variables are end-line survey-based measures. Robust standard errors clustered at community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects. All continuous variables are winsorized at the 1% level. Across Tables 3–4, β_1 (Upgrade alone) is rarely distinguishable from β_2 (Upgrade+Feedback) by the direct $\beta_1 = \beta_2$ test (all $p > 0.10$), yet β_2 alone survives Romano–Wolf correction on nearly every outcome while β_1 mostly does not – consistent with feedback adding power on top of upgrading, even where the two arms are not statistically distinguishable from each other. KLK Index has no MHT p-value (not in the Romano–Wolf family).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 5: Treatment Effects on Household Usage Intensity (Intensive Margin)

	Value: Outside Agt (SLL/90d) (1)	P2P: Value (SLL/90d) (2)	P2P: Attempts (ct./90d) (3)	Non-Cash Sent (SLL/90d) (4)	Non-Cash Received (SLL/90d) (5)	Sent+Recv. Sum (SLL/90d) (6)	Sent+Recv. Avg (SLL/90d) (7)	KLK Usage Index (8)
Upgrade (β_1)	-18.838	3.096	0.032	35.774	41.449	77.247	38.624	0.091
	(37.198)	(3.566)	(0.033)	(25.004)	(28.754)	(48.401)	(24.200)	(0.074)
Upgrade+Feedback (β_2)	-38.549	11.148***	0.109***	50.503**	42.372	93.045*	46.522*	0.186**
	(35.298)	(3.671)	(0.033)	(24.479)	(28.123)	(48.273)	(24.136)	(0.072)
Test: $\beta_1 = \beta_2$ (p)	0.457	0.052	0.049	0.579	0.977	0.768	0.768	0.243
MHT (RW p), β_1	0.606	0.606	0.493	0.177	0.177	–	–	–
MHT (RW p), β_2	0.159	0.007	0.002	0.010	0.083	–	–	–
Observations	2,618	2,619	2,619	2,619	2,618	2,618	2,618	2,619
Mean, dep. var. (control)	207.55	5.03	0.035	27.85	45.09	72.94	36.47	0.00
SD, dep. var. (control)	497.05	43.40	0.300	208.06	294.03	390.49	195.25	0.61
Data source	HH	HH	HH	HH	HH	HH	HH	HH

NOTE: Observations are at the household level. Dependent variables are end-line survey-based measures of intensive-margin digital-financial-services usage over the trailing 90 days, following up on the two columns of Table 4 (Columns 2–3) that measure usage on the extensive margin only. Robust standard errors clustered at the community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects, following the same specification as Table 4. All continuous variables, including Column 3's count variable, are winsorized at the 1% level. Column 6 (Sent+Recv. Sum) is the sum of Columns 4–5 (each already winsorized before summing); Column 7 (Sent+Recv. Avg) is Column 6 divided by two. Column 8 (KLK Usage Index) is a control-group-standardized z-score row-mean of Columns 1–5 (each winsorized before standardizing). As in Table 4, Upgrade alone (β_1) is rarely distinguishable from Upgrade+Feedback (β_2) by the direct test, yet β_2 is the coefficient that is significant on four of the five raw outcomes and on the Sent+Recv. and KLK composites, consistent with feedback adding power on top of upgrading. Column 1 (DFS Value Outside Local Agent) is the exception, with neither arm distinguishable from zero. MHT (Romano–Wolf) p-values for the five raw outcomes (Columns 1–5): Upgrade+Feedback survives correction on P2P Value, P2P Attempts, and Non-Cash Sent (Columns 2–4), but not on DFS Value Outside Agent or Non-Cash Received (Columns 1 and 5). Columns 6–8 are derived/composite measures and, following this paper's convention for such measures (e.g. KLK Index in Table 4), are excluded from the Romano–Wolf family.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 6: Effects on Usage by Transaction Channel

	Deposit (Cash-In) (1)	Withdraw (Cash-Out) (2)	Peer-to-Peer (3)	Savings Instr.: DFS (4)	Savings Instr.: Inf. (5)	KLK Index (6)
Upgrade (β_1)	0.042*	0.066	0.011	0.078***	-0.011	0.140**
	(0.025)	(0.043)	(0.012)	(0.023)	(0.042)	(0.056)
Upgrade+Feedback (β_2)	0.006	0.092**	0.046***	0.066***	-0.057	0.166***
	(0.024)	(0.037)	(0.013)	(0.024)	(0.042)	(0.053)
Test: $\beta_1 = \beta_2$ (p)	0.141	0.450	0.010	0.648	0.215	0.645
MHT (RW p), β_1	0.020	0.020	0.189	0.001	–	0.001
MHT (RW p), β_2	0.738	0.002	0.001	0.005	–	0.001
Observations	2,619	2,619	2,619	2,619	2,619	2,619
Mean, dep. var. (control)	0.089	0.274	0.017	0.125	0.596	-0.197
SD, dep. var. (control)	0.285	0.446	0.129	0.331	0.491	0.556
Data source	HH	HH	HH	HH	HH	HH

NOTE: Observations are at the household level. Dependent variables are end-line survey-based measures. Robust standard errors clustered at community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects. All continuous variables are winsorized at the 1% level. Savings Instr.: Informal was never part of the Romano–Wolf family, so it has no MHT p-value. The KLK Index is the predicted score from a principal-factor analysis of the other five columns (one retained factor), which already accounts for Savings Instr.: Informal moving in the opposite direction from the rest (informal saving falls as digital usage rises).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 7: Spillover Effects on VSLA Business

	Income/Wk	# Current	Contribution	% Digital	# Members	Default Risk (Count < 4)		KLK
	(SLL)	Members	Amount (SLL)	Contrib.	Defaulted	Low-Risk	High-Risk	Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Upgrade (β_1)	51.6	0.72	455.0	3.20	-0.55	-0.42	0.05	0.220
	(62.8)	(3.32)	(714.6)	(2.36)	(0.76)	(0.30)	(1.68)	(0.267)
Upgrade+Feedback (β_2)	159.7*	1.28	1009.0	6.88***	1.20	-0.72**	0.70	0.529
	(94.5)	(2.83)	(695.6)	(2.51)	(0.86)	(0.29)	(1.35)	(0.330)
Test: $\beta_1 = \beta_2$ (p)	0.239	0.856	0.457	0.182	0.007	0.092	0.698	0.361
MHT (RW p), β_1	0.897	0.897	0.897	0.747	0.897	–	–	–
MHT (RW p), β_2	0.409	0.637	0.409	0.146	0.409	–	–	–
Observations	172	174	174	173	174	120	54	174
Mean, dep. var. (control)	18.5	29.6	1694.1	0.00	2.6	1.03	6.73	0.00
SD, dep. var. (control)	82.1	12.5	2141.3	0.00	3.9	–	–	0.56
Data source	Secretary	Secretary	Secretary	Secretary	Secretary	Secretary	Secretary	Secretary

NOTE: Observations are at the VSLA/secretary level. Dependent variables are end-line survey-based measures. Income/Wk is the VSLA secretary's response to "What was the total business income earned during the last 7 days / last week after paying all expenses (profits)?" Robust standard errors clustered at community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects. All continuous variables are winsorized at the 1% level. VSLA stands for Village Savings and Loan Association. "Default Risk (Count < 4)" splits VSLAs by whether fewer than 4 members were in default or delayed contribution over the past 90 days (Low-Risk) versus 4 or more (High-Risk). This cutoff sits at the 75th percentile of the default-count distribution (the sample median is zero, since a majority of VSLAs report no defaulting members at all), just below the median of five among the subset of VSLAs with at least one defaulting member, striking a balance that leaves meaningful variation on both sides of the threshold for identifying heterogeneous effects. The KLK Index pools four of this table's first five outcomes (Business Income, # Current Members, Contribution Amount, # Members Defaulted), excluding % Digital Contribution: that variable has exactly zero variance in the control group, so standardizing it would divide by zero. Romano–Wolf MHT correction, by contrast, is unaffected by this exclusion and remains a 5-outcome family over all five raw outcomes (the index itself is not part of that family); neither default-risk subsample column is part of it either.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 8: Treatment Effects on Shock Mitigation

	Any Shock (90 Days) (1)	Unmitigated Shock (2)	Shock Index (Mean) (3)	Shock Index (Any) (4)
Upgrade (β_1)	0.006	0.042	0.024	0.021
	(0.025)	(0.041)	(0.028)	(0.025)
Upgrade+Feedback (β_2)	-0.036	-0.015	-0.025	-0.029
	(0.028)	(0.041)	(0.029)	(0.027)
Test: $\beta_1 = \beta_2$ (p)	0.087	0.069	0.035	0.037
MHT (RW p), β_1	0.998	0.529	0.755	0.834
MHT (RW p), β_2	0.611	0.912	0.806	0.766
Observations	2,619	2,619	2,619	2,619
Mean, dep. var. (control)	0.78	0.40	0.59	0.78
Data source	HH	HH	HH	HH

NOTE: Observations are at the household level. Dependent variables are end-line survey-based measures, recoded so that 1 = shock/unmitigated shock throughout. Robust standard errors clustered at community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects. Column 1 ("Any Shock") indicates having experienced any negative income or wealth shock in the past 90 days. Column 2 ("Unmitigated Shock") is a distinct survey item asking whether the household experienced a shock in the past 90 days that it was unable to mitigate or resolve financially, rather than Column 1's broader question of whether any shock occurred at all. Column 3 ("Shock Index (Mean)") is the simple average of Columns 1 and 2. Column 4 ("Shock Index (Any)") equals 1 if either Column 1 or Column 2 is 1. Romano–Wolf p-values are from a 14-outcome family that jointly includes all four shock measures here, household poverty likelihood, and all nine poverty-scorecard components (Section 5.A); because the four shock measures are re-codings of the same two underlying survey questions, pooling them into one family is a conservative choice.

Online Appendix

A. Appendix: Enrollment and Data Collection Timeline

Table A.1 reports the enrollment and data collection timetable by cohort. A year-0 pilot preceded staggered onboarding of the three cohorts. Each cohort had three rounds of brief customer surveys during the intervention, constructing the round-by-round consumer-review measures analyzed in Appendix D, followed by the endline surveys (agent, household, VSLA secretary) providing Section 6's primary outcomes. The staggered rollout from Section 3.B appears directly below: cohort 1 data collection began in March 2024 and cohort 3 endline concluded in January 2026.

Table A.1: Enrollment and Data Collection Timeline

Activity	Sample Size (1)	Start Date (2)	End Date (3)
Year 0 (Pilot)			
	—	11/1/2019	3/31/2020
Cohort 1			
Onboarding	64	11/1/2022	11/30/2023
Customer Surveys — R1	1,296	3/18/2024	5/16/2024
Customer Surveys — R2	540	7/1/2024	7/18/2024
Customer Surveys — R3	528	8/7/2024	8/23/2024
Endline Survey — Agent	47	11/2/2024	12/2/2024
Endline Survey — Household	926	10/31/2024	12/3/2024
Endline Survey — Secretary	50	11/2/2024	12/3/2024
Cohort 2			
Onboarding	65	5/1/2023	5/30/2024
Customer Surveys — R1	477	9/17/2024	10/7/2024
Customer Surveys — R2	452	1/13/2025	2/4/2025
Customer Surveys — R3	444	3/12/2025	3/29/2025
Endline Survey — Agent	48	10/7/2025	10/27/2025
Endline Survey — Household	955	10/6/2025	10/27/2025
Endline Survey — Secretary	64	10/6/2025	10/27/2025
Cohort 3			
Onboarding	66	12/1/2023	10/30/2024
Customer Surveys — R1	366	2/17/2025	3/4/2025
Customer Surveys — R2	438	6/16/2025	7/4/2025
Customer Surveys — R3	407	9/1/2025	9/22/2025
Endline Survey — Agent	94	11/26/2025	1/28/2026
Endline Survey — Household	1,004	11/26/2025	1/28/2026
Endline Survey — Secretary	67	11/26/2025	1/28/2026

NOTE: Sample size is the number of completed responses for that survey round/wave; for the three cohort Onboarding rows, it is instead the number of VSLAs onboarded that cohort, one per community (Control/Upgrade/Upgrade+Feedback: Cohort 1, 14/25/25; Cohort 2, 15/24/26; Cohort 3, 17/23/26; pooled, 46/72/77, for a study total of N=195). The Year-0 pilot preceded this randomization and has no associated community count. Customer Surveys are the brief, repeated (R1–R3) consumer-evaluation phone audits described in Section 3.B and analyzed in Appendix D; Endline Surveys are the single-wave, comprehensive surveys of agents, households, and VSLA secretaries analyzed throughout Section 6. Under Cohort 1, Customer Surveys — R1 were fielded to *all* treatment and control communities alike (i.e., prior to any differential feedback provision), unlike later rounds and cohorts.

B. Appendix: Proofs

Proof 1A: The Access Effect. This proof demonstrates that infrastructure is a necessary but insufficient condition for high-quality service provision. In the absence of accountability ($\theta = 0$), the provider's first-order condition (FOC) simplifies to:

$$\gamma A \tau T'(e) = \phi e \tag{5}$$

Since A is a binary indicator $\{0, 1\}$, if $A = 0$ (unserved communities), the marginal benefit of effort is zero, and the provider chooses the corner solution $e^* = 0$. When infrastructure is introduced ($A = 1$), the provider faces a trade-off between baseline commissions and effort costs. However, if the cost of effort ϕ (e.g., the burden of maintaining privacy in a village setting) is sufficiently large relative to the commission rate γ, the optimal effort e^* remains at a minimal level near the lower bound. This confirms that while infrastructure expansion moves the extensive margin by shifting baseline usage $u_0 > 0$, it does not inherently align the provider's incentives for privacy and consumer protection practices or trust-based intensive-margin usage. □

Proof 1B: The Access Effect. In the pre-intervention state, the cost of accessing formal services is c_{travel}, representing the distance to the nearest bank. A consumer i with demand ω_i will use formal finance only if $\omega_i - c_{travel} > 0$. In unserved areas, c_{travel} is high, resulting in usage $u = 0$ for the majority of the population. The upgrade of VSLAs to agents reduces transaction costs to $c_{local} \approx 0$. This shifts the baseline usage u_0:

$$\Delta u_{access} = \int \left(I[\omega_i - c_{local} > 0] - I[\omega_i - c_{travel} > 0] \right) dF(\omega) \tag{6}$$

Since $c_{local} < c_{travel}$, it follows that $\Delta u_{access} > 0$. However, without accountability ($\theta = 0$), the provider chooses effort $e^*_{low} = \frac{\gamma \tau}{\phi}$ (assuming $T(e) = e$). If ϕ is high (e.g., privacy costs), e^*_{low} remains minimal. Thus, infrastructure provides the *capacity* for usage, but accountability is required to move the intensive margin. □

Proof 2A: The Behavioral Shift. First, we establish the directional shift using the Implicit Function Theorem. Let F be the first-order condition $\gamma A \tau T'(e) + \theta V'(e) - \phi e = 0$. The comparative static for accountability is:

$$\frac{\partial e^*}{\partial \theta} = -\frac{\partial F / \partial \theta}{\partial F / \partial e} = -\frac{V'(e)}{\gamma A \tau T''(e) + \theta V''(e) - \phi} \tag{7}$$

The numerator $V'(e)$ is positive by definition. The denominator is the second-order condition (SOC) for a maximum, which is negative. Thus, $\frac{\partial e^*}{\partial \theta} > 0$, confirming that accountability unambiguously raises effort.

Proof 2B: Step-Function Jump vs. Learning. We contrast a *Learning* regime with an *Accountability* regime, nested in a single effort path so the two are falsifiable against one another rather than compared as separately specified stories. Let the *learning* component be driven not by a falling cost of effort ϕ (which stays fixed throughout this regime) but by the provider's growing *effective utilization* of the infrastructure A as experience accumulates. Formally, cumulative experience t raises effective utilization $a(t)$ from a baseline $a_0 \geq 0$

toward an asymptotic ceiling $a_\infty > a_0$, following the standard diminishing-returns experience curve:

$$a(t) = a_\infty - (a_\infty - a_0)e^{-\kappa t}, \qquad \kappa > 0. \tag{8}$$

With $\theta = 0$ (pre-accountability) and linear $T(e) = e$, the first-order condition $\gamma A\, a(t)\tau T'(e) - \phi e = 0$ solves to

$$e_L^*(t) = \frac{\gamma A \tau}{\phi}\, a(t) \equiv c \cdot a(t), \qquad c \equiv \frac{\gamma A \tau}{\phi} > 0, \tag{9}$$

so $e_L^*(t)$ is a positive scalar multiple of $a(t)$, not an independently assumed shape. Differentiating the composition directly:

$$e_L^{*\prime}(t) = c \cdot a'(t) = c\kappa(a_\infty - a_0)e^{-\kappa t} > 0, \tag{10}$$

$$e_L^{*\prime\prime}(t) = c \cdot a''(t) = -c\kappa^2(a_\infty - a_0)e^{-\kappa t} < 0 \quad \text{for every finite } t \geq 0, \tag{11}$$

confirming $e_L^*(t)$ is strictly increasing and strictly concave everywhere: curvature inherited intact from $a(t)$ through the first-order condition, not asserted for e^* directly, and strict (the second derivative is never zero, unlike a genuinely linear path). Writing $e_0 \equiv c \cdot a_0$ and $e_\infty \equiv c \cdot a_\infty$ recovers exactly the same path as before,

$$e_L^*(t) = e_0 + (e_\infty - e_0)(1 - e^{-\kappa t}), \qquad e_\infty > e_0 \geq 0, \tag{12}$$

gains are largest early and shrink as effort approaches its ceiling, consistent with standard diminishing-returns learning curves.

The *accountability* component is the discrete jump derived above: at t^*, the moment feedback is introduced, the provider's marginal utility of effort jumps by

$$\Delta \equiv e^*(\bar{\theta}) - e^*(0) \approx \frac{\bar{\theta} V'(e)}{\phi} > 0, \tag{13}$$

Because $\bar{\theta} V'(e) > 0$ is added to the FOC the moment the feedback mechanism is launched, the model predicts a discontinuous jump of this size. Nesting both components in one effort path,

$$e^*(t; \theta) = e_L^*(t) + \Delta \cdot \mathbf{1}\{t \geq t^*\}, \tag{14}$$

makes the two regimes directly comparable within a single model: under pure learning ($\theta = 0$ for all t), $e^*(t) = e_L^*(t)$ is smooth and strictly concave everywhere, with no discontinuity. Under accountability, $e^*(t)$ is continuous and concave everywhere *except* at t^*, where it jumps by exactly

$$\lim_{t \to t^{*+}} e^*(t) - \lim_{t \to t^{*-}} e^*(t) = \Delta = \frac{\bar{\theta} V'(e)}{\phi} > 0. \tag{15}$$

The two regimes are therefore distinguishable by a single testable statistic: a structural break in $e^*(t)$ at t^* versus smooth concavity throughout, rather than only by the surge's timing and magnitude. The immediacy and sustained linearity of the +20.5 pp surge in our data support this incentive shock over a gradual learning process. Table D.1 (Appendix D) implements exactly this comparison on the round-to-round consumer-review data: for the Core Review Index and Overall Satisfaction, the round-1-to-round-2 change is large and

the round-2-to-round-3 change is statistically indistinguishable from zero, consistent with a level shift at t^* rather than the continued, smaller gain a concave learning path $e_L^*(t)$ would predict; not every individual privacy item shows this shape, a caveat we report in full there rather than fold into the headline claim. □

Proof 3A: The Complementary Effect. We define infrastructure (A) and accountability (θ) as productive complements if the provider's *indirect* (optimized) utility, $V_p(A,\theta) \equiv U_p(e^*(A,\theta);A,\theta)$, is supermodular, i.e. $\frac{\partial^2 V_p}{\partial A \partial \theta} > 0$. This is the economically relevant object: A and θ enter the primitive $U_p = \gamma A[u_0 + \tau T(e)] - \frac{1}{2}\phi e^2 + \theta V(e)$ additively, with no direct $A\theta$ interaction term, so $\frac{\partial^2 U_p}{\partial A \partial \theta}\Big|_{e \text{ fixed}} = 0$ identically. Any complementarity between infrastructure and accountability must come through the endogenous response of effort, not through a direct interaction in the utility function itself.

By the envelope theorem, since e^* is chosen optimally, the indirect effect of θ through e^* vanishes at the first-order condition, leaving only the direct effect:

$$\frac{\partial V_p}{\partial \theta} = \frac{\partial U_p}{\partial \theta}\bigg|_{e=e^*} = V(e^*(A,\theta)). \tag{16}$$

Differentiating again with respect to A:

$$\frac{\partial^2 V_p}{\partial A \partial \theta} = V'(e^*) \cdot \frac{\partial e^*}{\partial A}. \tag{17}$$

$V'(e^*) > 0$ by assumption (reputational value is increasing in effort). From the first-order condition $\gamma A \tau T'(e) + \theta V'(e) - \phi e = 0$ and the implicit function theorem (the same comparative-statics argument as Proof 2A, applied to A in place of θ),

$$\frac{\partial e^*}{\partial A} = -\frac{\partial F/\partial A}{\partial F/\partial e} = -\frac{\gamma \tau T'(e)}{\gamma A \tau T''(e) + \theta V''(e) - \phi} > 0, \tag{18}$$

since the numerator is positive and the denominator is the (negative) second-order condition. Hence

$$\frac{\partial^2 V_p}{\partial A \partial \theta} = V'(e^*) \cdot \frac{\partial e^*}{\partial A} > 0 : \tag{19}$$

infrastructure and accountability are supermodular in the provider's optimized utility. Essentially, the infrastructure increases the "stakes" for the provider by linking effort to a larger pool of potential revenue γu, which raises the optimal response e^* to any given level of accountability. Consequently, the accountability mechanism becomes more effective at driving usage u and subsequent welfare gains. This synergy results in a total market development effect where $Y(A=1,\theta=1) > Y(1,0) + Y(0,1)$. This supermodularity explains the downstream spillovers into business revenues and household incomes, as the combined treatment optimizes both the *capacity* and the *quality* of intermediation. □

Proof 3B: Complementary Effects. Proof 3A establishes supermodularity of the *provider's own profit function* U_p in (A,θ). Extending this to household welfare requires an additional assumption: *household welfare $Y(A,\theta)$ is increasing and supermodular in usage u and effort e* ($\partial Y/\partial u > 0$, $\partial Y/\partial e > 0$, $\partial^2 Y/\partial u \partial e \geq 0$), so that Y inherits Proof 3A's

complementarity whenever u and e jointly determine welfare, as they do empirically through the digitization, trust, and default-rate channels documented in Section 6. Under this assumption, we define market development Y as a function of access (A) and accountability (θ). Complementarity exists if $\frac{\partial^2 Y}{\partial A \partial \theta} > 0$. As proved in the FOC, effort $e^*(\theta)$ is increasing in accountability. The total change in usage is:

$$\frac{du}{d\theta} = A \cdot \tau \frac{\partial T}{\partial e} \frac{\partial e^*}{\partial \theta} \tag{20}$$

If $A = 0$ (no infrastructure), then $\frac{du}{d\theta} = 0$. When $A = 1$, the marginal impact of accountability becomes positive. This supermodularity extends to household welfare: let household income I be a function of successful investment $I = f(u(1-\delta))$, where $\delta(e)$ is the default rate ($\delta' < 0$). Differentiating with respect to θ:

$$\frac{\partial I}{\partial \theta} = f' \cdot \left[(1-\delta) \frac{\partial u}{\partial \theta} + u \left(-\frac{\partial \delta}{\partial e} \frac{\partial e^*}{\partial \theta} \right) \right] > 0 \tag{21}$$

The first term is the *Usage Effect* (quantity) and the second is the *Efficiency Effect* (quality). Together, they produce $Y(1,1) > Y(1,0) + Y(0,1)$, confirming accountability as the "missing link" for growth. □

C. Appendix: Additional Tables

This appendix collects five tables referenced in the main text: the pre-intervention randomization-balance check, the endline sample-attrition breakdown, the treatment effects on upgrading, feedback, and provider behavior, and community-level (total and average) versions of Table 5's household usage-intensity outcomes, used to rule out a compositional/household-count explanation for that table's result.

Table C.1: Pre-Intervention Randomization Balance

	Control Mean (SD) (1)	Upgrade (2)	Upgrade+Feedback (3)	N (4)
Year of formation	2020.5 (2.8)	0.09	-0.14	195
		(0.29)	(0.28)	
Female members	16.3 (4.3)	-0.25	0.22	195
		(0.72)	(0.70)	
Total members	25.1 (0.7)	-0.13	-0.12	195
		(0.12)	(0.11)	
Current cash (SLL)	5935.7 (4302.7)	-299.7	599.4	190
		(711.9)	(708.2)	
Loan (SLL)	3408.6 (3781.7)	29.1	57.0	173
		(553.5)	(806.7)	
Total cash assets (SLL)	8892.0 (5715.5)	-220.3	948.7	195
		(692.1)	(909.8)	
Avg. weekly savings	173.9 (223.9)	-4.9	5.4	181
		(30.4)	(30.5)	
Community population	756.3 (318.4)	7.4	83.8	192
		(44.3)	(79.8)	
Joint test: covariates predict arm (mprobit), $\chi^2(14) = 16.95$, $p = 0.259$, $N = 156$				
Joint test: covariates predict pooled treatment (LPM), $F(7, 143) = 0.79$, $p = 0.599$, $N = 156$				

NOTE: Sample is the VSLA-level randomization frame (pooled across cohorts, prior to any endline attrition), with randomization-strata and cohort fixed effects and robust standard errors in parentheses. No individual covariate is significant at conventional levels. The joint test is a single Wald chi-squared test, from a multinomial probit (mprobit) model of the 3-category treatment arm on seven of the eight baseline covariates plus randomization-strata and cohort fixed effects, of the null that all 14 covariate coefficients are jointly zero. Total cash assets is excluded from this joint test only (it remains in its own row above) because it is an accounting identity of current cash plus loan and so is exactly collinear with them. The joint test's sample ($N = 156$) is smaller than the full 195-VSLA frame because it requires non-missing values on all seven included covariates simultaneously; Loan, the covariate with the most missingness on its own, has $N = 173$. Covariates do not jointly predict treatment arm at any conventional level. A companion linear-probability joint test, collapsing the three arms to a single pooled-treatment-vs-control indicator, gives the same conclusion and is reported in the row above.

Table C.2: Sample Attrition by Treatment Arm and Respondent Type

	Control (1)	Upgrade (2)	Upgrade+Feedback (3)
Panel A: Household (= Any Respondent under corrected linkage; see note)			
Baseline N (assigned)	35	63	73
Matched at endline	35	59	69
Attrition rate	0.0%	6.3%	5.5%
Differential attrition vs. control (pp)	–	7.3**	6.3**
		(3.4)	(2.9)
Attrition rate, generous†	0.0%	1.6%	0.0%
Differential, generous (pp)	–	1.7	0.0
		(1.7)	(0.2)
Panel B: VSLA/Secretary			
Baseline N (assigned)	35	63	73
Matched at endline	34	56	67
Attrition rate	2.9%	11.1%	8.2%
Differential attrition vs. control (pp)	–	10.2*	6.9
		(5.2)	(4.5)
Attrition rate, generous†	0.0%	4.8%	1.4%
Differential, generous (pp)	–	5.7*	1.8
		(3.1)	(1.6)
Panel C: Agent (no true control arm by design)			
Baseline N (assigned)	–	63	73
Matched at endline	–	56	69
Attrition rate	–	11.1%	5.5%
Differential attrition, U+F vs. Upgrade (pp)	–	–	-6.2
			(4.8)
Attrition rate, generous†	–	4.8%	0.0%
Differential, generous, U+F vs. Upgrade (pp)	–	–	-5.0*
			(2.8)

NOTE: Baseline N and matches are counted on the 171 unambiguous VSLAs described above; Panel C additionally restricts to the 136 VSLAs in the Upgrade and Upgrade+Feedback arms, since agents were not fielded in control communities by design. A VSLA counts as matched if its name, once normalized, matches an endline record in the same chiefdom. Panel A also serves as the "any respondent" measure: a household counts as unreached only if neither a household *nor* a secretary record was matched for it, under either threshold. Each differential row reports coefficients from a regression of a not-matched-at-endline indicator on arm dummies (Panels A–B) or a single Upgrade+Feedback indicator (Panel C), with randomization-strata fixed effects and robust standard errors in parentheses. Joint F-tests that both arm coefficients are zero (primary / generous): Panel A $F = 3.92$, $p = 0.022$ / $F = 0.51$, $p = 0.604$, $N = 171$; Panel B $F = 2.16$, $p = 0.118$ / $F = 1.96$, $p = 0.144$, $N = 171$; Panel C's single coefficient has $p = 0.195$ / $p = 0.080$, $N = 136$. Attrition does not appear to threaten the treatment effects that follow: it is not statistically distinguishable between the Upgrade and Upgrade+Feedback arms in any panel, and once strata fixed effects are included it is not distinguishable from zero either, except for the household/any-respondent measure under the primary threshold and the Upgrade-vs-Upgrade+Feedback agent gap under the generous threshold (significant only at the 10 percent level). Table C.1's covariates are unaffected, since balance is assessed before any attrition occurs.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table C.3: Upgrading, Feedback, and Provider Behavior

	A. Provider Access (Household)				B/C. Agent Feedback/Behavior	
	# Agents in Comm. (1)	Access: Inside (2)	Access: Outside (3)	KLK Index (4)	Received Feedback (5)	Made Changes (6)
Upgrade (β_1)	0.403***	0.188***	-0.100**	0.354***	–	–
	(0.096)	(0.047)	(0.050)	(0.082)		
Upgrade+Feedback (β_2)	0.508***	0.194***	-0.120***	0.406***	0.211**	0.205***
	(0.098)	(0.049)	(0.043)	(0.079)	(0.081)	(0.076)
Test: $\beta_1 = \beta_2$ (p)	0.185	0.888	0.585	0.424	–	–
MHT (RW p), β_1	0.001	0.001	0.001	–	–	–
MHT (RW p), β_2	0.001	0.001	0.001	–	0.047	0.047
Observations	2,613	2,619	2,619	2,613	136	137
Mean, dep. var. (ref. grp.)	0.54	0.06	0.32	-0.32	0.38	0.21
SD, dep. var. (ref. grp.)	0.75	0.23	0.47	0.51	0.49	0.41
Data source	HH	HH	HH	HH	Agent	Agent

NOTE: In Panel A, observations are at the household level, while in Panel B/C they are at the agent level. Dependent variables are end-line survey-based measures. Robust standard errors clustered at community level are reported in parentheses. All specifications include randomization strata and cohort fixed effects. In the agent data, there is no pure control group; comparisons in Panel B/C are between the Upgrade and Upgrade + Feedback groups only, so no $\beta_1 = \beta_2$ test is reported there and the coefficient's own significance (starred above) is the full story. All continuous variables are winsorized at the 1% level. The Panel B/C Romano–Wolf p-value comes from a combined 8-outcome family spanning both Panel B/C here and Table 2 Panels A and B (feedback, privacy, and consumer-protection/trust outcomes together) – see Table 2's note.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table C.4: Household Usage Intensity, Community-Level Totals

	Value: Outside Agt (SLL/90d) (1)	P2P: Value (SLL/90d) (2)	P2P: Attempts (ct./90d) (3)	Non-Cash Sent (SLL/90d) (4)	Non-Cash Received (SLL/90d) (5)	Sent+Recv. Sum (SLL/90d) (6)	Sent+Recv. Avg (SLL/90d) (7)	KLK Usage Index (8)
Upgrade (β_1)	-532.4	17.34	0.197	500.4	325.1	825.5	412.8	0.204
	(577.4)	(52.27)	(0.483)	(370.8)	(391.2)	(667.0)	(333.5)	(0.272)
Upgrade+Feedback (β_2)	-887.1*	136.4**	1.337***	571.4	338.1	909.6	454.8	0.580**
	(530.3)	(54.64)	(0.487)	(357.3)	(385.4)	(660.9)	(330.5)	(0.266)
Test: $\beta_1 = \beta_2$ (p)	0.372	0.047	0.048	0.855	0.976	0.911	0.911	0.227
MHT (RW p), β_1	–	–	–	–	–	–	–	–
MHT (RW p), β_2	–	–	–	–	–	–	–	–
Observations	171	171	171	171	171	171	171	171
Mean, dep. var. (control)	1978.34	48.73	0.34	269.78	436.79	706.57	353.28	0.00
SD, dep. var. (control)	2615.69	140.38	1.07	774.77	1210.76	1636.97	818.48	0.71
Data source	HH	HH	HH	HH	HH	HH	HH	HH

NOTE: This table reports the same five raw outcomes as Table 5 (Columns 1–5), plus their sum, average, and KLK index (Columns 6–8), but treats the community – rather than the household – as the unit of observation: household-level winsorized values are summed within each community, the community-level sums are then re-winsorized at 1%/99%, and the sum/average/KLK-index columns are recomputed from the re-winsorized community sums. This is the market-level framing commercial DFS providers would plausibly care about, as distinct from the household-level effects in Table 5; together with Table C.5, this table is used in the main text to rule out a compositional/household-count explanation for Table 5's result – see that discussion for how the two tables are used together. All regressions use the same covariates as Table 5 (randomization strata and cohort fixed effects); because the community is now the unit of observation (one summed value per community), standard errors are heteroskedasticity-robust (HC1) rather than cluster-robust. No communities are excluded from this analysis: all 171 surveyed communities, covering all 2,619 households in the analytic sample, are included. This table re-tests Table 5's five outcomes at the community level as a robustness check, rather than as a new family of hypotheses, so no separate Romano–Wolf p-values are reported here (see the discussion in the main text).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table C.5: Household Usage Intensity, Community-Level Averages

	Value: Outside Agt (SLL/90d) (1)	P2P: Value (SLL/90d) (2)	P2P: Attempts (ct./90d) (3)	Non-Cash Sent (SLL/90d) (4)	Non-Cash Received (SLL/90d) (5)	Sent+Recv. Sum (SLL/90d) (6)	Sent+Recv. Avg (SLL/90d) (7)	KLK Usage Index (8)
Upgrade (β_1)	-26.50	2.283	0.0220	32.58	30.20	62.79	31.39	0.185
	(40.31)	(3.709)	(0.0337)	(26.32)	(29.08)	(49.74)	(24.87)	(0.191)
Upgrade+Feedback (β_2)	-46.07	10.59***	0.100***	46.54*	31.56	78.11	39.05	0.404**
	(38.42)	(3.836)	(0.0337)	(25.81)	(28.77)	(50.03)	(25.02)	(0.188)
Test: $\beta_1 = \beta_2$ (p)	0.482	0.059	0.058	0.621	0.967	0.787	0.787	0.326
MHT (RW p), β_1	–	–	–	–	–	–	–	–
MHT (RW p), β_2	–	–	–	–	–	–	–	–
Observations	171	171	171	171	171	171	171	171
Mean, dep. var. (control)	191.62	4.88	0.03	27.26	27.01	54.27	27.13	0.00
SD, dep. var. (control)	237.61	19.83	0.13	88.88	77.42	126.37	63.19	0.56
Data source	HH	HH	HH	HH	HH	HH	HH	HH

NOTE: This table reports the same community-level construction as Table C.4, except that each community's five raw outcomes (and the sum/average/KLK-index columns built from them) are averaged across the community's surveyed households rather than summed, before re-winsorizing at the community level and regressing on the same covariates with HC1 robust standard errors. Because an average, unlike a sum, nets out any difference in how many households a community happens to contribute, this table is the more direct test of whether Table 5's household-level pattern survives the household-count imbalance discussed in the main text: it reproduces that table's pattern almost exactly, rescaled by each community's household count, with Upgrade+Feedback significant on the same three columns (P2P Value, P2P Attempts, KLK Index) and nothing else distinguishable from zero. All 171 surveyed communities are included, with no exclusions. As in Table C.4, this is a robustness check on Table 5's already-corrected five outcomes rather than a new family of hypotheses, so no separate Romano–Wolf p-values are reported here.

***Significant at the 1 percent level.

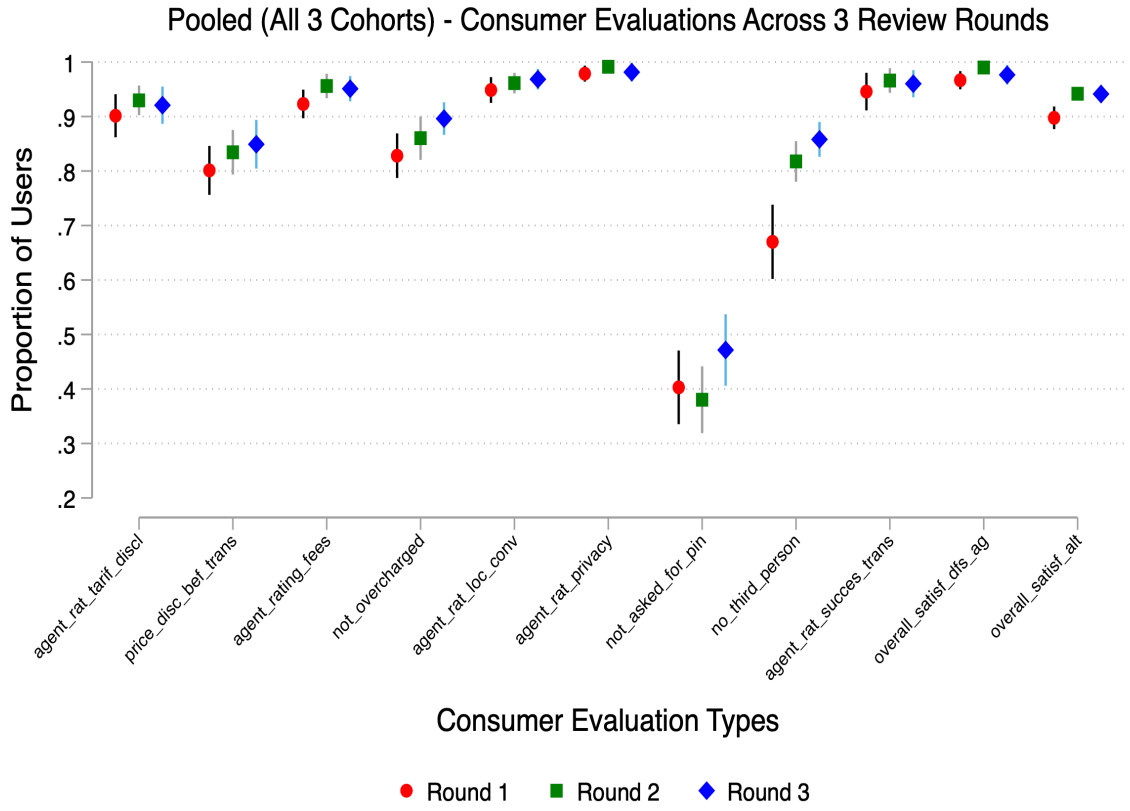
**Significant at the 5 percent level.

*Significant at the 10 percent level.

D. Appendix: Customer Reviews

This appendix collects supplementary analyses of the consumer-review data referenced in the discussion of Table C.3 and Table 2. It begins with round-by-round trends across the three endline survey rounds in twelve consumer-evaluation indicators, both in raw pooled form and residualized on cohort fixed effects, then turns to several targeted follow-up analyses – within-arm targeting, an accountability-versus-learning test, and open-ended complaint text – that this data can speak to but that go beyond the trend figures alone.

Figure D.1: Customer-Reported Evaluations Across Review Rounds: Pooled (All Cohorts)



NOTE: Mean (with 90% CI) of 12 consumer-evaluation indicators – not being overcharged, not being asked for a PIN, no third person present during transactions, overall satisfaction, and a behavioral alternative privacy measure averaging the not-asked-for-PIN and no-third-person items, among others – pooled across all three cohorts (76 communities per round: 25 + 25 + 26), across the 3 survey rounds. A cohort-residualized version of this figure (each indicator's cohort-specific mean subtracted and the pooled grand mean added back before computing round-level means and confidence intervals) produces an almost identical pattern of means and confidence intervals, confirming these trends are not an artifact of cohort composition.

D.1. Within-arm targeting.

The consumer-review data covers only the Upgrade+Feedback arm, which alone was surveyed this way. A cross-arm heterogeneity test (whether the Feedback mechanism's effect is larger where a round-1 baseline index was worse) is therefore not identifiable from this data alone. What can be tested within the Feedback arm is whether later-round improvement is concentrated among communities with worse round-1 scores. Pooling all three cohorts (70 of 76 communities matched across all three rounds; a baseline index averaging the six core review items, mean 0.893 (SD 0.088) at round 1, rising to 0.941 and 0.939 at rounds 2 and 3), a regression of the round-1-to-round-2 change on the round-1 level, with cohort fixed effects, gives a coefficient of -0.818 (robust SE 0.077, $p < 0.001$); the round-1-to-round-3 change gives -0.694 (SE 0.204, $p = 0.001$). Communities starting with worse review scores show much larger subsequent gains. However, the round-1 index already averages 0.89 against a hard ceiling of 1.0, so top performers mechanically have little room to improve; combined with regression to the mean, a coefficient this close to -1 is also what a ceiling effect with no real targeting would produce. A level-on-level check (round 3 on round 1) gives a slope of 0.306 (SE 0.204), consistent with convergence but not sharply estimated. Without a non-Feedback comparison arm in this data, a genuine within-arm targeting effect cannot be distinguished from this mechanical pattern.

D.2. Accountability (immediate response) versus learning (gradual, concave improvement).

This paper interprets the Upgrade+Feedback arm's effect through an accountability channel: agents respond to monthly consumer feedback immediately and substantially, rather than gradually adjusting behavior over repeated cycles the way a learning curve would predict. These two stories predict different shapes of round-to-round change within the Feedback arm. Under accountability, the round-1-to-round-2 change and the round-2-to-round-3 change should be similar in magnitude (a constant, "linear" rate of improvement) or the improvement should arrive as an immediate level shift that then holds (round-2-to-round-3 change close to zero). Under learning, later rounds should show smaller but still positive gains than earlier rounds (a decelerating, concave trajectory that keeps rising). Table D.1 tests this directly on the same panel of 70 communities matched across all three rounds used above, computing each community's two round-to-round changes and testing whether they differ. The rows below cover the two categories of outcome for which this consumer-review data supports a round-to-round test. Panel A reports the overall rating indices: our Core Review Index, the Overall Satisfaction item it is partly built from, and a Pooled 12-Item Index across every indicator in Figure D.1. Panel B reports the two privacy-relevant behavioral items (not ratings) also in the review data. This is not the full set of 12 pooled indicators taken individually, only these two groupings.

D.3. Accountability: what the round-to-round pattern shows.

For the overall rating index (Panel A), the pattern fits accountability's "immediate response" story better than a gradual learning curve: both the Core Review Index and Overall Satisfaction jump between round 1 and round 2 and then plateau, with the round-2-to-3 change statistically indistinguishable from zero – a one-time level shift that holds, not a curve that

Table D.1: Round-to-Round Dynamics in the Upgrade+Feedback Arm: Accountability vs. Learning

	Round 1	Round 2	Round 3	$\Delta(1\to2)$	$\Delta(2\to3)$	Test: $\Delta_{23} = \Delta_{12}$ Diff. (pp)	(SE)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Overall Rating Index							
Core Review Index (main text)	0.893	0.941	0.939	+0.048	-0.002	-0.052***	(0.019)
Overall Satisfaction (DFS Agent)	0.965	0.990	0.975	+0.025	-0.014	-0.046*	(0.025)
Pooled 12-Item Index	0.850	0.882	0.892	+0.032	+0.010	-0.025	(0.018)
B. Privacy-Relevant Outcomes							
No Third Person Present	0.657	0.815	0.856	+0.158	+0.041	-0.204**	(0.085)
Not Asked for PIN	0.394	0.382	0.460	-0.013	+0.079	+0.226***	(0.071)

NOTE: Sample is the same balanced panel of 70 communities (of 76) matched across all three review rounds used in the targeting check above; $N = 70$ throughout. Panel A reports the paper's overall rating index and the two other aggregate/summary ratings it draws on; Panel B reports the two privacy-relevant behavioral consumer-protection items (not ratings) that this review data separately supports a round-to-round test for. The Core Review Index reproduces this section's own baseline-index numbers exactly (a six-item average: tariff disclosure, fee rating, privacy rating, no-third-person, successful-transaction rating, and overall DFS-agent satisfaction). No Third Person Present and Not Asked for PIN are two of the individual items underlying that index and Figure D.1's pooled indicators. The Pooled 12-Item Index averages every indicator shown in Figure D.1. "Test: $\Delta_{23} = \Delta_{12}$" reports the coefficient (and robust SE) on a constant term from a regression of each community's (round-2-to-round-3 change − round-1-to-round-2 change) on cohort fixed effects: a negative, significant coefficient means the second-round-to-third-round gain was smaller than the first-round-to-second-round gain (deceleration), while a coefficient indistinguishable from zero means the two changes were of similar magnitude. See the discussion following this table for interpretation.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

keeps rising more slowly as genuine learning would predict. The Pooled 12-Item Index shows the same jump-then-plateau shape, but here the gap between the two round-to-round changes is only marginal ($p = 0.178$ implied by a coefficient of -0.025, SE 0.018). Taken together, the Core Review Index and Overall Satisfaction support reading the Upgrade+Feedback effect as agents responding immediately and substantially to being reviewed, rather than adjusting gradually over repeated cycles; but that reading does not extend uniformly to every individual item this data can speak to, and the one item (No Third Person Present) that does look like learning, and the one (Not Asked for PIN) that looks like neither story, are flagged here rather than folded into the headline claim.

D.4. Open-ended complaint text.

The customer questionnaire includes a free-text item ("Do you have any other feedback/comments/complain regarding this visit or your other visits to this agent over the past month?", variable `addit_feedback`), available only for the Upgrade+Feedback arm (the same limitation as

Table D.2: Open-Ended Complaint Themes by Review Round (Upgrade+Feedback Arm Only)

	Round 1 (1)	Round 2 (2)	Round 3 (3)
Responses (N)	3,532	2,021	2,296
Any substantive feedback (not blank/"no")	38.2%	43.7%	35.8%
Overcharging/fee complaints	0.51%	0.10%	0.09%
Distance/wait/network complaints	7.05%	4.95%	3.22%
Positive/thankful mentions	11.07%	7.08%	4.79%

NOTE: Simple keyword tabulation on the raw free-text field, not mutually exclusive categories. Overcharging and distance/wait/network complaints both fall sharply from round 1 to round 3, corroborating the paper's consumer-protection mechanism with the customers' own words rather than only the constructed binary indicators. Privacy- and PIN-specific keyword mentions (e.g., "pin", "third person") occur in well under 0.1 percent of responses in every round – too rare in free text to be informative on their own – so the structured yes/no items remain the right measure for that specific mechanism.

above, so this is a within-arm, across-round comparison, not a treatment-arm comparison). Table D.2 tabulates the share of responses raising specific themes, by round, pooled across all three cohorts.

E. Appendix: Survey Attrition

Table C.2 established that, even after correcting the community-name linkage problem described there, treatment-arm attrition remains modestly higher than control attrition: roughly 6–10 pp at the conservative (string-similarity ≥ 0.90) threshold used as this paper's primary estimate. This appendix asks the natural follow-up question directly: could this residual differential attrition, on its own, be driving the positive treatment effects reported in Tables 2–7, rather than the effects being real? Lee (2009) trimming bounds answer this by constructing the most pessimistic and most optimistic treatment effects consistent with the observed attrition rates, without requiring any assumption about *why* some VSLAs are missing from the endline sample.

E.1. Method.

For each treatment arm, the community-level retention rate (1 − the conservative attrition rate from Table C.2) is compared to its comparator's retention rate. For household- and secretary-level outcomes (Table C.3 Panel A, 3, 4, 6, and 7), the comparator is control, which is retained at a higher rate than either treatment arm in both the household sample (100.0% vs. 93.7% Upgrade / 94.5% Upgrade+Feedback) and the secretary sample (97.1% vs. 88.9% Upgrade / 91.8% Upgrade+Feedback), so control is the "over-retained" group there. For agent-level outcomes (Table C.3 Panels B–C and Table 2 Panel A), there is no control group; the comparison is Upgrade+Feedback versus Upgrade alone, and the retention ordering *reverses* relative to the household/secretary case – Upgrade+Feedback is retained

at a higher rate than Upgrade (94.5% vs. 88.9%), so in these agent-level columns it is Upgrade+Feedback's own distribution that gets trimmed to construct its bounds, not the comparator's. In every case, Lee bounds trim the over-retained group's community-level mean outcome down to the same effective retention rate as the under-retained group, by dropping the most extreme $(1-p)$ share of the over-retained group's community-level means, where p is the ratio of the lower retention rate to the higher one. Dropping the over-retained group's *lowest*-outcome communities raises that side of the comparison and yields the **lower** bound on the treatment effect (i.e., on β as defined by the main table, mean of arm-of-interest minus mean of comparator); dropping its *highest*-outcome communities lowers that side and yields the **upper** bound. Trimming and averaging are performed at the level of the randomization cohort – the same community grouping used for cluster-robust standard errors throughout this paper – rather than raw village geography, since a cohort can span villages in different chiefdoms or districts.

Every column of every regression table in this paper (Tables 2–7) is covered below. Each "Lee (2009) Bounds" row reported below is constructed on the simple, unadjusted community-level mean difference between the two compared arms (this underlying naive difference is not itself printed in the tables below, only its lower and upper bound), not on the covariate-adjusted, strata- and cohort-fixed-effect regression coefficient β_1/β_2 reported in the corresponding main-text table row directly above each bounds row; the two will generally differ in magnitude, though in every column checked they agree in sign. This is the standard, assumption-light implementation of Lee (2009): it answers whether differential attrition *on its own* could account for the sign and rough magnitude of a result, not a residualized re-estimate of the adjusted point estimate itself. For Table 7's two default-risk-restricted columns, the same overall secretary-level retention rates (97.1/88.9/91.8%) are applied to each restricted analytic subsample in turn, rather than a subsample-specific retention rate (not separately estimated in Table C.2); this is a simplifying approximation. One outcome, % Digital Contribution (Table 7), has a control-arm community-level distribution that is uniformly zero (every control VSLA reports 0% digital contribution), so trimming it changes nothing and its lower and upper bounds coincide exactly with the naive difference.

Each table below reproduces its corresponding main-text table's coefficients, standard errors, and $\beta_1 = \beta_2$/MHT test rows exactly as reported there, with three rows added directly beneath each coefficient's standard-error row: "Lee (2009) Bounds," reported as ⟨lower, upper⟩; "Trim %," the share of the over-retained group's communities dropped to equalize retention rates with the under-retained group (this depends only on the two arms' fixed retention rates from Table C.2, not on the outcome itself, so it repeats the same value across every household/secretary column for a given arm, and a different, agent-specific value across the agent-only columns); and "Imbens & Manski (2004) CS," reported as [lower, upper]. A column of "–" in any of the three rows matches the corresponding "–" in the coefficient row above it (no comparison exists for that arm/column, e.g. β_1 in the agent-only columns).

E.2. Imbens and Manski (2004) confidence set.

The Lee-bounds row alone is a point-identified range: no point estimate here carries a standard error, so it says nothing about how precisely L and U are estimated. Imbens and Manski (2004) address this by constructing a confidence set that is valid for the true

parameter – not just for the bounds – with the stated coverage probability no matter where in $[L, U]$ the truth lies. For each coefficient, L and U are re-estimated on 2,000 community-level bootstrap resamples (resampling communities with replacement, separately within each arm being compared, holding the arm's retention rate fixed at its Table C.2 value as in the point estimate), giving bootstrap standard errors $\hat{\sigma}_L$ and $\hat{\sigma}_U$. Writing $\hat{\sigma} = \max(\hat{\sigma}_L, \hat{\sigma}_U)$ and $\hat{\Delta} = (\hat{U} - \hat{L})/\hat{\sigma}$, the 95% Imbens-Manski critical value C_n solves $\Phi(C_n + \hat{\Delta}) - \Phi(-C_n) = 0.95$, and the reported confidence set is $[\hat{L} - C_n\hat{\sigma}_L,\ \hat{U} + C_n\hat{\sigma}_U]$; $C_n \to 1.960$ as $\hat{\Delta} \to 0$ (point identification) and $C_n \to 1.645$ as $\hat{\Delta} \to \infty$ (bounds much wider than their own sampling noise), so it is never a plain 95% normal interval on either endpoint alone. This construction is mechanically wider than the Lee-bounds row above it in every column, by design – it is answering a strictly harder question (does the confidence set, not just the point bounds, exclude zero) – so a column where the Lee bounds exclude zero but the Imbens-Manski CS includes it is not a contradiction, it is the CS doing the job the point bounds cannot: accounting for how much the trimmed community-level means could have come out differently by chance.

Table E.1: Lee Bounds for Upgrading, Feedback, and Provider Behavior

	A. Provider Access (Household)				B/C. Agent Feedback/Behavior	
	# Agents in Comm. (1)	Access: Inside (2)	Access: Outside (3)	KLK Index (4)	Received Feedback (5)	Made Changes (6)
Upgrade (β_1)	0.403***	0.188***	-0.100**	0.354***	–	–
	(0.096)	(0.047)	(0.050)	(0.082)		
Lee (2009) Bounds	⟨0.470, 0.653⟩	⟨0.163, 0.197⟩	⟨-0.067, 0.013⟩	⟨0.282, 0.397⟩	–	–
Trim %	6.3%	6.3%	6.3%	6.3%	–	–
Imbens & Manski (2004) CS	[0.276, 0.830]	[0.108, 0.244]	[-0.165, 0.111]	[0.171, 0.504]	–	–
Upgrade+Feedback (β_2)	0.508***	0.194***	-0.120***	0.406***	0.211**	0.205***
	(0.098)	(0.049)	(0.043)	(0.079)	(0.081)	(0.076)
Lee (2009) Bounds	⟨0.343, 0.490⟩	⟨0.178, 0.208⟩	⟨-0.124, -0.061⟩	⟨0.306, 0.400⟩	⟨0.189, 0.256⟩	⟨0.223, 0.291⟩
Trim %	5.5%	5.5%	5.5%	5.5%	5.9%	5.9%
Imbens & Manski (2004) CS	[0.186, 0.633]	[0.121, 0.258]	[-0.217, 0.031]	[0.204, 0.498]	[0.045, 0.400]	[0.087, 0.427]
Test: $\beta_1 = \beta_2$ (p)	0.185	0.888	0.585	0.424	–	–
MHT (RW p), β_1	0.001	0.001	0.001	–	–	–
MHT (RW p), β_2	0.001	0.001	0.001	–	0.047	0.047
Observations	2,613	2,619	2,619	2,619	136	137
Mean, dep. var. (ref. grp.)	0.54	0.06	0.32	-0.32	0.38	0.21
SD, dep. var. (ref. grp.)	0.75	0.23	0.47	0.51	0.49	0.41
Data source	HH	HH	HH	HH	Agent	Agent

NOTE: Coefficients, SEs, tests, and sample statistics reproduce Table C.3 exactly. Lee (2009) Bounds rows are unadjusted community-level mean-difference bounds as described in the Method paragraph above; in the Received Feedback/Made Changes columns (no control group), the comparator is Upgrade and it is Upgrade+Feedback's own distribution that is trimmed, since Upgrade+Feedback is the higher-retention arm there.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table E.2: Lee Bounds for Provider Effort and Consumer Trust

	A. Agent Privacy Practices			B. Consumer Trust In:					
	3rd Person Present (1)	PIN Req. (2)	KLK Privacy (3)	Family/ Friends (4)	Comm. People (5)	Local VSLA (6)	Momo Agent (7)	Banking Agent (8)	KLK Index (9)
Upgrade (β_1)	–	–	–	-0.028*	-0.009	-0.020	0.069**	0.009	-0.000
				(0.016)	(0.032)	(0.012)	(0.032)	(0.047)	(0.062)
Lee (2009) Bounds	–	–	–	⟨-0.081, -0.039⟩	⟨-0.046, 0.034⟩	⟨-0.065, 0.001⟩	⟨-0.080, 0.001⟩	⟨0.011, 0.091⟩	⟨-0.152, 0.031⟩
Trim %	–	–	–	6.3%	6.3%	6.3%	6.3%	6.3%	6.3%
Imbens & Manski (2004) CS	–	–	–	[-0.117, 0.013]	[-0.119, 0.111]	[-0.104, 0.060]	[-0.154, 0.081]	[-0.091, 0.193]	[-0.303, 0.199]
Upgr.+Feedback (β_2)	-0.132**	-0.160	-0.249**	-0.018	0.004	-0.025**	0.067**	-0.012	0.007
	(0.059)	(0.141)	(0.114)	(0.017)	(0.033)	(0.012)	(0.030)	(0.048)	(0.060)
Lee (2009) Bounds	⟨-0.146, -0.079⟩	⟨-0.022, 0.181⟩	⟨-0.170, 0.004⟩	⟨-0.060, -0.022⟩	⟨-0.034, 0.029⟩	⟨-0.039, 0.025⟩	⟨-0.036, 0.028⟩	⟨0.062, 0.126⟩	⟨-0.050, 0.101⟩
Trim %	5.9%	5.9%	5.9%	5.5%	5.5%	5.5%	5.5%	5.5%	5.5%
Imbens & Manski (2004) CS	[-0.258, 0.034]	[-0.288, 0.448]	[-0.379, 0.222]	[-0.095, 0.027]	[-0.111, 0.108]	[-0.081, 0.081]	[-0.111, 0.105]	[-0.041, 0.229]	[-0.189, 0.256]
Test: $\beta_1 = \beta_2$ (p)	–	–	–	0.381	0.536	0.724	0.951	0.608	0.892
MHT (RW p), β_1	–	–	–	–	0.849	0.190	0.002	0.849	–
MHT (RW p), β_2	0.131	0.663	–	–	0.899	0.062	0.003	0.899	–
Obs.	136	136	136	2,619	2,619	2,619	2,619	2,619	2,619
Mean (ref. grp.)	0.24	2.12	0.00	0.96	0.86	0.97	0.79	0.61	0.00
SD (ref. grp.)	0.43	0.95	0.77	0.20	0.35	0.17	0.41	0.49	0.58
Data source	Agent	Agent	Agent	HH	HH	HH	HH	HH	HH

NOTE: Coefficients, SEs, tests, and sample statistics reproduce Table 2 exactly. In Panel A (no control group), the comparator is Upgrade and Upgrade+Feedback's own distribution is trimmed, as in Table C.3 Panels B–C. PIN Req., Comm. People, Banking Agent, and KLK Index are the (unstarred, near-zero) columns whose Lee bounds include zero. Of this table's six starred coefficients, only two – 3rd Person Present and Family/Friends trust (Upgrade) – retain a Lee-bounds interval excluding zero; the remaining four (PCA Privacy, Local VSLA (Upgrade+Feedback), and Momo Agent trust under both arms) have Lee-bounds intervals that include zero outright. None of the six survives the stricter Imbens-Manski check – see the Interpretation and Imbens-Manski paragraphs below.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table E.3: Lee Bounds for Adoption and Formal Account Ownership

	Owns DFS Account (1)	DFS: Momo Indicator (2)	# Momo Accounts (3)	# Bank Accounts (4)	KLK Index (5)
Upgrade (β_1)	0.052*	0.044	0.083*	0.043***	0.111*
	(0.029)	(0.030)	(0.049)	(0.011)	(0.063)
Lee (2009) Bounds	⟨0.060, 0.141⟩	⟨0.038, 0.119⟩	⟨0.222, 0.369⟩	⟨0.095, 0.099⟩	⟨0.130, 0.302⟩
Trim %	6.3%	6.3%	6.3%	6.3%	6.3%
Imbens & Manski (2004) CS	[-0.032, 0.233]	[-0.055, 0.212]	[0.068, 0.509]	[0.057, 0.136]	[-0.061, 0.491]
Upgrade+Feedback (β_2)	0.091***	0.092***	0.179***	0.037***	0.202***
	(0.029)	(0.029)	(0.058)	(0.010)	(0.064)
Lee (2009) Bounds	⟨-0.023, 0.040⟩	⟨-0.035, 0.028⟩	⟨0.084, 0.211⟩	⟨0.047, 0.051⟩	⟨-0.046, 0.089⟩
Trim %	5.5%	5.5%	5.5%	5.5%	5.5%
Imbens & Manski (2004) CS	[-0.115, 0.132]	[-0.125, 0.118]	[-0.052, 0.337]	[0.024, 0.073]	[-0.241, 0.284]
Test: $\beta_1 = \beta_2$ (p)	0.185	0.116	0.118	0.623	0.161
MHT (RW p), β_1	0.045	0.101	0.101	0.006	–
MHT (RW p), β_2	0.002	0.002	0.002	0.003	–
Observations	2,619	2,619	2,613	2,618	2,612
Mean, dep. var. (control)	0.23	0.22	0.51	0.01	-0.19
SD, dep. var. (control)	0.42	0.42	0.73	0.10	0.88

NOTE: Coefficients, SEs, tests, and sample statistics reproduce Table 3 exactly. Three of this table's nine starred coefficients – Owns DFS Account, DFS: Momo Indicator, and KLK Index, each under the Upgrade+Feedback arm – now have Lee-bounds intervals that include zero; the remaining six retain the sign of the naive difference above them. Only three starred coefficients (# Momo Accounts and # Bank Accounts under Upgrade, and # Bank Accounts under Upgrade+Feedback) survive the stricter Imbens-Manski check – see the Interpretation and Imbens-Manski paragraphs below.

***

Table E.4: Lee Bounds for Usage and Adoption–Usage Differences

	Used FS	Used DFS	DFS: Momo	DFS: GT Bank	KLK Index	Adopt but Don't Use
	(1)	(2)	(3)	(4)	(5)	(6)
Upgrade (β_1)	0.028	0.025	0.043	0.014	0.073	-0.018
	(0.043)	(0.042)	(0.044)	(0.009)	(0.082)	(0.037)
Lee (2009) Bounds	⟨-0.014, 0.067⟩	⟨0.094, 0.175⟩	⟨0.108, 0.188⟩	⟨0.017, 0.020⟩	⟨0.195, 0.358⟩	⟨0.012, 0.073⟩
Trim %	6.3%	6.3%	6.3%	6.3%	6.3%	6.3%
Imbens & Manski (2004) CS	[-0.122, 0.175]	[-0.001, 0.270]	[0.014, 0.282]	[0.000, 0.036]	[0.010, 0.541]	[-0.073, 0.148]
Upgrade+Feedback (β_2)	0.067*	0.082**	0.104***	0.013*	0.190**	-0.062*
	(0.036)	(0.038)	(0.039)	(0.008)	(0.074)	(0.033)
Lee (2009) Bounds	⟨0.164, 0.227⟩	⟨0.029, 0.092⟩	⟨0.026, 0.089⟩	⟨0.010, 0.013⟩	⟨0.071, 0.199⟩	⟨-0.026, 0.035⟩
Trim %	5.5%	5.5%	5.5%	5.5%	5.5%	5.5%
Imbens & Manski (2004) CS	[0.061, 0.330]	[-0.071, 0.191]	[-0.074, 0.189]	[0.002, 0.019]	[-0.112, 0.382]	[-0.107, 0.105]
Test: $\beta_1 = \beta_2$ (p)	0.211	0.134	0.118	0.945	0.117	0.119
MHT (RW p), β_1	0.470	0.588	0.288	0.182	–	0.644
MHT (RW p), β_2	0.005	0.005	0.002	0.090	–	0.090
Observations	2,619	2,619	2,619	2,619	2,619	833
Mean, dep. var. (control)	0.66	0.42	0.35	0.005	-0.17	0.13
SD, dep. var. (control)	0.47	0.49	0.48	0.068	0.91	0.33

NOTE: Coefficients, SEs, tests, and sample statistics reproduce Table 4 exactly. Adopt-but-Don't-Use (β_2) is the only starred coefficient in this table whose Lee-bounds interval includes zero – it is not, however, the only such case paper-wide; eight other starred coefficients across the other Lee-bounds tables share this outcome (see the Interpretation paragraph below). Used DFS, DFS: Momo, and KLK Index (β_2 in each case) survive the Lee-bounds check but not the stricter Imbens-Manski one; only Used FS and DFS: GT Bank (β_2) survive both. Adopt-but-Don't-Use (β_1, unstarred) additionally has a naive-difference sign opposite its (unstarred, not significant) main-text coefficient. See the Interpretation and Imbens-Manski paragraphs below.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table E.5: Lee Bounds for Usage by Transaction Channel

	Deposit (Cash-In) (1)	Withdraw (Cash-Out) (2)	Peer-to-Peer (3)	Savings Instr.: DFS (4)	Savings Instr.: Inf. (5)	KLK Index (6)
Upgrade (β_1)	0.042*	0.066	0.011	0.078***	-0.011	0.140**
	(0.025)	(0.043)	(0.012)	(0.023)	(0.042)	(0.056)
Lee (2009) Bounds	⟨0.131, 0.153⟩	⟨0.120, 0.201⟩	⟨0.031, 0.053⟩	⟨0.179, 0.237⟩	⟨-0.117, -0.037⟩	⟨0.333, 0.434⟩
Trim %	6.3%	6.3%	6.3%	6.3%	6.3%	6.3%
Imbens & Manski (2004) CS	[0.083, 0.200]	[0.030, 0.288]	[-0.007, 0.079]	[0.110, 0.294]	[-0.218, 0.064]	[0.209, 0.539]
Upgrade+Feedback (β_2)	0.006	0.092**	0.046***	0.066***	-0.057	0.166***
	(0.024)	(0.037)	(0.013)	(0.024)	(0.042)	(0.053)
Lee (2009) Bounds	⟨0.031, 0.049⟩	⟨0.039, 0.102⟩	⟨0.016, 0.036⟩	⟨0.106, 0.159⟩	⟨-0.008, 0.056⟩	⟨0.141, 0.228⟩
Trim %	5.5%	5.5%	5.5%	5.5%	5.5%	5.5%
Imbens & Manski (2004) CS	[-0.001, 0.079]	[-0.045, 0.186]	[-0.014, 0.050]	[0.041, 0.214]	[-0.105, 0.153]	[0.030, 0.318]
Test: $\beta_1 = \beta_2$ (p)	0.141	0.450	0.010	0.648	0.215	0.645
MHT (RW p), β_1	0.020	0.020	0.189	0.001	–	0.001
MHT (RW p), β_2	0.738	0.002	0.001	0.005	–	0.001
Observations	2,619	2,619	2,619	2,619	2,619	2,619
Mean, dep. var. (control)	0.089	0.274	0.017	0.125	0.596	-0.197
SD, dep. var. (control)	0.285	0.446	0.129	0.331	0.491	0.556

NOTE: Coefficients, SEs, tests, and sample statistics reproduce Table 6 exactly (see that table's note on the KLK Index's construction from the five usage-channel variables). Savings Instr.: Informal (β_1) is the (unstarred) column where the Lee-bounds interval includes zero; every starred coefficient in this table, including the KLK Index under both arms, excludes zero at the Lee-bounds level. At the stricter Imbens-Manski level, however, Withdraw (Cash-Out) and Peer-to-Peer under the Upgrade+Feedback arm now include zero, while Deposit (Cash-In) (Upgrade), Savings Instr.: DFS (both arms), and the KLK Index (both arms) still exclude zero at both levels.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table E.6: Lee Bounds for Spillover Effects on VSLA Business

	Income/Wk	# Current	Contribution	% Digital	# Members	Default Risk (Count < 4)		KLK
	(SLL) (1)	Members (2)	Amount (SLL) (3)	Contrib. (4)	Defaulted (5)	Low-Risk (6)	High-Risk (7)	Index (8)
Upgrade (β_1)	51.6	0.72	455.0	3.20	-0.55	-0.42	0.05	0.220
	(62.8)	(3.32)	(714.6)	(2.36)	(0.76)	(0.30)	(1.68)	(0.267)
Lee (2009) Bounds	⟨66.2, 86.7⟩	⟨-0.72, 3.72⟩	⟨430.8, 1185.0⟩	⟨2.54, 2.54⟩	⟨-1.21, -0.05⟩	⟨-0.59, -0.24⟩	⟨-0.62, 1.18⟩	⟨0.26, 0.47⟩
Trim %	8.4%	8.4%	8.4%	8.4%	8.4%	8.4%	8.4%	8.4%
Imbens & Manski (2004) CS	[8.7, 139.7]	[-5.57, 7.71]	[-544.2, 2059.2]	[-0.50, 5.58]	[-2.43, 0.75]	[-1.10, 0.27]	[-3.55, 3.45]	[-0.02, 0.74]
Upgrade+Feedback (β_2)	159.7*	1.28	1009.0	6.88***	1.20	-0.72**	0.70	0.529
	(94.5)	(2.83)	(695.6)	(2.51)	(0.86)	(0.29)	(1.35)	(0.330)
Lee (2009) Bounds	⟨230.9, 249.9⟩	⟨0.41, 4.06⟩	⟨866.7, 1468.7⟩	⟨5.64, 5.64⟩	⟨0.32, 1.32⟩	⟨-0.97, -0.75⟩	⟨-0.06, 1.74⟩	⟨0.74, 0.92⟩
Trim %	5.5%	5.5%	5.5%	5.5%	5.5%	5.5%	5.5%	5.5%
Imbens & Manski (2004) CS	[42.1, 436.5]	[-4.25, 7.98]	[-202.5, 2455.0]	[1.74, 9.53]	[-1.05, 2.35]	[-1.45, -0.26]	[-3.15, 4.18]	[0.14, 1.52]
Test: $\beta_1 = \beta_2$ (p)	0.239	0.856	0.457	0.182	0.007	0.092	0.698	0.361
MHT (RW p), β_1	0.897	0.897	0.897	0.747	0.897	–	–	–
MHT (RW p), β_2	0.409	0.637	0.409	0.146	0.409	–	–	–
Observations	172	174	174	173	174	120	54	174
Mean, dep. var. (control)	18.5	29.6	1694.1	0.00	2.6	1.03	6.73	0.00
SD, dep. var. (control)	82.1	12.5	2141.3	0.00	3.9	–	–	0.56

NOTE: Coefficients, SEs, tests, and sample statistics reproduce Table 7 exactly. Lee bounds for the two default-risk-restricted columns apply the same overall secretary-level retention rates to each restricted subsample (see Method paragraph). % Digital Contribution's bounds coincide exactly with its naive difference because control's community-level distribution is uniformly zero (a degenerate comparator, not an error). # Current Members (both β_1 and β_2, unstarred) and Default Risk: High (Count, both arms, unstarred) are the columns whose Lee-bounds interval includes zero; every starred coefficient in this table excludes zero at both the Lee-bounds and Imbens-Manski levels.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

E.3. Interpretation.

Across the roughly eighty coefficient-level Lee bounds reported in Tables E.1–E.6, the large majority of starred coefficients retain a Lee-bounds interval that excludes zero and shares its sign, but nine do not. (Throughout this and the following subsection, a specific Lee-bounds interval or Imbens-Manski confidence set discussed alongside a main-text table's name is reported in that table's appendix counterpart, not in the main-text table itself: Table E.1 for Table C.3, Table E.2 for Table 2, Table E.3 for Table 3, Table E.4 for Table 4, Table E.5 for Table 6, and Table E.6 for Table 7 – each appendix table's own note confirms its point estimates reproduce its main-text counterpart exactly.) The nine are: Table C.3's Access: Outside effect under Upgrade alone ($\beta_1 = -0.100$, ⟨-0.067, 0.013⟩); Table 2's KLK Privacy ($\beta_2 = -0.249$, ⟨-0.170, 0.004⟩), Local VSLA trust ($\beta_2 = -0.025$, ⟨-0.039, 0.025⟩), and Momo Agent trust under both arms ($\beta_1 = 0.069$, ⟨-0.080, 0.001⟩; $\beta_2 = 0.067$, ⟨-0.036, 0.028⟩); Table 3's Owns DFS Account, DFS: Momo Indicator, and KLK Index, all under Upgrade+Feedback (⟨-0.023, 0.040⟩, ⟨-0.035, 0.028⟩, and ⟨-0.046, 0.089⟩ respectively); and Table 4's Adopt-but-Don't-Use ($\beta_2 = -0.062$, significant at 10%, ⟨-0.026, 0.035⟩), which is also the weakest (10%) significance tier in its table and already fails Romano–Wolf multiple-hypothesis correction at conventional levels (Table 4's own MHT row). Every other starred coefficient across Tables E.1–E.6 – including Table 4's Used FS ($\beta_2 = 0.067$, ⟨0.164, 0.227⟩) and Table 7's count-based Default Risk: Low result ($\beta_2 = -0.72$, ⟨-0.968, -0.746⟩) – retains a Lee-bounds interval excluding zero. Residual differential attrition, at the modest 5.5–8.4 pp range documented in Table C.2, is therefore not large enough on its own to manufacture the paper's central treatment effects at the point-identified Lee-bounds level, for all but these nine columns; for them, as the stricter check below also confirms, attrition cannot be as confidently ruled out. This exercise is a complementary, assumption-light check on the main tables' covariate-adjusted specifications, not a substitute for them.

E.4. Imbens-Manski confidence sets: a stricter check.

The Lee-bounds row just discussed ignores sampling uncertainty in L and U; the Imbens-Manski CS row beneath it does not, and is mechanically wider in every column for that reason (see the Imbens-Manski Method paragraph above). This is a materially stricter test: of the 41 starred coefficients across Tables E.1–E.6, 20 – just under half – now have a confidence set that includes zero.

Table C.3 is robust apart from one outcome: Access: Outside fails under both arms (Upgrade, CS $[-0.165, 0.111]$; Upgrade+Feedback, CS $[-0.217, 0.031]$); the other eight starred coefficients (# Agents in Comm., Access: Inside, and KLK Index under both arms, and Received Feedback/Made Changes under Upgrade+Feedback) survive. Table 2 is where this stricter check matters most: of its six starred coefficients, none retains a confidence set excluding zero – 3rd Person Present ($[-0.258, 0.034]$), KLK Privacy ($[-0.379, 0.222]$), Family/Friends trust under Upgrade ($[-0.117, 0.013]$), Local VSLA trust under Upgrade+Feedback ($[-0.081, 0.081]$), and Momo Agent trust under both arms (Upgrade, $[-0.154, 0.081]$; Upgrade+Feedback, $[-0.111, 0.105]$) all include zero. Table 3 fares better but is not fully robust: of its nine starred coefficients, three survive (# Momo Accounts and # Bank Accounts under Upgrade, and # Bank Accounts under Upgrade+Feedback), while the other six – Owns DFS

Account under both arms, DFS: Momo Indicator and KLK Index under Upgrade+Feedback, and # Momo Accounts under Upgrade+Feedback – include zero (e.g. Owns DFS Account, Upgrade+Feedback, $[-0.115, 0.132]$; KLK Index, Upgrade+Feedback, $[-0.241, 0.284]$). Table 4 splits evenly: Used FS and DFS: GT Bank (both Upgrade+Feedback) survive, while Used DFS ($[-0.071, 0.191]$), DFS: Momo ($[-0.074, 0.189]$), KLK Index ($[-0.112, 0.382]$), and Adopt-but-Don't-Use ($[-0.107, 0.105]$), all Upgrade+Feedback, include zero. Table 6 is mostly robust: five of its seven starred coefficients survive (Deposit under Upgrade, and Savings Instr.: DFS and KLK Index under both arms), while Withdraw and Peer-to-Peer, both under Upgrade+Feedback, newly include zero ($[-0.045, 0.186]$ and $[-0.014, 0.050]$ respectively). Table 7 is fully robust: all three of its starred coefficients – Business Income, % Digital Contribution, and the count-based Default Risk: Low result ($[-1.454, -0.261]$) – keep a confidence set excluding zero.

None of this means the individual-item point estimates are wrong; it means that, once the bounding exercise's own estimation uncertainty is accounted for at the stringent, uniformly-valid level Imbens-Manski (2004) requires, differential attrition cannot be as confidently ruled out as a partial explanation for a substantial share of this paper's reported effects – concentrated especially in Table 2's consumer-trust results and, to a lesser extent, Table 3's adoption results – while the VSLA-business spillover results in Table 7 and most of the usage-channel results in Table 6 hold up fully. As with the Lee-bounds row, this is a complementary robustness check layered on top of the main tables' own covariate-adjusted specifications, not a re-estimate of them, and a coefficient failing this check is not thereby shown to be attrition-driven – only that this check, unlike the point Lee bounds above it, cannot rule attrition out as a contributor.

F. Appendix: Cross-Market Spillovers

Section 7.C motivates this check and defines the concern; this appendix presents the full chiefdom-level placebo regressions.

F.1. Method

Markets are grouped by *chiefdom* rather than district for this check. Chiefdoms are the administrative tier directly beneath district in Sierra Leone, and this sample's three districts (Kambia, Port Loko, Kenema) contain eight chiefdoms between them (four, two, and two, respectively), each with its own communities and its own realized treated share (ranging from 62% to 86% across chiefdoms, versus 75–79% at the coarser district level). For each community i in chiefdom c a leave-one-out treated-market share is constructed: the fraction of *other* communities in i's own chiefdom that were assigned to Upgrade or Upgrade+Feedback, excluding i itself. This is then interacted with i's own treatment status in the regression given by Equation 4 in Section 7.C above, with standard errors clustered at the community level, matching every other regression in this paper. Since $\texttt{Treated}_i = 0$ for control units, β_2 is exactly the marginal association between a control unit's own outcome and how treatment-saturated its chiefdom is – "does Control look significantly different where more of its chiefdom was treated" – which is what this appendix reports, now run across every outcome in Tables E.1–E.6 rather than one headline column per table. Agent-level outcomes

(Table C.3 Panels B/C, Table 2 Panel A) are excluded from this particular check: the agent sample has no control arm (every agent surveyed is in either Upgrade or Upgrade+Feedback), so a test of whether *control* units drift toward treated levels as their area's treatment saturation rises does not have a well-defined control group to apply to on that sample; extending this idea to a treated-treated comparison would be a different, more speculative exercise, and is left to future work. Markets are grouped by chiefdom rather than district – eight units instead of three – because the leave-one-out treated-share regressor then varies *within* district as well as between districts (Kambia's own four chiefdoms range from 75% to 80% treated, for instance), so β_2 is not mechanically collinear with a district fixed effect the way it would be at the coarser district level; it remains a modest number of independent geographic units to lean on, and chiefdom-level confounders unrelated to spillover (market access, agent density, local economy) remain a possible alternative explanation for a significant β_2. Because chiefdoms nest inside districts, adding an explicit district fixed effect alongside the chiefdom-level share provides a genuine test—rather than a tautological one, which it would be if both terms were defined at the district level—of whether a chiefdom-level result reflects within-district variation or is itself picking up broader district-level differences, which is why both specifications are reported below.

F.2. Results

The chiefdom-level specification above is estimated separately for each of the 37 household and secretary outcomes in Tables E.1–E.6 (74 regressions in total), without and with an added district fixed effect, reporting $\hat{\beta}_2$ in both cases. The last four rows below are the shock-mitigation measures of Table 8, added to this check's outcome list after the original 34-outcome version of this table was written:

Table F.1: Chiefdom-Level Placebo Regressions: Effect of Local Treatment Saturation on Control-Group Outcomes (Full 37-Outcome Set)

	No district FE			With district FE		
Outcome	$\hat{\beta}_2$ (1)	SE (2)	p (3)	$\hat{\beta}_2$ (4)	SE (5)	p (6)
# Agents in Comm. (Table C.3)	2.173	1.560	0.166	2.158	1.568	0.171
Access: Inside (Table C.3)	0.746	0.577	0.198	0.802	0.588	0.174
Access: Outside (Table C.3)	-0.130	0.622	0.835	0.161	0.524	0.759
KLK Index (Access) (Table C.3)	1.407	1.050	0.182	1.252	1.045	0.233
Trust: Family/Friends (Table 2)	0.331	0.259	0.202	0.219	0.254	0.391
Trust: Comm. People (Table 2)	0.564	0.497	0.259	0.149	0.438	0.735
Trust: Local VSLA (Table 2)	0.121	0.120	0.316	0.119	0.129	0.358
Trust: Momo Agent (Table 2)	0.506	0.379	0.183	0.413	0.392	0.293
Trust: Banking Agent (Table 2)	0.393	0.574	0.495	0.144	0.607	0.812
KLK Index (Trust) (Table 2)	1.456	0.805	0.072	0.828	0.758	0.277
Owns DFS Account (Table 3)	-0.497	0.423	0.241	-0.296	0.419	0.480
DFS: Momo Indicator (Table 3)	-0.565	0.424	0.184	-0.334	0.406	0.412
# Momo Accounts (Table 3)	-0.894	0.978	0.362	-0.454	0.899	0.614
# Bank Accounts (Table 3)	0.170	0.091	0.062	0.162	0.117	0.167
KLK Index (Adoption) (Table 3)	-1.106	0.902	0.222	-0.641	0.877	0.466

continued on next page

Table F.1 – continued from previous page

Outcome	No district FE			With district FE		
	$\hat{\beta}_2$ (1)	SE (2)	p (3)	$\hat{\beta}_2$ (4)	SE (5)	p (6)
Used FS (Table 4)	0.469	0.557	0.401	0.543	0.547	0.322
Used DFS (Table 4)	-0.289	0.553	0.602	-0.183	0.500	0.715
DFS: Momo (Table 4)	0.327	0.622	0.600	0.538	0.594	0.365
DFS: GT Bank (Table 4)	-0.018	0.061	0.765	-0.037	0.077	0.636
KLK Index (Usage) (Table 4)	0.180	1.051	0.865	0.505	0.976	0.606
Adopt-Use Gap (Table 4)†	0.719	0.287	0.013	0.761	0.268	0.005
Deposit (Cash-In) (Table 6)	-0.558	0.435	0.201	-0.568	0.436	0.195
Withdraw (Cash-Out) (Table 6)	0.805	0.662	0.226	0.795	0.703	0.260
Peer-to-Peer (Table 6)	-0.017	0.091	0.856	-0.119	0.114	0.297
Savings Instr.: DFS (Table 6)	0.561	0.382	0.143	0.740	0.384	0.055
Savings Instr.: Inf. (Table 6)	1.124	0.770	0.146	1.355	0.768	0.080
KLK Index (Channels) (Table 6)	0.465	0.622	0.455	0.457	0.642	0.478
Business Income (Table 7)	-217.1	395.9	0.584	-849.0	639.2	0.186
# Current Members (Table 7)	14.947	23.477	0.525	-2.692	25.560	0.916
Contribution Amount (Table 7)	1864.6	4449.9	0.676	-4377.0	5148.9	0.396
% Digital Contrib. (Table 7)	-6.598	10.915	0.546	-6.926	10.964	0.528
# Members Defaulted (Table 7)	-7.772	9.501	0.414	-4.178	9.558	0.662
KLK Index (Business) (Table 7)	0.351	1.621	0.829	-2.869	2.459	0.245
Any Shock (Table 8)	-0.638	0.367	0.084	-0.487	0.362	0.180
Unmitigated Shock (Table 8)	-0.516	0.543	0.344	-0.522	0.529	0.324
Shock Index: Mean (Table 8)	-0.577	0.342	0.093	-0.505	0.323	0.120
Shock Index: Any (Table 8)	-0.625	0.363	0.087	-0.503	0.362	0.167

The overall pattern is reassuring: 36 of 37 outcomes show no significant relationship between chiefdom-level treatment saturation and control-group outcomes at conventional levels, with or without the district fixed effect – this includes all four shock-mitigation rows, none of which is significant at conventional levels in either specification (Any Shock comes closest, at $p = 0.084$ without the district fixed effect, but loses even that at $p = 0.180$ once it is added). In particular, DFS: Momo Indicator and Used DFS—two outcomes that show a spurious relationship with treatment saturation when it is measured at the coarser district level—are *not* significant at the finer chiefdom level ($p = 0.184$ and $p = 0.602$ respectively), consistent with the coarser measure's apparent signal being an artifact of collinearity between district-level treatment saturation and the district fixed effects themselves, rather than genuine spillover.

One outcome is the exception: Adopt-Use Gap shows a positive, significant $\hat{\beta}_2$ both without a district fixed effect (0.719, $p = 0.013$) and, if anything more strongly, with one added (0.761, $p = 0.005$). Because this result survives conditioning on district – unlike a spurious signal, which the Method above predicts would be absorbed by the district fixed effect – it is a genuine candidate for chiefdom-level spillover. It should be read in context: the underlying sample for this outcome is restricted to `dfs_own==1` households and is accordingly smaller than the other columns, and a single significant result out of 37 checked outcomes is within the range one might expect from chance alone even after accounting for its survival under the district-FE check. It does not overturn any headline conclusion in this paper. †Flagged

above as the one outcome not explained by the chiefdom-level data's residual coarseness (see Method).

G. Appendix: Hawthorne and Demand Effects

Section 7.D motivates this concern (Hawthorne and John Henry effects) and summarizes why it is limited here. This appendix does not run a new statistical test for these effects – unlike attrition or spillovers, there is no dataset-internal placebo or bounding exercise available to test for experimenter demand directly – but lays out the full reasoning behind Section 7.D's summary, and where the limits of that argument are.

G.1. Why Hawthorne effects are a limited concern here

Several design features work against a pure Hawthorne story, in which effects arise merely from being observed rather than from the intervention itself. First, the treatment is a physical, structural change – installing or formalizing a DFS/mobile-money agent relationship for a VSLA, with associated training and (in the Upgrade+Feedback arm) a feedback mechanism – not a motivational message, a request to change behavior, or anything a respondent could simply decide to perform without the underlying infrastructure actually being present. A household cannot credibly report using a mobile-money agent it has no access to, and administrative/behavioral outcomes like DFS transaction usage (Table 4) or a secretary's recorded VSLA business figures (Table 7) are harder to fabricate under observation than an attitudinal survey item. Second, all three arms – including Control – were surveyed with the identical endline instrument, by the same enumerator pool, on the same timeline; the survey itself does not tell a respondent which arm they are in or what the "expected" answer might be, unlike a lab or framed-field experiment where subjects are explicitly told they are receiving a treatment. Third, enumerators were kept blind to treatment status during data collection, and so the channel by which an enumerator might consciously or unconsciously cue "expected" responses was closed off. Fourth, the pattern of results itself is not what a pure demand story would predict: if simply being enrolled in a study of DFS adoption made respondents report or behave in ways pleasing to the research team, the Upgrade-alone arm – which received the identical survey attention as Upgrade+Feedback – should show similarly inflated effects across the board; instead, Upgrade-alone is frequently insignificant while Upgrade+Feedback is not (Tables 3–6), a pattern that tracks the added feedback mechanism rather than mere awareness of being studied.

G.2. Why John Henry effects are a limited concern here

A John Henry effect would require the control arm to compensate for being denied the intervention by unilaterally working harder to close the gap. Two features make this less plausible here. First, the intervention is again structural rather than motivational: a Control VSLA cannot simply try harder to become a formal DFS agent site, since that requires an actual agent relationship and equipment the study did not provide to them; consistent with this, the endline data show no instance of a Control-arm VSLA having become a mobile money

or banking retail agent by endline – unsurprising given the structural nature of the intervention, but a useful direct check against this concern nonetheless. Second, randomization was conducted at the VSLA/community level, not within a shared community, so Control and Treatment groups are physically located in different villages; a Control respondent has no obvious daily point of comparison against a specific treated peer group, reducing the salience needed for a competitive response. (Note that this is a different concern from Appendix F's cross-market spillovers, above, which asks whether treated *markets* economically affect nearby control markets through agent traffic or information diffusion, not whether control *respondents* psychologically compete with treated ones.)

H. Appendix: Field Exhibits and Data Collection

This appendix provides visual context on the study setting and intervention, reports sample survey items from the household, secretary, and agent endline instruments, and reproduces excerpts of the standard agent agreement form used to onboard VSLAs. Full instruments and agreement forms are available from the authors upon request.

H.1. Setting: Terrain and the VSLA Money Box

Figure H.1: Field Setting: Terrain and the VSLA Money Box



(a) Inaccessible terrain poses significant access challenges, frequently requiring river crossings to reach isolated communities.



(b) The VSLA "money box" comprises three padlocks held by the VSLA chairman, the DFS agent, and the community chief. Each holds their own key, and the box can be opened only when all three parties are present. Members deposit their contributions into the box weekly, every Friday.

H.2. Treatment Implementation: DFS Agents and Community Barrays

Figure H.2: Treatment Implementation: DFS Agents and Community Barrays



(a) A mobile-money and bank agent created to serve as a financial champion for the VSLA and to provide formal financial services to the wider community. Each agent is supplied with a solar panel, two phones, a loudspeaker, an umbrella, a bracelet, a table, and initial capital of NLe 1,500.



(b) A community-court "barray" gathering to implement the evaluation and feedback mechanism. In towns and villages across Sierra Leone, the barray functions as a community-based public forum for discussion, dispute resolution, and collective decision-making, akin to an informal court system.

H.3. Survey Instruments: Select Measurement Questions

Items below are grouped by outcome domain and reproduced verbatim from the underlying instruments.

A. Digital Financial Services: Usage and Adoption (Table 4)

Qx. [DFS account ownership] Which of the following DFS accounts do you have? (DFS include mobile money and bank accounts) SELECT ALL APPLICABLE (MULTIPLE): 1=momo: Orange Money, 2=momo: Africell, 3=momo: Airtel Money, 4=momo: Qcell, 5=bank, 6=other, specify

Qx. [Momo account duration] How long have you had a mobile money account? (consider the first momo account) … *years/months*

Qx. [DFS transactions, local agent (90 days)] In the past 90 days, about how many times have you attempted a DFS transaction at agent outlets located in your community? … *number (estimate)*

Qx. [DFS transaction value, local agent (90 days)] In the past 90 days / past 3 months, what is the common DFS transaction value you transacted at agent outlets located in your community? … *SLL*

Qx. [Total DFS value (30 days)] In the past 30 days / past 1 month, what is the total value of all your DFS transactions conducted (retail agent + peer to peer + bank agent)? … *estimated value, SLL*

B. VSLA Business and Digitization: Spillover Effects (Table 7)

Qx. [Business income (7 days)] What was the total business income earned during the last 7 days / last week after paying all expenses (profits)? *(Note: think about total profits, i.e. MOMO/bank commissions plus non-MOMO profits, minus wage expenses to the worker(s).)* … *SLL*

Qx. [Current VSLA members] How many members does your VSLA currently have? … *number*

Qx. [Contribution amount (90 days)] Total value of VSLA member contribution in the past 90 days / 3 months … *SLL*

Qx. [% digital contribution] What percent of VSLA total contribution in the past 90 days / 3 months came in as cash vs. digital/non-cash? … *% cash vs. % digital*

Qx. [Members defaulted (90 days)] How many members defaulted and/or delayed in contributing in the past 90 days / 3 months? … *number (0 if none)*

C. Consumer Trust and Provider Practices (Table 2)

Qx. [Trust: local VSLA] How much do you trust each of the following, or haven't you heard enough to say? Local VSLA … *1=not at all, 2=very little, 3=somewhat, 4=a lot, 5=haven't heard enough to say*

Qx. [Trust: momo agent] How much do you trust each of the following, or haven't you heard enough to say? Other momo agent … *1=not at all … 5=haven't heard enough to say*

Qx. [Trust: banking agent] How much do you trust each of the following, or haven't you heard enough to say? Other banking agent … *1=not at all … 5=haven't heard enough to say*

Qx. [Third person present] During a transaction, apart from the consumer and the agent, was there any third person (including another consumer in line) present who might have overheard the transaction details? … *1=yes, 2=no (agent-reported)*

Qx. [PIN requested] Have momo agents ever asked you to share your personal PIN during a transaction? … *1=yes, 2=no*

D. Household Shocks and Welfare (Table 8; Poverty and Shocks, Section 5.A)

Qx. [Negative shock experienced (90 days)] In the past 90 days / past 3 months, have you experienced ANY negative income/wealth shock (e.g., lost work/wages, illness/death, high prices, bad weather, etc.)? … *1=yes, 2=no*

Qx. [Unremedied shock (90 days)] In the past 90 days / past 3 months, have you experienced ANY negative income/wealth shock that you could NOT financially remedy? … *1=yes, 2=no*

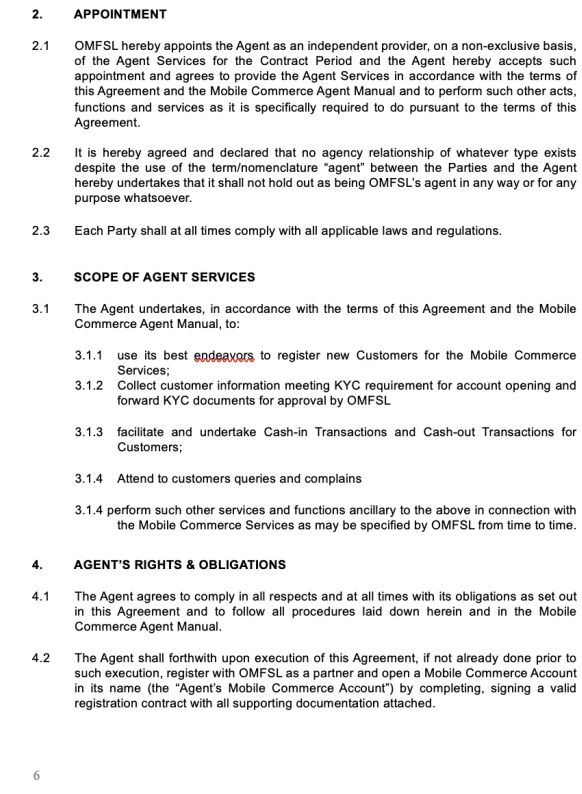
Qx. [Response to strongest shock] Consider the strongest negative shock, how did you respond to it? … *1=reduce expenditure, 2=labor supply response, 3=schooling response, 4=financing response (family/friends), 5=financing response (VSLA), 6=financing response (other financial institutions)*

Qx. [Total household expenses (30 days)] What is the total household expenses (i.e., food, bills, education, health, durables/appliances/accessories, personal care) made over the last 30 days / last month by your household? … *SLL*

H.4. Sample Agent Agreement Forms

Excerpts below are reproduced from the standard Agent Agreement that onboarded VSLAs sign with Orange Mobile Finance (SL) Limited (OMFSL), the Orange Money provider used in the randomized market-creation intervention. Panel (a) shows the scope of services an onboarded agent undertakes to provide; panel (b) shows the agreement's fees-and-commissions terms, which prohibit the agent from charging customers beyond the published tariff and bar any extra cash-in fee. The full agreement, including all clauses, the parties and signature pages, and supporting annexes, is available from the authors upon request.

Figure H.3: Sample Agent Agreement: Scope of Services, and Fees and Commissions



2. APPOINTMENT

2.1 OMFSL hereby appoints the Agent as an independent provider, on a non-exclusive basis, of the Agent Services for the Contract Period and the Agent hereby accepts such appointment and agrees to provide the Agent Services in accordance with the terms of this Agreement and the Mobile Commerce Agent Manual and to perform such other acts, functions and services as it is specifically required to do pursuant to the terms of this Agreement.

2.2 It is hereby agreed and declared that no agency relationship of whatever type exists despite the use of the term/nomenclature "agent" between the Parties and the Agent hereby undertakes that it shall not hold out as being OMFSL's agent in any way or for any purpose whatsoever.

2.3 Each Party shall at all times comply with all applicable laws and regulations.

3. SCOPE OF AGENT SERVICES

3.1 The Agent undertakes, in accordance with the terms of this Agreement and the Mobile Commerce Agent Manual, to:

3.1.1 use its best endeavors to register new Customers for the Mobile Commerce Services;
3.1.2 Collect customer information meeting KYC requirement for account opening and forward KYC documents for approval by OMFSL

3.1.3 facilitate and undertake Cash-in Transactions and Cash-out Transactions for Customers;

3.1.4 Attend to customers queries and complains

3.1.4 perform such other services and functions ancillary to the above in connection with the Mobile Commerce Services as may be specified by OMFSL from time to time.

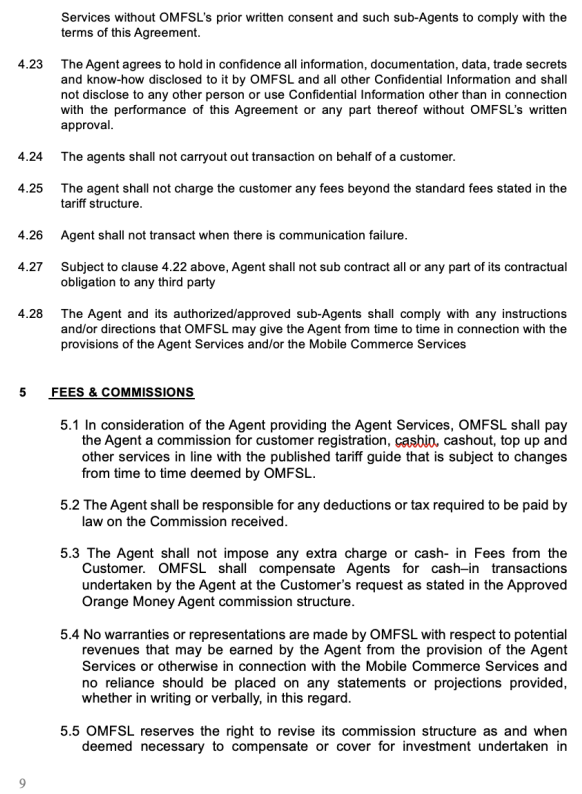
4. AGENT'S RIGHTS & OBLIGATIONS

4.1 The Agent agrees to comply in all respects and at all times with its obligations as set out in this Agreement and to follow all procedures laid down herein and in the Mobile Commerce Agent Manual.

4.2 The Agent shall forthwith upon execution of this Agreement, if not already done prior to such execution, register with OMFSL as a partner and open a Mobile Commerce Account in its name (the "Agent's Mobile Commerce Account") by completing, signing a valid registration contract with all supporting documentation attached.

6

(a) Scope of Agent Services: the agent undertakes to register new customers, collect KYC documentation, and facilitate Cash-in and Cash-out transactions for customers.



Services without OMFSL's prior written consent and such sub-Agents to comply with the terms of this Agreement.

4.23 The Agent agrees to hold in confidence all information, documentation, data, trade secrets and know-how disclosed to it by OMFSL and all other Confidential Information and shall not disclose to any other person or use Confidential Information other than in connection with the performance of this Agreement or any part thereof without OMFSL's written approval.

4.24 The agents shall not carryout out transaction on behalf of a customer.

4.25 The agent shall not charge the customer any fees beyond the standard fees stated in the tariff structure.

4.26 Agent shall not transact when there is communication failure.

4.27 Subject to clause 4.22 above, Agent shall not sub contract all or any part of its contractual obligation to any third party

4.28 The Agent and its authorized/approved sub-Agents shall comply with any instructions and/or directions that OMFSL may give the Agent from time to time in connection with the provisions of the Agent Services and/or the Mobile Commerce Services

5 FEES & COMMISSIONS

5.1 In consideration of the Agent providing the Agent Services, OMFSL shall pay the Agent a commission for customer registration, cashin, cashout, top up and other services in line with the published tariff guide that is subject to changes from time to time deemed by OMFSL.

5.2 The Agent shall be responsible for any deductions or tax required to be paid by law on the Commission received.

5.3 The Agent shall not impose any extra charge or cash- in Fees from the Customer. OMFSL shall compensate Agents for cash–in transactions undertaken by the Agent at the Customer's request as stated in the Approved Orange Money Agent commission structure.

5.4 No warranties or representations are made by OMFSL with respect to potential revenues that may be earned by the Agent from the provision of the Agent Services or otherwise in connection with the Mobile Commerce Services and no reliance should be placed on any statements or projections provided, whether in writing or verbally, in this regard.

5.5 OMFSL reserves the right to revise its commission structure as and when deemed necessary to compensate or cover for investment undertaken in

9

(b) Fees and Commissions: the agent is compensated via commission on the published tariff guide and is explicitly barred from imposing any extra charge or cash-in fee on customers.

H.5. Cost-Effectiveness: Back-of-the-Envelope Calculations

This subsection details the back-of-the-envelope cost-effectiveness calculations summarized in Section 7.E. Program costs come from the IPA–Sierra Leone subaward budget to CAFOD: the Community Engagement Meetings line item funding onboarding/enrollment ($16,087), plus, for Upgrade+Feedback only, the monthly consumer-feedback calls (10 enumerators per round at $400, 3 rounds, $12,000) and the respondent incentive paid for each completed call (10 SLL in airtime credit per respondent, valued here as the opportunity cost of participating in a brief, under-10-minute phone call, rather than as a cash transfer with its own welfare value). The respondent count is the number of completed Customer Surveys by round (Table A.1): 540, 540, and 528 for Cohort 1 (the published 1,296 for Cohort 1's Round 1 includes control- and Upgrade-arm respondents surveyed before the feedback mechanism existed; 540 is the Upgrade+Feedback-arm subset that actually received this incentive), 477, 452, and 444 for Cohort 2, and 366, 438, and 407 for Cohort 3 – 4,192 calls in total, or 41,920 SLL (≈$1,999 at the Table C.3 exchange rate), bringing the Upgrade+Feedback total to $30,086. This is a deliberately narrow accounting of costs (excludes monitoring, equipment, office, and indirect costs) and of benefits (three outcomes only, no externalities), in the same conservative spirit as Annan (2025).

Table H.1: Program Cost per Respondent

	Upgrade	Upgrade+Feedback
Total program cost (one-time)	$16,087	$30,086
per agent (N=189)	$85.12	$159.19
per VSLA (N=181)	$88.88	$166.22
per household (N=2,885)	$5.58	$10.43

NOTE: The three per-respondent rows divide the same one-time total cost by three different respondent-population sizes (Table A.1); they are not additive. Upgrade+Feedback's total adds the recurring feedback-call cost and the respondent airtime incentive (see text above for the respondent-count derivation) to the Upgrade cost.

Table H.2: Cost-Effectiveness Ratios by Respondent Type

Respondent (Outcome)	Effect (β_1)	Effect (β_2)	Ratio, Upgrade	Ratio, Upgrade+ Feedback
Agent	–	+21.1 pp** (8.1)	–	\$7.54 per pp
VSLA/Secretary	+\$127.96/yr (155.7)	+\$396.03/yr* (234.4)	144.0%	238.3%
Household	+\$7.37/yr (4.62)	+\$8.87/yr* (4.61)	132.1%	85.1%
Aggregate (VSLA income + household usage value only; see note)				
Total annualized \$ benefit	\$44,417	\$97,284		
Aggregate return			**276.1%**	**323.4%**

NOTE: Agent outcome is Table C.3 Col. 5; VSLA/Secretary is Table 7 Col. 1; Household is Table 5 Col. 7 (non-cash sent+received avg). Ratio is cost per respondent (Table H.1) divided by the effect: a dollar cost per pp for the Agent row, an annualized percent return for the other two. Standard errors in parentheses; stars denote each effect's own regression p-value (Romano–Wolf-corrected p-values appear in the main tables cited above). Agents have no control arm (Table C.3's note), so the Agent row is excluded from the aggregate return below, which sums only the VSLA-income and household-usage rows over total program cost. VSLA income and household usage are annualized ($\times 52$, $\times 4$) from the weekly/90-day effects in Tables 7 and 5; dollar figures use US\$1=SLL20.969 (May 2025). ***1%, **5%, *10% significance.