No. 2026-011

Scaling Support Programs:

Recurrent Demand, Capacity, and Liquidity

September 7, 2026

Takanori Suzuki

Scaling Support Programs:

Recurrent Demand, Capacity, and Liquidity

Takanori Suzuki[1]

September 7, 2026

Research Institute of Business Administration (Sanken), Waseda University
Working Paper Series, No. 2026-011

Working paper note. This paper is the sixth in a research program on mental quieting, self-referential psychological memory, entry into attention, experience-dependent learning, and assisted transition. The immediate foundations are RIBA Working Paper No. 2026-009, *When Openings Change Future Openings: A Fast–Slow Model of Experience-Dependent Transition Laws in Self-Referential Cognition*, and No. 2026-010, *The Economics of Assisted Transition: Learning Dividends, Re-closure Risk, and Optimal Withdrawal*. The present paper studies population-level recurrent demand, service capacity, service-credit obligations, liquidity, and provider governance. This working-paper version is based on Core Version 0.2.4 and consolidates the main text and technical supplement. A corresponding manuscript is being prepared for submission to *Service Science*. The two versions share the core model, analytical results, numerical inputs, and illustrations. The numerical settings are synthetic and do not establish empirical effectiveness, participant safety, or readiness for implementation.

Abstract

An individually efficient support policy need not be feasible when delivered to a heterogeneous population. We study support programs whose objective includes reducing future dependence on active assistance, and we ask when a population of such transitions can be delivered within staff capacity, service-credit obligations, and cash reserves. The model takes a policy-induced Markov-reward spell as its input, separates recurrence within a support spell from re-enrollment after withdrawal, and carries calendar-age profiles through population aggregation into physical, credit, and cash accounts. Four results organize the analysis. First, lifetime requirements are an expectation of a nonlinear type-specific load; the error of a two-mean representative calculation decomposes into a convexity term and an account-specific load–return covariance whose sign can be negative. Second, the minimum initial reserve that avoids an interim shortfall equals the maximum liability-adjusted cumulative deficit along the delivery path, so rollout schedules with identical lifetime totals differ in capacity, exposure, and liquidity requirements. Third, when waiting deters re-enrollment, the feedback path is unique and computable by a forward recursion, and the fixed-design offered-load path bounds it from above for every nonincreasing deterrence law: the operational screens computed without feedback are conservative, and a companion lower path supplies a necessity check for a specified law. Fourth, a stopping-only provider continuation incentive raises the withdrawal threshold, shifts the binding constraint, and at sufficient strength produces a closed active class for which finite-spell accounting is unavailable. Mechanism-based numerical comparisons

[1]Faculty of Commerce, Waseda University. Email: suzuki@waseda.jp. ORCID: 0009-0008-8955-7003.

with explicitly synthetic institutional conversions illustrate each result, and one-at-a-time rescalings of each conversion-coefficient group show which conclusions are preserved under the stated kernels and funding calendar (the ordering of schedules and the direction of the governance effect) and which depend on the assumed bounds (pass–fail outcomes). Operational self-liquidation—safely reducing future support dependence—is distinct from financial self-sufficiency.

Keywords: Service systems; recurrent demand; service credits; liquidity; provider incentives.

1 Introduction

A support policy can be efficient for an individual and nevertheless be infeasible for the institution that delivers it to a population. An additional period of assistance may improve a participant's subsequent transition prospects. Across participants, however, the resulting service time, repeated enrollment, future service entitlements, and cash payments must be delivered on a shared calendar. A favorable individual benefit–cost comparison leaves open whether sufficient staff, platform capacity, or liquid funds will be available when required. We study this gap for support programs whose intended outcome includes reducing future dependence on active assistance.

For concreteness, consider a publicly funded program that helps individuals through a transition, such as a return to employment, a move to independent living, or the completion of a training pathway. A participant enrolls, receives assistance from caseworkers and from a monitoring platform, and holds restricted service credits that can be redeemed with contracted providers. Support tapers and is withdrawn once the participant's capacity to proceed without it is judged sufficient; some participants later re-enroll. The program's operating plan must specify how many participants to admit and when, how many staff hours and platform slots to hold, how many credits may remain outstanding, and how large a cash reserve to keep between the receipt of external funding and the later arrival of any savings captured by the program. Each of these quantities is a commitment on the same calendar, and each is generated by the same participant histories.

Three mechanisms make the timing and composition of these commitments material. Earlier expenditure and later benefits are misaligned, and benefits accruing to participants or society need not return as cash to the program. Repeated enrollment concentrates demand among types with high service loads. Credits issued during support create obligations that outlive the activity that generated them, and a provider rewarded for maintaining active relations may delay withdrawal, which changes the process that produces future work rather than adding a fixed administrative charge.

Service operations research has established that service policy shapes subsequent demand. Chan, Yom-Tov, et al. (2014) show that speeding service can increase returns and exacerbate congestion; Chan, Farias, et al. (2012) and Shi et al. (2021) optimize discharge decisions against readmission risk and congestion; KC and Terwiesch (2012) document empirically that early discharge raises readmission. We build on the withdrawal–return link identified in that work and add the two commitment accounts that a support program carries alongside its capacity: service-credit obligations, whose financial relevance is established for loyalty points by Chun et al. (2020), and a cash reserve that must remain sufficient at every date of the delivery path.

Our input is an individual policy-induced Markov-reward process. The policy can be a first-best benchmark or a separately specified implementation. One active-support spell begins at enrollment and ends at withdrawal; reversals before withdrawal are inside that spell, and only re-enrollment after withdrawal starts another spell. We retain the initial distribution, the within-spell transition matrix, the exit distribution, and dated account profiles instead of reducing the participant to a few totals. Post-withdrawal renewal and population selection then generate recurrent demand, and explicit conversion maps produce physical work, credit issuance, operating cash, and captured cash saving as distinct outputs.

Four results provide the paper's substantive content. First, lifetime burden is an expectation of a nonlinear, type-specific load rather than a ratio of two means; its difference from a representative-person calculation contains a convexity term and a load–return covariance term, and the latter can reverse the sign of the error and differs across accounts (Proposition 1). Second, the minimum initial reserve that avoids an interim shortfall is the maximum liability-adjusted cumulative deficit along the delivery path (Lemma 1); consequently, rollout schedules with identical participation and identical lifetime totals differ in peak capacity, peak credit exposure, and required reserve. In the numerical example, a staged schedule is liquid with an initial reserve of 6,000 synthetic currency units but violates capacity and credit-exposure bounds, whereas a uniform schedule satisfies all three operational conditions. Third, when waiting deters re-enrollment, the feedback path is unique and is computed by a forward recursion, and the offered-load path computed for a fixed design bounds it from above for every nonincreasing deterrence law, so the screens computed without feedback are conservative; a companion lower path supplies a necessity check for a specified law (Proposition 2), and a reactive multiplier cannot avert the first congestion, so a zero-backlog capacity outcome is invariant to it (Corollary 1). Fourth, a stopping-only continuation incentive raises the implemented withdrawal threshold, moves the binding constraint from credit exposure to capacity, and, on the finite state grid of the example, produces a reachable closed active class for which a finite expected spell cost is unavailable (Proposition 3).

The mathematical components—Markov renewal, convolution, nonnegative workload reflection, survival accounting, a cumulative-deficit identity, and a causal forward construction with an order argument for an antitone map—are standard. The contribution is the service-design comparison they support when one implemented process and one set of recognition conventions are carried through every account. The institution is screened against jointly specified constraints; a social objective can then be optimized over the surviving designs, and no favorable social score offsets an unmet service or cash commitment.

The numerical study uses the individual transition mechanism of Suzuki (2026), summarized in Section S1 of the online supplement, together with transparent synthetic population and institutional conversions. One-at-a-time rescalings of each conversion-coefficient group separate the conclusions preserved under the stated kernels and funding calendar from those that depend on the assumed bounds. Sections 2–3 locate the question and define the system; Sections 4–6 establish the four results; Section 7 evaluates them numerically; Section 8 states managerial implications, an application procedure, and the scope of the analysis. The online supplement supplies the supporting arguments and the reproducibility details.

2 Related Work

The closest operating comparison is a service system with returns. Chan, Yom-Tov, et al. (2014) study a state-dependent queueing network in which service times and return probabilities depend on congestion, and show why current speed need not reduce long-run congestion. Chan, Farias, et al. (2012) formulate intensive-care discharge as a control problem in which early discharge raises readmission, and KC and Terwiesch (2012) estimate that relation from patient-flow data. Shi et al. (2021) develop a data-integrated decision framework for how many and which patients to discharge, balancing readmission risk and congestion. Those papers endogenize the link between service decisions and repeated demand through a channel in which early discharge under high occupancy raises subsequent returns; they are the closest operating comparators, and they are not evidence for the opposite-signed deterrence channel, in which waiting suppresses re-enrollment, that Section 5 analyzes. We complement them along two dimensions: the participant's return is generated by a policy-induced spell process with an explicit exit distribution, and capacity is evaluated jointly with credit obligations and liquidity. Section 5 characterizes when the fixed-design offered-load calculation used here brackets the equilibrium path of a system with waiting feedback.

The capacity calculations use the standard distinction between flow and occupancy associated with Little (1961), and the service-operations emphasis on staffing and time-varying demand in Gans et al. (2003). Uncertain arrival rates and supply availability motivate caution about nominal utilization (Whitt, 2006). Markov renewal (Pyke, 1961) supplies the representation for repeated spells with state-dependent starting conditions and delays. Keeping within-spell transitions distinct from between-spell returns prevents a specific double count when the individual input already includes reversals during support.

The financial side has two nearby literatures. Chun et al. (2020) treat loyalty points as promises of future service and study their value jointly with financial performance. Our credits have a fixed, explicitly specified service-settlement convention in the baseline, and an age ledger distinguishes outstanding units, gross settlement exposure, and actual cash redemption. Cash-balance management has a long history, including Miller and Orr (1966). Our minimum-reserve calculation is a pathwise cumulative-deficit identity whose role is to show why aggregated program totals or a positive net present value do not determine the liquidity needed by a particular rollout; stochastic cash management remains a complementary layer.

Heterogeneous policy assignment makes the distinction between the target population and the selected population essential (Manski, 2004; Kitagawa and Tetenov, 2018), and the warning against representative-agent aggregation is established (Kirman, 1992). Here the assigned unit carries a lifetime process: one-spell load, return likelihood, exit state, and timing jointly determine institutional burden. The covariance decomposition below identifies an account-specific aggregation error; estimation of the participant law is a separate empirical task.

Finally, multitask incentives and imperfect performance measures can redirect effort away from the principal's objective (Holmstrom and Milgrom, 1991; Baker, 1992). Performance measurement and mission accountability are consequential in public and hybrid organizations (Speklé and Verbeeten, 2014; Ebrahim et al., 2014). We isolate a narrower channel: a continuation-value wedge changes stopping and thereby changes the demand-generating process, and its downstream implications are computed.

Participation constraints, hidden action, and the cash price of a corrective contract belong to a fuller contracting model. This positions operational self-liquidation as a service-design objective rather than as an argument against outcome-based contracting or measurement.

3 From Individual Support to Population Commitments

3.1 Design, selection, and behavioral admissibility

An institution chooses a design γ: eligibility and access rules, support implementation, credit terms, funding arrangements, and a rollout calendar. Its first-best individual policy is a benchmark; the implemented policy is a separately recorded object. A case $z = (\theta, \omega, \alpha)$ specifies transition characteristics, mission weights, and an initial state law. A change in the provider contract belongs to γ, whereas the participant type z is held fixed.

Let F_0 be the target-population distribution and $w_\gamma(z) \in [0, 1]$ the eligibility times conditional uptake/access weight. For $q_\gamma = \int w_\gamma \, dF_0 > 0$, the selected law is

$$F_{P,\gamma}(dz) = \frac{w_\gamma(z)}{q_\gamma} F_0(dz), \qquad \int g w_\gamma \, dF_0 = q_\gamma \mathbb{E}_{F_{P,\gamma}}[g]. \tag{1}$$

Participation scale and composition are therefore distinct. Excluding a type changes the population measure while leaving that person's underlying requirements unchanged. If $q_\gamma = 0$, no participant arrives and the conditional law is undefined.

Behavioral admissibility is a separate restriction on the implementation. For a scalar rule x, ordered effectiveness and pressure margins define an interval $[L_{z,\gamma}, H_{z,\gamma}]$ of admissible rules, possibly empty. A common rule covers every selected type exactly when nonempty intervals occur almost surely and $\operatorname{ess\,sup} L_{Z,\gamma} \le \operatorname{ess\,inf} H_{Z,\gamma}$; cohort menus improve coverage but cannot cover an empty interval without changing the design. We carry a required coverage level $\underline{\Gamma}$ and a design coverage Γ_γ as a screen. These are model-defined conditions, and the main numerical experiment is conditional on them; Section 7 keeps the behavioral illustration separate from the financial–operational population.

3.2 A spell and two distinct recurrence operators

An active-support spell runs from first enrollment or re-enrollment to withdrawal. Let $Q_{z,\gamma}$ be the policy-induced substochastic matrix on a finite state set, α_z the initial row vector, and $A_{z,\gamma}$ the withdrawal-to-exit-state map. A withdrawal row of Q is zero and its active rewards are zero; withdrawal is an exit, not zero-intensity continuation. The ordinary finite-spell domain requires spectral radius $\rho(Q) < 1$, withdrawal probability $\alpha(I - Q)^{-1} A\mathbf{1} = 1$, and no unclassified exit. For a nonnegative state-account vector y^q, zero at withdrawal,

$$N = (I - Q)^{-1}, \qquad Y^q_{\text{spell}} = \alpha N y^q, \qquad y^q(n) = \alpha Q^n y^q. \tag{2}$$

The identity $\sum_n y^q(n) = Y^q_{\text{spell}}$ preserves both totals and episode-age profiles. Reversals and re-entry before withdrawal are already counted by Q.

A second operator describes return *after* withdrawal. With K possible starting classes, let $\mathsf{R}_{ij}(da)$ give the probability that a spell started in class i is followed by a

Table 1: Different accounts require different units and recognition events. The same population operator does not make the outputs interchangeable.

Account	Recognition	Required separation
Physical capacity	Service use by resource and date	Elapsed participant time is not staff time.
Service credits	Posted entitlement, then redemption, expiry, or cancellation	Support intensity is not an issued credit count.
Operating cash	Cash paid for specified inputs	Excludes redemption cash under the baseline convention.
Redemption cash	Validated service claim and its settlement date	Outstanding exposure is not itself a cash payment.
Captured saving	Cash actually returned to the fund	Social benefit is not automatically fund revenue.

spell in class j after start-to-start calendar lag a. It includes the current spell's duration and the post-withdrawal waiting time. Its total-mass matrix is $P_{ij} = \mathsf{R}_{ij}(\mathbb{R}_+)$. The gate for re-enrollment includes need, eligibility, uptake, access, and admission; autonomous need alone does not start a spell.

For nonexplosive regeneration, $\rho(P) < 1$, and a vector of one-spell calendar-age measures $\boldsymbol{\mu}^q$, the lifetime profile is

$$\boldsymbol{L}^q = \sum_{j=0}^{\infty} \mathsf{R}^{*j} * \boldsymbol{\mu}^q, \qquad \widetilde{Y}^q = \eta(I - P)^{-1} y^q_{\text{spell}}. \tag{3}$$

Here η selects the initial class and convolution acts on age and class. A positive lower bound on the start-to-start lag is sufficient for nonexplosion in the discrete numerical example. If every new spell regenerates the same distribution and returns with probability $r_z^{\text{post}} < 1$, (3) reduces to $Y^q_{z,\text{spell}}/(1 - r_z^{\text{post}})$; the renewal factor uses the post-withdrawal return probability, never a within-spell reversal probability. The supplement proves the profile-total identity and states the signed-account integrability qualification.

3.3 Calendar and recognition maps

An economic episode is not automatically a day or a worker-hour. Explicit duration and recognition maps transform episode profiles into physical service work, actual credit issuance, and cash events. Denote these maps by M_T, M_U, M_C, respectively. Social benefit M_B and cash saving actually captured by the fund M_S are separate maps. Individual discounting β applies to the participant's value problem; physical capacity and issued units are undiscounted, and any financial discount is attached separately to the calendar. Learning-value diagnostics and the individual's withdrawal value are decision objects, not cash assets.

Let $\Lambda_{0,\gamma}(ds, dz)$ be the mean measure of *first-spell* starts. For a nonnegative account q and a calendar set B,

$$\mathcal{Y}^q_\gamma(B) = \int\int \mathbf{1}\{s + a \in B\} L^q_{z,\gamma}(da)\Lambda_{0,\gamma}(ds, dz). \tag{4}$$

Measurability and local integrability of the displayed expression are assumed, and no Poisson arrival assumption is needed for this mean accounting. Equivalently, all-spell arrivals may be convolved with *one-spell* profiles (Route B); convolving all-spell arrivals with lifetime profiles would count returns twice. Initial backlogs or legacy credits enter through separate opening-state accounts; the primary example starts empty.

Equation (4) is a fixed-design offered-load calculation. If waiting changes uptake, return gating, the support policy, or cash recognition, the affected process and profiles must be regenerated jointly with the backlog they create. Section 5.5 shows that when waiting deters re-enrollment, the fixed-design path is an upper bound for every such equilibrium path, and Route B is the computational form in which the feedback is evaluated. An overload path is a diagnosis of commitments that cannot be met on schedule, not a claim that unavailable services were delivered.

4 Heterogeneity Changes Account Requirements

Fix the selected law $F_{P,\gamma}$ and consider the identical-spell baseline. For an account q, write $Y_q \geq 0$ for one-spell load, $R = r_Z^{\text{post}} \in [0, 1)$ for the post-withdrawal return probability, and $O_R = R/(1 - R)$ for return odds. Assume $\mathbb{E}[Y_q]$, $\mathbb{E}[O_R]$, and $\mathbb{E}[Y_q O_R]$ are finite. The population mean is $M_q = \mathbb{E}[Y_q/(1-R)]$, and the two-mean approximation is $M_q^{\text{mean}} = \mathbb{E}[Y_q]/(1 - \mathbb{E}[R])$.

Proposition 1 (Account-specific aggregation error). *Under these conditions,*

$$M_q = \mathbb{E}[Y_q]\{1 + \mathbb{E}[O_R]\} + \operatorname{Cov}(Y_q, O_R), \tag{5}$$

$$M_q - M_q^{\text{mean}} = \mathbb{E}[Y_q]\left\{\mathbb{E}[O_R] - \frac{\mathbb{E}[R]}{1 - \mathbb{E}[R]}\right\} + \operatorname{Cov}(Y_q, O_R). \tag{6}$$

The first term on the right of (6) *is nonnegative; the total error can take either sign.*

Proof. Use $1/(1 - R) = 1 + O_R$, expand the expectation of the product, and apply Jensen's inequality to the convex function $r/(1 - r)$ on $[0, 1)$. A sufficiently negative covariance dominates the nonnegative convexity term. $\square$

For example, mean one-spell load and mean return are identical when $Y = 1$ and either $R = 0.5$ surely or $R \in \{0, 0.8\}$ with probabilities 3/8 and 5/8, yet the lifetime means are 2 and 3.5. Pointwise subcriticality is also weaker than population integrability: $Y = 1$ and a uniform R on $(0, 1)$ yield $M = \infty$. A uniform bound $R \leq \bar{r} < 1$ with integrable Y suffices, and the direct integrability condition can be used instead.

An explicit co-loading benchmark makes the sign transparent. Let $X \sim N(\mu, \Sigma)$, $Y_q = Y_{q0} \exp(a_q' X)$, and $R = \operatorname{logit}^{-1}(\beta_0 + \beta' X)$. Gaussian exponential moments give

$$M_q = \mathbb{E}[Y_q]\{1 + \mathbb{E}[O_R] \exp(\kappa_q)\}, \qquad \kappa_q = a_q' \Sigma \beta. \tag{7}$$

Positive κ_q means that high-load types are also high-return types in this family; negative κ_q gives the opposite alignment. The loading vector a_q differs between staff time, credits, and cash, so a single multiplier is valid for all accounts only under additional restrictions. In the fixed-marginal synthetic comparison of Section 7, the exact means are 1.43, 1.58, and 1.80 at $\kappa_q = -0.4, 0, 0.4$, whereas the two-mean approximation is 1.51 throughout. The supplement derives (7) and records its numerical verification.

With class-changing returns, the general quantity is $\mathbb{E}[\eta_Z(I - P_Z)^{-1} y_{q,Z}]$, and expectation is taken after inversion. Selection adds a further distinction: for integrable lifetime load G, $\mathbb{E}_{F_{P,\gamma}}[G] = \mathbb{E}_{F_0}[G] + \mathrm{Cov}_{F_0}(w_\gamma, G)/q_\gamma$. Two designs with the same number of participants can therefore have different requirements because access selects different lifetime-load types. These are accounting consequences of the specified joint distribution; which groups actually participate is an empirical question.

5 Capacity, Service Obligations, and Liquidity

5.1 Capacity is a vector and a path

Let $H_k(t)$ be offered physical work per calendar unit for resource k, generated by (4), and $K_k(t)$ its service capacity. Staff, monitoring, and platform resources are separate unless a substitution or routing technology is explicitly supplied. With zero initial backlog and locally integrable rates, the work-conserving fluid backlog is

$$B_k(t) = \sup_{0 \le s \le t} \left\{ \int_s^t [H_k(u) - K_k(u)]\, du \right\}. \tag{8}$$

The choice $s = t$ includes zero. This is a workload, not a number of waiting participants or a waiting-time distribution. At constant participant inflow Λ_P from the remote past, $\overline{H}_k = \Lambda_P \mathbb{E}_{F_{P,\gamma}}[\widetilde{T}^k]$. Strict headroom requires $\overline{H}_k < K_k$ for every resource, and the associated stationary rate supremum is $\min_k K_k/\mathbb{E}[\widetilde{T}^k]$ for positive finite denominators. It is a mean-load screen; transient overload is read from (8). Spare platform capacity does not cancel a shortage of qualified staff.

5.2 Credits create a separate obligation path

The baseline instrument is a restricted, nontransferable service credit. Posting an entitlement is issuance; validated eligible service extinguishes units by redemption; expiration and valid cancellation extinguish units without creating cash receipts. Participants cannot exchange the credit for cash. "Liability" refers to the economic service obligation; classification under a financial-reporting standard is a separate legal determination.

Let $I(t)$ be actual issuance and $S_L(a)$ the survival probability that an issued unit remains outstanding at age a. In the fixed-terms, empty-start continuous benchmark,

$$L(t) = \int_0^t I(t-a) S_L(a)\, da, \quad m_L = \int_0^\infty S_L(a)\, da, \quad L(t) \le \overline{I} m_L \tag{9}$$

when $I(t) \le \overline{I}$ and $m_L < \infty$. The discrete analogue replaces the integral by a sum under a specified within-tick recognition order. Finite expiry bounds outstanding life even without redemption. With no extinguishment and sustained positive issuance, outstanding units grow without bound; a finite cohort instead leaves a finite stranded balance.

An age/vintage measure $\mathcal{L}_t(da, dv)$ is retained because equal totals can have different future redemptions or expiries. Gross exposure is $V(t) = \int p(t,a,v) \mathcal{L}_t(da, dv)$, whereas redemption cash is generated by actual settlements. A bound on L bounds V together with a bound on the settlement price p; neither bound establishes that cash is available at settlement. The capacity and cash scopes must also be reconciled: credit-settled services cannot be charged again as the same operating input, and any shared physical resources must enter the capacity map. The numerical example uses distinct cash blocks and leaves additional redemption-provider capacity to field specification.

5.3 Reserve requirements depend on recognition time

Let $F(t)$ denote external funding, $S(t)$ captured cash saving, $O(t)$ operating cash excluding credit settlement, and $C^{\text{red}}(t)$ redemption cash. Define net cash $f(t) = F(t) + S(t) - O(t) - C^{\text{red}}(t)$. With no borrowing or investment return in the baseline,

$$R(t) = R_0 + \int_0^t f(s)\,ds. \tag{10}$$

A reserve floor $\underline{R}(t) \geq 0$ is a specified requirement, for example $\kappa_R V(t)$, and it is a stock condition rather than a second cash payment. If redemption is included in an all-inclusive operating account, the separate settlement term is removed. Social benefit enters (10) only through a separately identified transfer that actually reaches the fund.

Lemma 1 (Reserve identity for a specified delivery path)**.** *For integrable cash flows on* $[0, H]$ *and a finite prescribed floor, define*

$$R_0^{\min}(H) = \max\left\{0, \sup_{0\leq t\leq H}\left[\underline{R}(t) + \int_0^t \{O + C^{\text{red}} - F - S\}(s)\,ds\right]\right\}. \tag{11}$$

Then $R(t) \geq \underline{R}(t)$ *for every* $t \leq H$ *if and only if* $R_0 \geq R_0^{\min}(H)$*.*

Proof. Substitute (10) into each floor inequality, rearrange, and take the supremum, also imposing $R_0 \geq 0$. □

The identity is elementary; its service-design content comes from regenerating every cash path from the implemented support process and settlement ledger. Equal lifetime cash totals need not give equal maxima in (11). Even fixing both total net cash and discounted net cash leaves the maximum interim deficit unidentified; a constructive example appears in the supplement. Long-run nonnegative cash drift and positive social value are likewise distinct from no-interim-shortfall feasibility. If borrowing, interest, or restricted funding is introduced, its timing and terms enter the ledger directly.

5.4 Feasibility is an intersection

For a fixed horizon and specified limits, define the admissible set by

$$\mathcal{G}^{\text{joint}}(H) = \mathcal{G}_B \cap \mathcal{G}_R \cap \mathcal{G}_C(H) \cap \mathcal{G}_L(H) \cap \mathcal{G}_F(H). \tag{12}$$

These predicates require, respectively, behavioral coverage $\Gamma_\gamma \geq \underline{\Gamma}$; proper spells and integrable renewal loads; $B_k(t) \leq \bar{B}_k$ for all k, t; well-defined credit obligations and any exposure cap $V(t) \leq \bar{V}$; and (11). Bounded issuance and finite expected outstanding life are additionally required when long-run credit boundedness is asserted. Equation (12) organizes the necessary simultaneous tests. A social objective ranks designs inside this set; it cannot turn a failed service commitment into a feasible one by placing sufficient weight on another benefit, and a fully funded design need not have positive social value. Where behavioral safety, redemption-provider supply, or a cash-capture map is unmeasured, a numerical pass is conditional on the supplied design assumptions.

5.5 Waiting feedback: the fixed-design path as a bracket

The offered-load calculation in (4) takes the re-enrollment gate as given. In a congested program, waiting itself can deter re-enrollment. The following result shows that the fixed-design calculation remains informative in exactly that case, and that the resulting feedback path is itself simple to compute.

Consider the discrete-tick form used in the numerical study. Let $\lambda_t \in \mathbb{R}_+^K$ be first-spell admissions by class at tick $t = 0, \dots, H$, $\mathsf{R}(a)$ the $K \times K$ return kernel at start-to-start lag $a \geq 1$, and $\mu_i^k(a) \geq 0$ the one-spell physical profile of resource k for a spell started in class i. A congestion multiplier $e : \mathbb{R}_+^m \to [0, 1]$, with $e(0) = 1$, is evaluated on the backlog vector b_{t-1} of m designated resources at the end of the previous tick. Given a backlog path $b = (b_t)$, spell starts and offered loads are

$$\Lambda_t(b) = \lambda_t + e(b_{t-1}) \sum_{s<t} \Lambda_s(b)\, \mathsf{R}(t-s), \qquad H_t^k(b) = \sum_{s \leq t} \Lambda_s(b)\, \mu^k(t-s), \tag{13}$$

with $b_{-1} = 0$, and the backlog map Φ returns the Lindley recursion $\Phi(b)_t^k = \max\{0, \Phi(b)_{t-1}^k + H_t^k(b) - K_t^k\}$. A feedback path is a fixed point $b^* = \Phi(b^*)$, and the fixed-design path of (4) is $\Phi(0)$, computed with the multiplier at its no-waiting value.

The following conditions are maintained. The horizon is finite; first-spell admissions, return kernels, one-spell account kernels, and capacity paths are finite and nonnegative; the return kernel has no mass at lag zero; initial backlogs are zero; and e is a single-valued function, nonincreasing in each coordinate, evaluated only at the preceding tick's backlog. The return kernel already contains the baseline, no-congestion return probabilities, so e scales those probabilities rather than implying that every participant returns when queues are empty; denied return mass does not reappear at a later lag. One-spell policies, recognition kernels, prices, capacity paths, and credit-extinguishment terms do not change with backlog.

Proposition 2 (Causal deterrence bounds)**.** *Under the maintained conditions:*

(i) Φ *is antitone:* $b \leq b'$ *componentwise implies* $\Lambda(b) \geq \Lambda(b')$, $H(b) \geq H(b')$, *and* $\Phi(b) \geq \Phi(b')$.

(ii) The feedback system has exactly one nonnegative path $b^* = \Phi(b^*)$ *on* $0, \dots, H$*; it is obtained by a forward recursion in* t*, and no continuity of* e *is required.*

(iii) Writing $b^U = \Phi(0)$, $b^L = \Phi(b^U)$, $\Lambda^U = \Lambda(0)$, *and* $\Lambda^L = \Lambda(b^U)$,

$$b^L \leq b^* \leq b^U, \qquad \Lambda^L \leq \Lambda(b^*) \leq \Lambda^U.$$

Every account generated by a fixed nonnegative linear map of spell starts—offered physical work, credit issuance, outstanding units, exposure, operating cash, and redemption cash under the maintained specifications—is ordered between its lower and upper paths. Running maxima and Lindley backlogs preserve these orders as monotone, generally nonlinear, maps. The upper bounds hold for every nonincreasing e *and therefore do not require the deterrence law to be known.*

(iv) If, in addition, the funding and captured-cash paths are the same under the three calculations and the reserve floors follow a common rule $\underline{R}_t^j = r_t^0 + \kappa_R V_t^j$ *with* $r_t^0 \geq 0$ *and* $\kappa_R \geq 0$ *fixed (or any common order-preserving rule), then* $R_{0,L}^{\min}(H) \leq R_{0,*}^{\min}(H) \leq R_{0,U}^{\min}(H)$.

(v) *For the whole-path iteration* $b^{[n+1]} = \Phi(b^{[n]})$ *from any initial path,* $b_t^{[n]} = b_t^*$ *whenever* $n \geq t+1$ *in exact arithmetic, so the finite-horizon path is recovered by at most* $H+1$ *sweeps; from* $b^{[0]} = 0$*, the even and odd iterates are successively nested lower and upper bounds.*

Proof sketch. Induction on t in (13) gives $\Lambda(b) \geq \Lambda(b')$ when $b \leq b'$, because the multiplier is nonincreasing and all kernels are nonnegative; nonnegative convolution orders offered work, and the Lindley recursion is monotone in its input, so Φ is antitone. Because $\Phi(b)_t$ depends on b only through $b_0, \ldots, b_{t-1}$, the path is constructed forward: $\Lambda_0^* = \lambda_0$ fixes b_0^*; given the path through $t-1$, the return candidates, the multiplier $e(b_{t-1}^*)$, Λ_t^*, H_t^*, and b_t^* are determined in turn, and any fixed point must coincide with this construction coordinate by coordinate. Antitonicity applied to $0 \leq b^*$ gives $b^* \leq b^U$, and applied again gives $b^* \geq \Phi(b^U) = b^L$. Fixed nonnegative linear account maps preserve the order; maxima and reflection preserve it by monotonicity. With common funding, captured cash, and an order-preserving floor, each cumulative deficit is ordered pointwise, and the supremum and positive part order the reserves. The iteration statement follows from the same causal structure. The supplement gives the full argument. □

Two features of the result matter for planning. First, because the upper bounds hold for every nonincreasing multiplier, a fixed-design pass certifies the capacity, exposure, and liquidity screens for the feedback path without any knowledge of how strongly waiting deters return; the fixed-design calculation is the right first computation whenever deterrence is the only feedback channel. Second, for a specified multiplier the feedback path can be computed in one causal forward pass, without an iterative fixed-point solver, and a failure of a screen on the lower path rules out that screen under that multiplier. The result concerns suppression of return demand with within-spell delivery and recognition held fixed. It is not a model of queue-induced delivery delay, and a reduction in re-enrollment is foregone return demand rather than evidence of improved participant outcomes; feedback that merely *defers* return, shifting the kernel to later lags without reducing its mass, and feedback through the support policy itself are outside its scope.

A reactive multiplier also has a limit that follows from the same causal structure.

Corollary 1 (Reactive deterrence cannot prevent the first congestion)**.** *Under the maintained conditions of Proposition 2, suppose the capacity screen requires zero backlog at every date for every resource entering the congestion multiplier. Let* $b^U = \Phi(0)$ *and* $\tau = \inf\{t \in \{0, \ldots, H\} : b_t^U \neq 0\}$*, with* $\inf \varnothing = +\infty$*. For every admissible multiplier with* $e(0) = 1$*, the feedback path has the same spell starts, offered loads, and backlogs as the fixed-design path through date* $\min\{\tau, H\}$*. If* $\tau \leq H$*, its first positive monitored backlog occurs at* τ*; if* $\tau > H$*, its entire path agrees with the fixed-design path. Hence the zero-backlog screen on the monitored resources passes under feedback if and only if it passes under the fixed design, and the same equivalence holds for a zero-backlog screen on all resources.*

Proof. Before the first fixed-design positive backlog, the monitored backlog vector is zero, so induction in time gives the same multiplier $e(0) = 1$, the same return candidates, and the same starts and loads as the fixed design. At the first overload date the multiplier still observes the preceding zero backlog, so the load and the positive backlog at that date coincide as well. If there is no such date, the induction covers the whole horizon. For a screen that also includes unmonitored resources, either a monitored resource becomes congested, which already fails the screen, or the whole start and load path is unchanged and every other resource has the same outcome. The same prefix argument applies to the lower path: its prescribed input b^U_{t-1} is zero up to the first monitored overload date, so it has the same first monitored overload and the same zero-backlog screen outcome. □

The corollary concerns a zero-backlog screen, zero initial backlogs, a purely reactive lagged multiplier normalized by $e(0) = 1$, and fixed first admissions, return kernels, one-spell profiles, and capacities. It does not apply to a positive backlog tolerance, to preventive admission control, to a multiplier with $e(0) < 1$, to forecasts of next-period load, or to redesigned capacity; subsequent backlog magnitudes can still be reduced substantially. Removing every return is the distinct counterfactual $e \equiv 0$ and violates $e(0) = 1$. Two consequences follow for the screens. A zero-backlog capacity screen is invariant across the fixed-design, lower, and feedback paths, so deterrence can mitigate congestion after it first appears but cannot avert it; and the lower path retains its role only for the exposure and liquidity screens.

6 Timing and Provider Retention as Design Levers

6.1 Changing rollout without changing participants

For a fixed case mixture and lifetime kernel $g_q(a)$, the population path is $Y_q = \lambda_0 * g_q$. Tonelli's theorem gives $\int Y_q = (\int \lambda_0)(\int g_q)$ for nonnegative integrable functions or measures. Reallocating the same admissions in time therefore preserves total work and issuance while generally changing peaks, overlaps with funding, and the minimum reserve. With bounded arrival density, $\|\lambda_0 * g_q\|_\infty \leq \|\lambda_0\|_\infty \|g_q\|_1$; the analogous bound applies to discrete periods.

Calendar separation is a mechanism for reducing overlapping commitments, and lifetime resource use is unchanged by it. If nonnegative cohort requirement profiles have disjoint supports, the peak of their sum equals the largest individual peak; for overlapping profiles it lies between that value and the sum of peaks. Slower rollout does not lower reserves in every case: delaying a cohort also delays its captured cash, and the funding calendar may not move with it. Delay, idle resources, and equitable access enter the choice among feasible schedules. We compare fixed schedules; optimizing their sequence is a separate problem.

6.2 A stopping-only continuation wedge

For a post-entry state s, let $D_z(s) = W_z(s) - Q^{C,*}_z(s)$ be the first-best advantage of withdrawal over active continuation. The continuation comparison is held at the benchmark for this reduced-form exercise. Let ϖ be a provider's continuation-value advantage and b a sunset credit in the same decision-value units. The implemented withdrawal test is

$$D_z(s) + b \geq \varpi, \qquad w = \varpi - b. \tag{14}$$

This is a specified stopping distortion; a fully dynamic provider principal–agent problem would determine ϖ and b endogenously. The active policy is unchanged until a benchmark withdrawal is replaced by continuation; a new implemented matrix $Q_{z,w}$ is then constructed.

Proposition 3 (Conditional stopping and load effects). *On a safe-capable region where $D_z(\ell)$ is continuous and strictly increasing, an interior boundary satisfies $\ell_w = D_z^{-1}(w)$. If $D_z \in C^1$ and $D_z' > 0$, then $d\ell_w/dw = 1/D_z'(\ell_w) > 0$. For $w \geq 0$, coupling a retention path to the benchmark up to benchmark withdrawal makes every nonnegative active-episode one-spell load weakly larger. Holding P fixed with $\rho(P) < 1$ preserves this weak ordering for lifetime totals.*

Proof. The threshold result is inversion, with the derivative following from the inverse-function rule. Under the coupling, cumulative loads are identical until benchmark stopping; the retention path then adds only nonnegative active loads. The fixed-renewal extension follows from $(I - P)^{-1} = \sum_{j\geq 0} P^j \geq 0$. $\square$

Setting $b = \varpi$ restores the benchmark *stopping criterion* within this restricted comparison. Intensity, reporting, access, and safe exit are separate margins, and b becomes a cash cost only through a contract-to-cash map whose payment enters the fund ledger exactly once. Longer support could also improve exit states and reduce subsequent return; the numerical exercise holds the return matrix fixed to isolate the stopping channel, and the ordering in Proposition 3 is conditional on that restriction.

Retention can also change the domain of the analysis. A reachable closed active class with positive continuing resource load has no finite expected withdrawal spell, and truncating its simulation would misreport a finite lifetime requirement. A hard duration or no-progress gate creates a different process and ensures operational finiteness; whether the resulting exit is safe or durable is a separate question. This distinction becomes material in the numerical example.

7 Mechanism-Based Numerical Comparisons

7.1 What is generated and what is assumed

The transition input comes from the individual assisted-transition model of Suzuki (2026); Section S1 of the supplement summarizes its structure. Its state is a support phase (pre-entry or re-closed versus post-entry) together with a 51-point grid for *operator legibility*, a slow state describing how structurally visible the recurrent operator becomes in later episodes and thereby altering transition conditions, rather than a generic capability stock (Supplement S1.1). The policy-induced matrix therefore has 102 states and includes reversals during active support. At the baseline low start, $\rho(Q) = 0.687$, eventual withdrawal probability is one, expected active episodes are 9.788, normalized resource use is 0.908, normalized support is 1.616, and mean exit legibility is 0.663. These are mechanism outputs in normalized units.

Three initial cases begin at legibility 0.06, 0.26, and 0.50 under the same transition mechanism. Synthetic target weights $(0.55, 0.30, 0.15)$, eligibility probabilities $(0.90, 0.70, 0.95)$, and conditional uptake/access probabilities $(0.50, 0.65, 0.90)$ produce $q_\gamma = 0.512$ and selected weights $(0.483, 0.266, 0.250)$. The between-spell return matrix is

$$P = \begin{pmatrix} .25 & .15 & .05 \\ .10 & .15 & .07 \\ .04 & .07 & .09 \end{pmatrix}, \qquad \rho(P) = 0.355, \tag{15}$$

with start-to-start waits of three to seven ticks after withdrawal (Supplement S7). Its entries are stress parameters. Calendarization assigns two synthetic ticks to a pre-entry

Table 2: The principal institutional settings are dimensioned synthetic stress inputs. Section 7.6 scales each group by 0.5 and 1.5.

Setting	Numerical assumption
Resource capacities per tick	Direct 0.52; monitoring 0.50; platform 4.75
Operating cash per resource unit	Direct 120; monitoring 80; platform 12
Credit terms	Redemption 0.10; cancellation 0.02 per tick; expiry at 30 ticks
Settlement and exposure	25 currency units per credit; exposure ceiling 1,200
Liquidity	Initial reserve 6,000; floor $0.4V(t)$; no borrowing
Cash capture	220 per selected participant, received 50 ticks after admission

or re-closed episode and one to a post-entry episode.

The four lifetime basis profiles are active elapsed mass h^A, normalized support h^U, spell starts h^S, and withdrawals h^W. Explicit bridges produce direct work $0.04h^A + 0.30h^U + 0.10h^S$, monitoring work $0.06h^A + 0.12h^S + 0.08h^W$, platform work $0.75h^A + 0.05h^U$, and issued credits $6h^U + 4h^S$. The selected per-participant totals are, respectively, 1.515, 1.447, 14.3, and 18.2 synthetic resource-hour or credit units.

Operating cash pays the modeled support inputs; redemption cash is a separate settlement block. New credits are posted at tick end and become redeemable in the next tick. The fund uses period-end netting in the baseline. All three schedules below share the same external funding calendar, totaling 26,100. The full coefficient, funding, and recognition specifications appear in the supplement and companion inputs.

7.2 Aggregation can change both the amount and the direction of error

The Gaussian comparison in Section 4 holds marginal load and return distributions fixed while changing their alignment. The representative value 1.51 overstates the opposed-loading population mean 1.43 and understates the aligned mean 1.80. A one-million-draw simulation in the source experiment reproduces the exact values within 0.0014; the exact formulas establish the sign comparison.

The three-initial-case process provides a second audit of recurrence. For the low start, the Markov-renewal normalized resource total is 1.456. Applying the within-spell re-closure diagnostic (0.450, hence a factor 1.82) as another geometric factor would report 2.649: the same recurrence counted twice. Neither number is cash, because the cash bridge is applied separately. A representative participant and a single untyped recurrence multiplier are therefore inadequate inputs for a service-system budget.

7.3 Equal participation and equal totals do not give equal feasibility

We compare 74 target-population opportunity units, giving expected selected participation 37.9 under the frozen selection law. The front-loaded design admits all units at tick zero. The staged design admits at rates 0.2 for 20 ticks, 0.5 for 20, 1.0 for 40, 0.7 for 20, and 0.3 for 20. The uniform design admits 74/120 per tick for 120 ticks. These are alternative fixed schedules.

Table 3: The same participants and lifetime amounts produce different operational outcomes. “Yes” requires all modeled capacity, exposure, and liquidity bounds; behavioral admissibility is a separate condition.

Schedule	Peak platform	Peak exposure	$R_0^{\min}$	All three?
Front-loaded	28.81	4,928	22,081	No
Staged	6.77	1,717	5,709	No
Uniform	4.51	1,171	4,251	Yes

All three generate total operating cash 17,793, redemption cash 14,085, issued credits 691, captured saving 8,339, and terminal cash surplus 2,562. An optional synthetic social valuation also preserves its total across schedules and is kept outside the cash identity. The supporting computations use a 1,400-tick reporting grid that includes the long tails of the frozen lifetime profiles, and the all-spell-start route reproduces the first-spell route to within 5×10^{-8} on every account (Supplement S7).

The front-loaded schedule exceeds all three operational bounds, although its fund NPV at the synthetic calendar discount 0.999 is positive, 641. The staged schedule passes the cash test because its minimum reserve is below 6,000, and it fails the capacity and exposure limits. The uniform schedule passes all three specified operational tests (Table 3). A liquidity-only approval would therefore admit a schedule whose services or credits cannot be supported within the other stated limits.

Figure 1 isolates the cash mechanism. Earlier admissions bring expenditure forward while the external funding calendar remains unchanged and captured saving follows admissions with a delay. Panel (b) magnifies the two schedules whose margins stay positive: the staged minimum of 291 at tick 106 is the visible counterpart of $R_0^{\min} = 5{,}709 < 6{,}000$, and the uniform minimum of 1,749 at tick 121 corresponds to $R_0^{\min} = 4{,}251$. A positive terminal balance cannot repair an earlier violation without additional financing. The plotted paths are planned commitment paths; where capacity is violated, realized delivery would be delayed, and that delay is outside the fixed-design calculation. Section 7.4 evaluates a narrower feedback in which backlog suppresses subsequent re-enrollment while within-spell delivery and recognition are held fixed.

Within-period recognition matters as well. If each tick’s outflows occur before that tick’s inflows, required reserves rise from 22,081 to 22,231, from 5,709 to 5,972, and from 4,251 to 4,470 for the three schedules. This documented alternative order preserves terminal cash while changing liquidity, which is why event timing is retained below the level of lifetime totals.

7.4 The fixed-design screens are conservative under deterrence feedback

Proposition 2 is evaluated with the gate $e(B) = \exp(-B/\hat{B})$ applied to re-enrollments, where B is the platform backlog at the end of the previous tick and $\hat{B} = 10$ platform resource-hour equivalents, about two ticks of capacity. Table 4 reports, for each schedule and for the weaker deterrence $\hat{B} = 40$, the fixed-design path $\Phi(0)$, the lower path generated with the multiplier evaluated on the fixed-design backlog, and the feedback path computed by the forward recursion of Proposition 2(ii); the whole-path sweeps $\Phi^n(0)$ reproduce it within 12 sweeps.

For the uniform schedule the platform backlog is zero, so feedback is inactive and

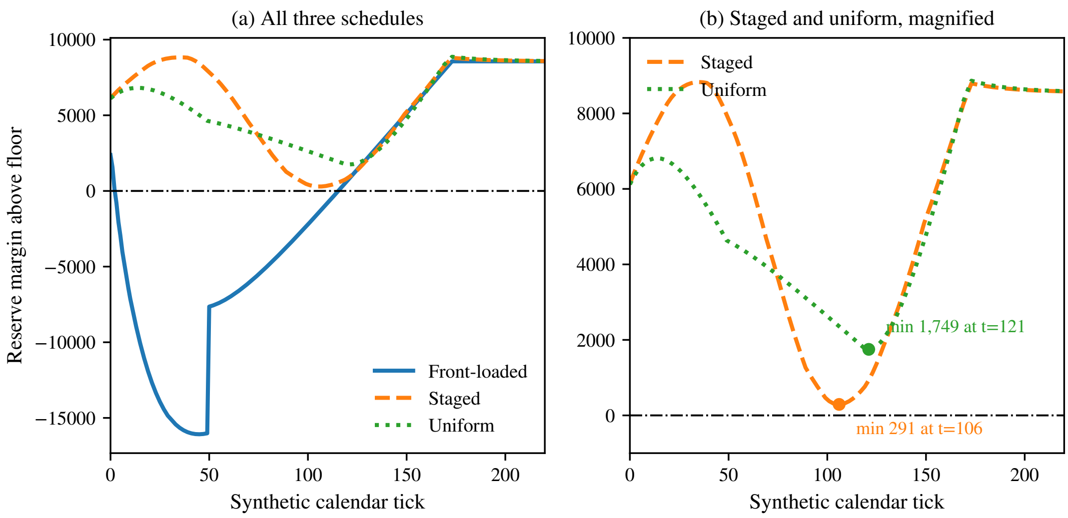

Figure 1: Reserve margin above the specified floor for equal-total rollout schedules with an initial reserve of 6,000 synthetic currency units. (a) All three schedules. (b) The staged and uniform schedules magnified, with their minima marked. A negative margin is an interim cash failure even when final net cash is positive.

Table 4: Deterrence feedback lowers every requirement relative to the fixed-design path and changes no screen outcome. Screens are capacity (C), exposure (E), and liquidity (L); P passes and F fails. Spell starts include first spells and re-enrollments.

$\hat{B}$	Schedule	Spell starts		Peak platform		Peak exposure		$R_0^{\min}$		Screens (C/E/L)		
		fixed	equil.	fixed	equil.	fixed	equil.	fixed	equil.	fixed	lower	equil.
10	Front-loaded	58.65	37.91	28.81	28.81	4,928	4,928	22,081	15,901	FFF	FFF	FFF
10	Staged	58.65	48.43	6.77	5.63	1,717	1,424	5,709	2,567	FFP	FFP	FFP
10	Uniform	58.65	58.65	4.51	4.51	1,171	1,171	4,251	4,251	PPP	PPP	PPP
40	Front-loaded	58.65	38.00	28.81	28.81	4,928	4,928	22,081	15,923	FFF	FFF	FFF
40	Staged	58.65	50.70	6.77	6.05	1,717	1,549	5,709	3,636	FFP	FFP	FFP
40	Uniform	58.65	58.65	4.51	4.51	1,171	1,171	4,251	4,251	PPP	PPP	PPP

the equilibrium equals the fixed design. For the staged schedule, waiting removes about half of the re-enrollments (58.7 to 48.4 spell starts, of which 37.9 are first spells), lowers peak platform load from 6.77 to 5.63, and lowers the minimum reserve from 5,709 to 2,567. The capacity and exposure failures persist at the feedback path, and Corollary 1 explains why the capacity outcome cannot change: the first platform overload occurs at tick 50 with the same offered load, 4.875 against capacity 4.75, as in the fixed design, because the multiplier still observes the preceding zero backlog. Even the first-spell-only reference path exceeds the exposure ceiling (peak 1,220 against 1,200); its platform peak of 4.71 against 4.75 is not attainable by any admissible multiplier, since eliminating every return is the counterfactual $e \equiv 0$ rather than a gate with $e(0) = 1$. Deterrence reduces subsequent congestion and the financial requirements; it does not turn a zero-backlog capacity failure into a pass. For the front-loaded schedule, the backlog suppresses almost all re-enrollments (37.9 spell starts, essentially the first spells alone) and the minimum reserve falls from 22,081 to 15,901, which remains far above 6,000. In every case the feedback path lies between the lower path and the fixed-design path, as Proposition 2 requires. The fixed-design pass certifies the uniform

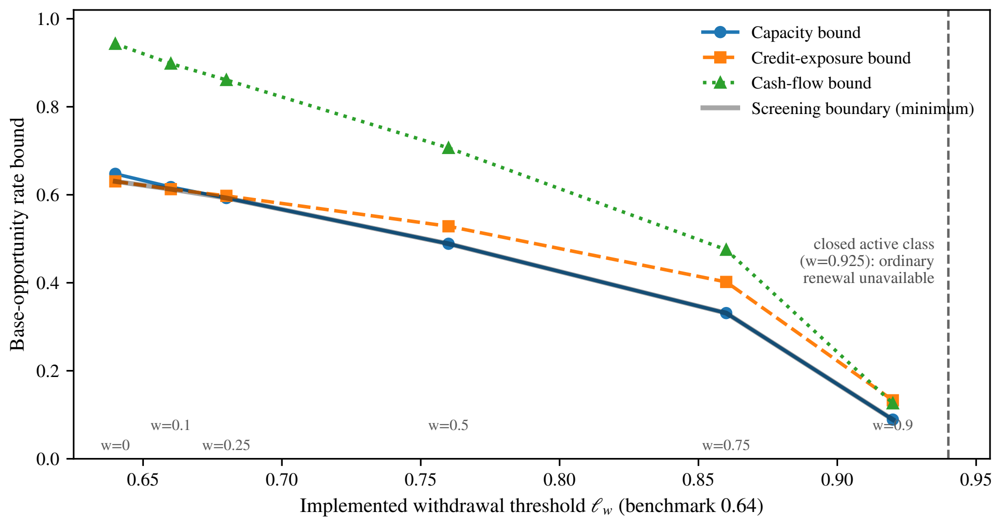

Figure 2: Stationary screening bounds against the implemented withdrawal threshold ℓ_w induced by the stopping-only retention wedge w. The binding account moves from credit exposure to capacity as the threshold rises. At $w = 0.925$ (threshold 0.94) the implemented process has a closed active class and ordinary renewal accounting is unavailable. Return probabilities, bridges, and captured saving are held fixed.

schedule's three screens for every nonincreasing multiplier; the capacity outcomes coincide across the three paths in every row by Corollary 1, and the lower-path failures rule out the staged and front-loaded exposure screens under the specified multipliers. The suppressed re-enrollments are foregone return demand under waiting, not evidence that the deterred participants fared better without support.

7.5 Continuation incentives shift the bottleneck and the model domain

The stopping comparison holds active actions and the return matrix (15) fixed and changes only the first crossing $D_z(\ell) \geq w$. For the stationary screen, external funding is fixed at 300 per tick and the exposure ceiling at 1,200. For each proper case we compute the capacity rate supremum, the credit-exposure rate limit, and the break-even cash-flow rate; their minimum is a stationary screening boundary.

Figure 2 reports the three bounds against the implemented withdrawal threshold ℓ_w, expressed in model-state units, with the corresponding wedge values marked; observing the threshold in the field requires a calibrated measurement map, which is not supplied here. At $w = 0$ the threshold is the benchmark 0.64, the boundary is 0.630 target-opportunity units per tick, and credit exposure binds. At $w = 0.25$ the threshold is 0.68, the boundary falls to 0.593, and capacity binds. At $w = 0.50$, 0.75, and 0.90 the thresholds are 0.76, 0.86, and 0.92 and the boundaries 0.489, 0.331, and 0.089. Retention changes which resource limits expansion, in addition to raising the operating bill.

At $w = 0.925$ the nominal crossing is 0.94, and the 102-state implemented matrix has a reachable closed active class of 94 states with no reachable withdrawal state from the low start. This graph property independently establishes $\rho(Q) = 1$, so finite-spell accounting is excluded rather than extrapolated as a small positive throughput. A revised exit gate would need its own process, safety assessment, and cash accounting.

Longer support may improve the exit state and reduce return need. The fixed-P comparison isolates the stopping channel; with endogenous returns the direction of the lifetime effect depends on the exit-state response, and a sunset credit that removes w carries an uncalibrated cash cost.

7.6 Sensitivity of the conclusions to the synthetic conversions

Each institutional coefficient group in Table 2, together with the physical-work and credit-issuance coefficient groups, was scaled by 0.5 and by 1.5, one group at a time with all others at baseline, and the three schedules were recomputed (Table 5); joint variation of several groups and the full parameter range were not explored. Two conclusions hold in all 23 tested cases: the minimum reserves and the peak loads are ordered front-loaded $\geq$ staged $\geq$ uniform, and the front-loaded schedule fails every screen. These orderings are preserved under every tested one-at-a-time rescaling with the stated kernels and funding calendar; they are not a universal ranking, since a different funding calendar can reverse the reserve ordering of two schedules (Section 6). The pass–fail pattern at the stated bounds is, by contrast, a property of those bounds: the uniform schedule passes all three screens in 15 of the 23 cases, and the staged schedule passes liquidity alone in 11 cases. The same distinction applies to governance: the screening boundary of Figure 2 is decreasing in w under every tested rescaling of capacities, ceiling, price, and funding, whereas the wedge at which capacity replaces credit exposure as the binding account depends on the relative levels of the capacity and exposure limits (Supplement Table S4).

7.7 Robustness and numerical scope

The online supplement reports exact tail checks, return and timing boundaries, the full stopping table, and a common-shock sensitivity. The latter uses a synthetic AR(1) common factor and two candidates fixed after an earlier exploration. In a separate 10,000-path run, scales 0.87 and 0.88 have success proportions 0.961 and 0.943, with simultaneous one-sided 95% binomial lower bounds 0.957 and 0.938; only the former meets a 0.95 lower-bound screen. A separate behavioral safe-window ensemble, summarized in Supplement S3, illustrates the coverage restriction imposed by common rules; its cases are distinct from the three-case financial population, so the operational passes above are conditional on the behavioral screen rather than a joint certification.

8 Implications, Application, and Scope

8.1 Managerial implications

The model suggests four rules for managing a support program at scale. First, recurrent work should be budgeted by selected population and by account: the lifetime requirement of an account is $\mathbb{E}[Y_q/(1-R)]$, and the cost of an average participant multiplied by an average return factor omits both the convexity of return loading and the load–return covariance, whose sign differs across staff time, credits, and cash. Selection changes both the level and the composition of requirements, so a low utilization count is evidence of durable independence only when access and coverage are held fixed.

Second, expansion should be evaluated on a common calendar across capacity, service commitments, and cash. The equal-total comparison identifies a concrete decision

Table 5: Sensitivity of the rollout comparison to the synthetic conversions. Each coefficient group is scaled by 0.5 and 1.5 with all others at baseline. F, S, and U denote the front-loaded, staged, and uniform schedules; screens are capacity/exposure/liquidity with P for pass and F for fail.

Coefficient group scaled	$R_0^{\min}$ (F / S / U)			Screens C/E/L (F / S / U)		
Baseline	22,081	5,709	4,251	FFF	FFP	PPP
Physical-work coefficients ×0.5	14,094	0	0	FFF	PFP	PPP
Physical-work coefficients ×1.5	30,283	13,156	12,117	FFF	FFF	FPF
Credit-issuance coefficients ×0.5	15,816	459	0	FFF	FPP	PPP
Credit-issuance coefficients ×1.5	28,508	11,473	10,385	FFF	FFF	PFF
Operating cash per resource unit ×0.5	14,094	0	0	FFF	FFP	PPP
Operating cash per resource unit ×1.5	30,283	13,156	12,117	FFF	FFF	PPF
Settlement price ×0.5	15,816	459	0	FFF	FPP	PPP
Settlement price ×1.5	28,508	11,473	10,385	FFF	FFF	PFF
Redemption and cancellation rates ×0.5	20,439	4,289	2,651	FFF	FFP	PFP
Redemption and cancellation rates ×1.5	22,407	6,044	4,588	FFF	FPF	PPP
Expiry age ×0.5	20,624	4,554	2,903	FFF	FFP	PPP
Expiry age ×1.5	22,243	5,837	4,417	FFF	FFP	PPP
Captured saving per participant ×0.5	22,081	7,727	6,807	FFF	FFF	PPF
Captured saving per participant ×1.5	22,081	4,180	1,798	FFF	FFP	PPP
Capture delay ×0.5	19,649	3,046	2,513	FFF	FFP	PPP
Capture delay ×1.5	22,081	8,330	5,988	FFF	FFF	PPP
Reserve-floor ratio ×0.5	21,915	5,454	4,033	FFF	FFP	PPP
Reserve-floor ratio ×1.5	22,255	5,970	4,471	FFF	FFP	PPP
Resource capacities ×0.5	22,081	5,709	4,251	FFF	FFP	FPP
Resource capacities ×1.5	22,081	5,709	4,251	FFF	PFP	PPP
Exposure ceiling ×0.5	22,081	5,709	4,251	FFF	FFP	PFP
Exposure ceiling ×1.5	22,081	5,709	4,251	FFF	FPP	PPP

error: a staged design can pass a reserve test and still fail delivery constraints. Reserve consumption, credit exposure, and provider-hour demand are distinct requirements with distinct recognition events, and the admissible designs are those in the intersection of the corresponding screens. The scope of the service being purchased, in particular whether credit-funded activities share a constrained provider pool, is part of that specification.

Third, when waiting deters re-enrollment, the fixed-design offered-load path is the right first computation: it bounds every requirement of the feedback path from above for any nonincreasing deterrence law, so a design that passes without feedback passes with it, and a zero-backlog capacity failure cannot be cured by reactive deterrence, which acts only after congestion has first appeared (Corollary 1). For a specified law the feedback path is one forward recursion, and the lower path obtained by evaluating the multiplier on the fixed-design backlog is a necessity check. Re-enrollment suppressed by waiting is foregone return demand and should be recorded separately from durable independence; a lower requirement produced by deterrence is not a better outcome.

Fourth, governance belongs upstream of budgeting. A retention incentive raises the withdrawal threshold and hence the quantities to be staffed and funded, and it can change which resource binds. A corrective sunset arrangement should be assessed by its effect on the implemented process and by its cash terms; a hard cap that makes a spell finite is operationally different from a safe, durable exit. These qualifications define *operational self-liquidation*: the intended reduction of future support dependence is distinct from cash self-financing and from maximizing withdrawals.

8.2 An application procedure

The results translate into a five-step screening procedure for a proposed design γ and rollout calendar. (1) Export the implemented spell process (α, Q, A) and its one-spell account profiles for each participant class, and verify $\rho(Q) < 1$ with withdrawal probability one; a design that fails this step is outside finite-spell accounting. (2) Specify the return kernel and the selection weights, and compute the lifetime profiles by (3), taking expectations after inversion. (3) Convolve first-spell admissions with the lifetime profiles, or all-spell starts with one-spell profiles, and generate the physical, credit-vintage, and cash paths under the stated recognition conventions. (4) Evaluate the capacity, exposure, and reserve screens on the fixed-design path; a design that passes is admissible under any nonincreasing deterrence law. Where a deterrence law is specified, compute the feedback path by the forward recursion of Proposition 2 and the lower path; a design that fails on the lower path is inadmissible under that law. (5) Rescale the conversion-coefficient groups and record which conclusions survive, distinguishing orderings preserved under the tested perturbations from bound-dependent outcomes, before ranking the admissible designs by a social objective.

8.3 Scope

Several boundaries define the contribution. The individual process is a model input, and the population, resource, credit, funding, and shock settings are synthetic; the numerical boundaries are conditional on the stated parameters, and the sensitivity analysis characterizes that dependence rather than removing it. Behavioral coverage is illustrated on a separate ensemble without a person-level crosswalk to the operational example. The feedback result covers deterrence of re-enrollment with within-spell delivery and recognition held fixed; queue-induced delivery delay, deferral of return, feedback through the support policy, and stochastic queueing are complementary layers. The sensitivity analysis is a set of one-at-a-time rescalings, not a characterization of the full parameter space. Provider governance is a specified stopping distortion, with participation constraints, hidden action, and the cash price of a corrective contract left to a fuller contracting model. Legal classification of service credits is a separate determination. We compare fixed rollout alternatives; optimizing admission sequences and certifying stochastic safety are further steps. Financial results require a cash-capture mechanism, because social value enters the fund only through an explicit transfer.

Within these boundaries, the analysis shows why individual efficiency is insufficient for population delivery and what replaces it. The relevant comparisons change when recurring loads are correlated, when the same commitments are rearranged in time, when waiting feeds back into re-enrollment, and when continuation incentives alter the process that produces the commitments. Familiar accounting and stochastic-process tools become useful together because they prevent an unrelated surplus from masking an unmet commitment. A support policy can make its own continuation less necessary without making its delivery automatically feasible; both questions must be answered, and the screens developed here answer the second.

References

Baker, G. P. (1992). "Incentive Contracts and Performance Measurement". *Journal of Political Economy* 100.3, pp. 598–614. DOI: 10.1086/261831.

Chan, C. W., V. F. Farias, N. Bambos, and G. J. Escobar (2012). "Optimizing Intensive Care Unit Discharge Decisions with Patient Readmissions". *Operations Research* 60.6, pp. 1323–1341. DOI: 10.1287/opre.1120.1105.

Chan, C. W., G. Yom-Tov, and G. Escobar (2014). "When to Use Speedup: An Examination of Service Systems with Returns". *Operations Research* 62.2, pp. 462–482. DOI: 10.1287/opre.2014.1258.

Chun, S. Y., D. A. Iancu, and N. Trichakis (2020). "Loyalty Program Liabilities and Point Values". *Manufacturing & Service Operations Management* 22.2. Issue year 2020; first published online October 22, 2019, pp. 257–272. DOI: 10.1287/msom.2018.0748.

Ebrahim, A., J. Battilana, and J. Mair (2014). "The Governance of Social Enterprises: Mission Drift and Accountability Challenges in Hybrid Organizations". *Research in Organizational Behavior* 34, pp. 81–100. DOI: 10.1016/j.riob.2014.09.001.

Gans, N., G. Koole, and A. Mandelbaum (2003). "Telephone Call Centers: Tutorial, Review, and Research Prospects". *Manufacturing & Service Operations Management* 5.2, pp. 79–141. DOI: 10.1287/msom.5.2.79.16071.

Holmstrom, B. and P. Milgrom (1991). "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design". *The Journal of Law, Economics, and Organization* 7.special issue, pp. 24–52. DOI: 10.1093/jleo/7.special_issue.24.

KC, D. S. and C. Terwiesch (2012). "An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit". *Manufacturing & Service Operations Management* 14.1, pp. 50–65. DOI: 10.1287/msom.1110.0341.

Kirman, A. P. (1992). "Whom or What Does the Representative Individual Represent?" *Journal of Economic Perspectives* 6.2, pp. 117–136. DOI: 10.1257/jep.6.2.117.

Kitagawa, T. and A. Tetenov (2018). "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice". *Econometrica* 86.2, pp. 591–616. DOI: 10.3982/ECTA13288.

Little, J. D. C. (1961). "A Proof for the Queuing Formula: $L = \lambda W$". *Operations Research* 9.3, pp. 383–387. DOI: 10.1287/opre.9.3.383.

Manski, C. F. (2004). "Statistical Treatment Rules for Heterogeneous Populations". *Econometrica* 72.4, pp. 1221–1246. DOI: 10.1111/j.1468-0262.2004.00530.x.

Miller, M. H. and D. Orr (1966). "A Model of the Demand for Money by Firms". *The Quarterly Journal of Economics* 80.3, pp. 413–435. DOI: 10.2307/1880728.

Pyke, R. (1961). "Markov Renewal Processes: Definitions and Preliminary Properties". *The Annals of Mathematical Statistics* 32.4, pp. 1231–1242. DOI: 10.1214/aoms/1177704863.

Shi, P., J. E. Helm, J. Deglise-Hawkinson, and J. Pan (2021). "Timing It Right: Balancing Inpatient Congestion vs. Readmission Risk at Discharge". *Operations Research* 69.6, pp. 1842–1865. DOI: 10.1287/opre.2020.2044.

Speklé, R. F. and F. H. M. Verbeeten (2014). "The Use of Performance Measurement Systems in the Public Sector: Effects on Performance". *Management Accounting Research* 25.2, pp. 131–146. DOI: 10.1016/j.mar.2013.07.004.

Suzuki, T. (2026). *The Economics of Assisted Transition: Learning Dividends, Re-closure Risk, and Optimal Withdrawal.* Working Paper 2026-010. September 4. Public working-paper version based on canonical Draft Version 0.4.1. Research Institute of Business Administration, Waseda University.

Whitt, W. (2006). "Staffing a Call Center with Uncertain Arrival Rate and Absenteeism". *Production and Operations Management* 15.1, pp. 88–102. DOI: 10.1111/j.1937-5956.2006.tb00005.x.

Technical Supplement

This supplement is the technical companion to the Core text in this working paper. The S-prefixed sections, equations, tables, and figures retain the identifiers of the separate journal supplement. References to the Core refer to the preceding main text.

S1 Process Inputs and Finite-Spell Accounting

This supplement is sufficient to check the analytical claims and reconstruct the numerical exhibits of the Core manuscript. The individual policy is an input, not re-optimized in the population model. Subscripts for case and design are suppressed only inside proofs. All empirical interpretations remain subject to the limitations stated in the Core.

S1.1 The individual input process in brief

The population model receives an initial law α, a policy-induced within-spell matrix Q, a withdrawal map A, and separate account vectors; net individual payoff need not be nonnegative. This subsection summarizes the numerical input mechanism of the upstream individual model (Suzuki 2026, cited in the Core); the population analysis uses the exported objects and does not re-optimize the individual policy.

State and within-episode variables. The beginning-of-episode economic state is (J, ℓ), with $J \in \{R, A\}$ and 51 equally spaced values of $\ell \in [0, 1]$. R is unresolved before qualified entry or re-closed after it; A is post-entry while the active-support decision remains open, and does not mean permanent independence. Operator legibility ℓ parameterizes how structurally visible the recurrent operator becomes in later episodes and thereby changes transition conditions. It is not a generic skill stock, an output capacity, or a clinically estimated score.

Within an episode, the fast state is (q, v, g, κ): curiosity, visibility, evaluative pressure, and object sensitivity. These variables are distinct from the slow state ℓ. To obtain a Markov model in (J, ℓ), the numerical baseline resets the fast state at each economic decision point, to $(0, 0, 0, 0)$ in R and to $(0.18, 0.25, 0.05, 0.12)$ in A; only (J, ℓ) is carried between economic decision points.

Episodes, experience, and capture. The economic action grids are $\{0, 0.025, \ldots, 0.25\}$ in R and $\{0, 0.025, \ldots, 0.15\}$ in A. The fixed (J, ℓ, u) determines a fast path over horizon $H_{\text{fast}} = 12$, integrated at step 0.10, and the cause-specific intensities of qualified entry, fusion, and re-closure. Fusion is an adverse within-episode event, not an absorbing economic failure state; a fusion-ending episode is mapped to R at the next decision point. Writing τ for within-episode time and $S(\tau)$ for its event state,

$$I_A = \int_0^{H_{\text{fast}}} \mathbf{1}_{\{S(\tau)=A\}}\, v(\tau)\kappa(\tau)\, d\tau, \qquad Y = 1 - e^{-I_A},$$

$$G = 1 - \exp\Big\{ -0.0018 \int_0^{H_{\text{fast}}} g(\tau)\, d\tau \Big\}.$$

Y is qualified experience and G is the episode-level evaluative-capture summary; G is not the instantaneous fast variable $g(\tau)$. The slow update is

$$\ell' = \ell + \eta Y(1 - \ell) - (\delta_\ell + \chi G)\ell, \qquad (\eta, \delta_\ell, \chi) = (0.22, 0.02, 0.70),$$

and $Y, G \in [0, 1]$ with the baseline parameters ensure state-space invariance. The joint law of the next economic phase and legibility, (J', ℓ'), is mapped to the economic grid by mass-preserving interpolation, using the joint terminal-state and exposure distribution before the slow-state update rather than replacing the next state by its mean.

Value and withdrawal. Normalized current payoff values qualified entry and current qualified experience, subtracting fusion, re-closure, and unresolved-state damage, direct resources, and current capture burden. In the numerical baseline the withdrawal branch is

$$W(\ell) = \frac{w_0 + b_P\, P_{\text{persist}}(\ell) - d_R\, r_0^A(\ell)}{1-\beta}, \qquad (w_0, b_P, d_R, \beta) = (0.45, 1.80, 1, 0.94),$$

the autonomous continuation branch valued at withdrawal, not a cash asset. The Bellman system compares continued active support with this branch in A; withdrawal is unavailable in R. The first baseline grid withdrawal point is $\ell = 0.64$. Withdrawal rows of Q and active rewards there are zero, and A records exit legibility. Governance variants change the stopping gate while retaining the declared continuation actions.

Boundary. Calendar duration, provider hours, actual service-credit issuance, cash, and post-withdrawal re-enrollment are separate institutional inputs. The economic state and fast-system terms above retain their upstream meanings; they are not empirical proxies for all employment, training, or independent-living programs. The exported finite matrices allow downstream reconstruction without rerunning the individual solver.

S1.2 Finite-spell accounting

Let $\mathcal{X}$ be a finite state set, $\mathcal{W} \subset \mathcal{X}$ the withdrawal set, and Q a nonnegative substochastic matrix whose withdrawal rows are zero. The map A assigns each withdrawal row to its exit state. The initial law α is a row probability vector. An active reward y is zero on $\mathcal{W}$. For $\rho(Q) < 1$, the Neumann series converges and gives

$$N = (I - Q)^{-1} = \sum_{n \ge 0} Q^n, \qquad \alpha N y = \sum_{n \ge 0} \alpha Q^n y. \tag{S1}$$

The formula counts active rewards and any zero-reward visit to a withdrawal row. The active-duration reward is $\mathbf{1}_{\mathcal{X} \setminus \mathcal{W}}$. Withdrawal probability is $p_W = \alpha N A \mathbf{1}$ and its conditional exit law is $\alpha N A / p_W$ when $p_W > 0$. Ordinary withdrawal-to-withdrawal renewal requires $p_W = 1$. A finite matrix resolvent alone is insufficient if mass can leave through an unclassified route.

For any truncation m, the exact remaining reward is

$$\sum_{n=m}^{\infty} \alpha Q^n y = \alpha Q^m N y. \tag{S2}$$

This follows by factoring Q^m out of the Neumann series. It distinguishes a small numerical tail from a finite-horizon truncation of an improper process. A reachable closed stochastic active class prevents almost-sure withdrawal. If it also has a positive lower bound on a required per-period load, positive probability of entry implies infinite expected total load for that account. A zero-load closed class need not make every reward infinite, but still does not constitute a proper withdrawal spell.

Calendarization maps episode occupancy into dated physical or cash events. If the same nonnegative account is redistributed over calendar age by a mass-preserving recognition kernel, its total is unchanged. Converting normalized units into different economic units is a separate operation with an explicitly specified coefficient or map; it need not preserve a numerical total. The numerical example uses duration two ticks before entry or after re-closure and one tick after entry. The return-start laws, not just the withdrawal threshold, determine which exit and re-entry distributions are used.

S2 Renewal, Population Convolution, and Discounting

For each spell-start class i, let μ_i^q be the expected one-spell measure and R_{ij} the defective measure of time to the next spell start and its class j. Conditional on the new class, assume regeneration of the future process; any relevant history must therefore be included in that class. Let $P_{ij} = \mathsf{R}_{ij}(\mathbb{R}_+)$, $\rho(P) < 1$, and assume nonexplosion. Iterating the renewal equation gives

$$\boldsymbol{L}^q = \boldsymbol{\mu}^q + \mathsf{R} * \boldsymbol{L}^q = \sum_{j\geq 0} \mathsf{R}^{*j} * \boldsymbol{\mu}^q. \tag{S3}$$

Every term is nonnegative. Monotone convergence therefore identifies the minimal nonnegative solution and permits integration term by term. The total mass of R^{*j} is P^j, so $\boldsymbol{L}^q(\mathbb{R}_+) = (I-P)^{-1}\boldsymbol{\mu}^q(\mathbb{R}_+)$. The solution is finite for finite one-spell totals. Signed accounts require absolute integrability before using these interchanges; the primary implementation instead keeps cash inflows and outflows as separate nonnegative accounts.

If there is only one regenerating class with return probability $r < 1$, the formula is $Y_{\text{spell}}/(1-r)$. The renewal event occurs only after withdrawal. Applying a second multiplier based on a within-spell reversal changes the model and double-counts events already included in Q. The two operators have different state spaces and different economic meanings.

For a locally finite first-start measure Λ_0, require

$$\int_{[0,H]\times\mathcal{Z}} L_z^q([0, H-s])\Lambda_0(ds, dz) < \infty \tag{S4}$$

for every finite H. Then the population formula in the Core defines a locally finite measure by Tonelli's theorem. Let $\Lambda^{\text{all}} = \Lambda_0 * \sum_{j\geq 0} \mathsf{R}^{*j}$, with the class coordinates retained. Associativity and monotone convergence yield

$$\Lambda_0 * \boldsymbol{L}^q = \Lambda^{\text{all}} * \boldsymbol{\mu}^q. \tag{S5}$$

These are the two valid routes. A first-start measure counts new participants, whereas Λ^{all} counts all spells, including later spells of the same person.

For constant arrivals from the remote past and an integrable lifetime profile, a change of variables gives the stationary mean $\overline{Y}^q = \Lambda_P \mathbb{E}_{F_P}[\widetilde{Y}^q]$. For an empty discrete system starting at zero, the path is the partial age sum. Its distance from the stationary level at tick t is exactly $\Lambda_P \mathbb{E}[\sum_{a>t} h_Z^q(a)]$. Thus the stationary mean does not substitute for the startup path.

A calendar discount δ transforms each measure by $\int \delta^a \, d\mu(a)$. Write $P_{ij}^\delta = \int \delta^a \mathsf{R}_{ij}(da)$ and $y_i^{q,\delta} = \int \delta^a \mu_i^q(da)$. When the transform converges, the lifetime PV is $\eta(I-P^\delta)^{-1} y^{q,\delta}$. The lag is start-to-start, including current service duration and return delay. It is not generally correct to divide the individual's β-discounted episode value

Table S1: With fixed marginals, load–return alignment reverses the two-mean approximation error. Monte Carlo values are inherited numerical checks, not estimates.

Alignment	Exact mean	Simulation mean	Two-mean value
Opposed	1.432840	1.432678	1.508608
Orthogonal	1.580236	1.581455	1.508608
Aligned	1.800125	1.801502	1.508608

by $1 - r^{\text{post}}$. Neither discount is applied to physical workload or issued credit counts.

S3 Heterogeneity and Behavioral-Rule Qualifications

The covariance decomposition in Core Proposition 1 is exact under the three finite-moment conditions stated there. Uniform subcriticality $R \le \bar{r} < 1$ and $\mathbb{E}[Y] < \infty$ imply $\mathbb{E}[Y/(1-R)] \le \mathbb{E}[Y]/(1-\bar{r})$. The uniform-return counterexample in the Core shows that pointwise subcriticality is not enough. For an account with nonnegative load, an infinite population mean precludes a positive constant arrival rate served by finite constant capacity in that account.

For the Gaussian benchmark, the exponential-moment formula gives

$$\mathbb{E}[Y_q] = Y_{q0} \exp(a_q'\mu + \tfrac{1}{2}a_q'\Sigma a_q), \tag{S6}$$

$$\mathbb{E}[O_R] = \exp(\beta_0 + \beta'\mu + \tfrac{1}{2}\beta'\Sigma\beta), \tag{S7}$$

$$\mathbb{E}[Y_q O_R] = \mathbb{E}[Y_q]\mathbb{E}[O_R]\exp(a_q'\Sigma\beta). \tag{S8}$$

Substituting the last line proves the closed form in the Core. The covariance has the sign of $a_q'\Sigma\beta$. These sign statements are properties of the specified family, not distribution-free claims about all support programs. In mixtures, the law of total covariance separates within-component alignment from between-component alignment. Selection reweighting follows directly by expanding $\mathbb{E}_{F_0}[wG] = q\mathbb{E}_{F_0}[G] + \mathrm{Cov}_{F_0}(w, G)$.

For completeness, ordered continuous behavioral margins define an effectiveness lower set boundary L_z and a pressure upper boundary H_z. Their intersection is $[L_z, H_z]$ when $L_z \le H_z$ and empty otherwise. A common x is feasible almost surely exactly when nonempty sets have probability one and $\operatorname{ess\,sup} L_z \le \operatorname{ess\,inf} H_z$. A menu covers a type when at least one of its points belongs to that type's interval; hence expanding a menu cannot lower oracle coverage. This is not a guarantee that an imperfect assignment procedure attains the oracle bound.

The separate inherited 5,000-case example has 2,804 nonempty intervals. Coverage for one through six rules is 0.5280, 0.5524, 0.5580, 0.5594, 0.5606, and 0.5608, respectively. It cannot be joined person by person to the current three-case financial inputs. The Core accordingly reports financial–operational feasibility conditional on a behavioral design screen, rather than certifying that these specific participants satisfy the behavioral conditions. The missing crosswalk remains missing, not zero.

S4 Workload Reflection and Timing Comparisons

Let $X_k(t) = \int_0^t [H_k(s) - K_k(s)]ds$ and define the regulator $Y_k(t) = -\min\{0, \inf_{u\le t} X_k(u)\}$. Then $B_k = X_k + Y_k \ge 0$, Y_k is nondecreasing, and it increases only when $B_k = 0$. This constructs the one-dimensional work-conserving reflection. Any smaller regulator would leave a negative backlog at a running minimum, and any larger increase away from zero would violate minimality. Since $X_k(0) = 0$, the construction equals the supremum formula in the Core.

For constant net input $d_k = \overline{H}_k - K_k$, a positive initial workload drains when $d_k < 0$, remains unchanged when $d_k = 0$, and grows linearly when $d_k > 0$. Equality has no headroom to remove inherited backlog. A resource-weighted sum is not a substitute for componentwise tests: $K = (1, 100)$ and $\overline{H} = (1.1, 0.1)$ have aggregate slack but overload the first resource. No distribution of individual waiting times follows from a workload reflection without service requirements, routing, and a service discipline.

For a fixed nonnegative integrable kernel g, Tonelli gives $\int(\lambda * g) = (\int \lambda)(\int g)$. Also $(\lambda * g)(t) \leq \|\lambda\|_\infty \int g$, proving the peak bound used in the Core. If nonnegative cohort requirement envelopes $v_c(t - \tau_c)$ have disjoint supports, at most one is positive at any time; their aggregate maximum is $\max_c \|v_c\|_\infty$. With overlaps the maximum is between the largest individual peak and the sum of individual peaks. This last statement concerns nonnegative envelopes. A signed net-cash profile may contain offsetting surpluses, so the same peak equality cannot be applied to arbitrary cash flows by silently clipping negative components.

The numerical capacity paths are offered-load schedules. If backlog delays delivery, recognition of wages, credit redemption, outcomes, and captured saving may also change, and a capacity-admission rule may change Λ_0 or P. Core Proposition 2, proved next, shows that when the feedback acts by deterring re-enrollment with within-spell timing held fixed, the fixed-path screens bracket the resulting feedback path; delivery delay and other feedback channels require the paths to be regenerated jointly.

S4.1 Proof of Core Proposition 2 (causal deterrence bounds)

Adopt the discrete-tick setting and maintained conditions of Core Section 5.5 with $t = 0, \ldots, H$: $\lambda_t \geq 0$, $\mathsf{R}(a) \geq 0$ for $a \geq 1$ with no mass at lag zero, $\mu_i^k(a) \geq 0$, finite nonnegative capacity paths, zero initial backlogs, and a single-valued multiplier $e : \mathbb{R}_+^m \to [0, 1]$ with $e(0) = 1$, nonincreasing in each coordinate, evaluated only at b_{t-1} (with $b_{-1} = 0$). Paths are ordered componentwise.

(i) Antitonicity. Let $b \leq b'$. We show $\Lambda_t(b) \geq \Lambda_t(b')$ by induction on t. At $t = 0$ both equal λ_0. If the claim holds for all $s < t$, then $\sum_{s<t} \Lambda_s(b)\mathsf{R}(t - s) \geq \sum_{s<t} \Lambda_s(b')\mathsf{R}(t - s) \geq 0$ because $\mathsf{R} \geq 0$, and $e(b_{t-1}) \geq e(b'_{t-1}) \geq 0$ because e is nonincreasing; the product of ordered nonnegative quantities is ordered, and adding λ_t preserves the order. Hence $\Lambda(b) \geq \Lambda(b')$, and $H^k(b) \geq H^k(b')$ because $\mu^k \geq 0$. The Lindley map $h \mapsto (\max\{0, x_{t-1} + h_t - K_t\})_t$ is monotone in h by induction on t, so $\Phi(b) \geq \Phi(b')$.

(ii) Existence and uniqueness by forward construction. The component $\Phi(b)_t$ depends on b only through $b_0, \ldots, b_{t-1}$, because Λ_t uses $e(b_{t-1})$ and earlier starts, H_t uses starts up to t, and the Lindley recursion uses its already computed output $\Phi(b)_{t-1}$ together with $H_t(b)$. Construct the path forward: at $t = 0$, $\Lambda_0^* = \lambda_0$, which determines H_0^* and b_0^*. Given starts and backlogs through $t - 1$, compute the return candidates $\sum_{s<t} \Lambda_s^* \mathsf{R}(t - s)$, evaluate $e(b_{t-1}^*)$, obtain Λ_t^*, then H_t^* and b_t^*. Every operation uses only values already determined and each is single-valued. This constructs one path b^* with $\Phi(b^*) = b^*$, and any fixed point must agree with it coordinate by coordinate by the same induction. No continuity of e and no fixed-point theorem is used.

(iii) Brackets. Since $b^* \geq 0$, antitonicity gives $b^* = \Phi(b^*) \leq \Phi(0) = b^U$; applying

antitonicity to $b^* \leq b^U$ gives $b^* = \Phi(b^*) \geq \Phi(b^U) = b^L$. The same two applications yield $\Lambda^U = \Lambda(0) \geq \Lambda(b^*) \geq \Lambda(b^U) = \Lambda^L$ and $H(0) \geq H(b^*) \geq H(b^U)$. Offered physical work, credit issuance, outstanding units, exposure, operating cash, and redemption cash are each the image of the spell-start path under a fixed nonnegative linear map (nonnegative one-spell convolution, nonnegative bridge coefficients, vintage survival, and prices), so each is ordered in the same way. Running maxima and Lindley backlogs are monotone but not linear maps of these accounts; they preserve the order by monotonicity. Since the upper bounds use only $b^* \geq 0$ and antitonicity, they hold for every nonincreasing multiplier.

(iv) Reserves. Suppose funding F and captured cash S are the same under the three calculations and the floors follow a common rule $\underline{R}_t^j = r_t^0 + \kappa_R V_t^j$ with $r_t^0 \geq 0$, $\kappa_R \geq 0$, or any common order-preserving rule. Then the deficit $\underline{R}_t^j + \sum_{s \leq t}(O_s^j + C_s^{\mathrm{red},j} - F_s - S_s)$ is pointwise ordered across $j \in \{L, *, U\}$ because O, C^{red}, and V are ordered and the remaining terms coincide; taking the maximum over t and then the positive part orders the minimum reserves.

(iv′) First congestion. Core Corollary 1 follows from the same causal structure: for $t < \tau$ the monitored backlog of the fixed design is zero, so an induction identical to (ii) shows that any multiplier with $e(0) = 1$ produces the same starts, loads, and backlogs, and at $t = \tau$ the multiplier still evaluates $e(0) = 1$, so the first positive backlog coincides. In the numerical study the first platform overload of the staged schedule occurs at tick 50 with offered load 4.875 in the fixed design and under every tested multiplier.

(v) Whole-path iteration. Let $b^{[n+1]} = \Phi(b^{[n]})$ from any initial path. By (ii), $b_0^{[1]} = \Phi(b^{[0]})_0$ does not depend on $b^{[0]}$ and equals b_0^*; if $b_s^{[n]} = b_s^*$ for $s \leq t-1$, then $b_t^{[n+1]} = \Phi(b^{[n]})_t = b_t^*$. Hence $b_t^{[n]} = b_t^*$ for $n \geq t+1$, and the path on $0, \ldots, H$ is recovered after at most $H+1$ sweeps. From $b^{[0]} = 0$, antitonicity and $0 \leq \Phi^2(0) \leq \Phi(0)$ give nondecreasing even iterates and nonincreasing odd iterates, every even iterate lying below every odd iterate. $\square$

Remarks. (a) Uniqueness here comes from strict causality (the one-tick lag in the multiplier and the absence of return mass at lag zero), not from antitonicity: for a general antitone map a unique fixed point does not imply convergence of the alternating iteration (for $f(x) = 1 - x$ on $[0, 1]$ the iteration from 0 alternates between 0 and 1). A multiplier evaluated on the same tick's backlog would reintroduce a genuine simultaneous fixed-point problem. (b) If the multiplier depends on the backlogs of several resources, the argument is unchanged with the componentwise order on $\mathbb{R}_+^m$. (c) The argument retains first-spell admissions as exogenous and uses a multiplier that suppresses return *mass* at the stated lag. Deferral of returns and endogenous first-spell uptake are outside the maintained model and need separate order and cash-capture analyses; they need not, in every specification, destroy antitonicity. Uptake that is nonincreasing in the prescribed backlog can act in the same order direction as the deterrence multiplier. For deferral kernels or uptake that increases with backlog, this order comparison is not generally guaranteed; no general extension is claimed here. If first-spell admissions become endogenous, the shared funding and captured-cash paths required by part (iv) also need to be established separately. (d) $e(0) = 1$ normalizes the multiplier to the baseline return kernel, which already contains the no-congestion return probabilities.

S5 Vintage Conservation and the Cash Counterexample

In discrete time, let $l_{t,a}$ be the closing outstanding balance of age a. Under the numerical convention, opening balances face redemption rate d, cancellation rate b_L, and survival $q_L = 1 - d - b_L$. Survivors of the maximum age expire, after which new issuance is posted:

$$D_t = d \sum_a l_{t-1,a}, \qquad B_t^{\text{burn}} = b_L \sum_a l_{t-1,a}, \tag{S9}$$

$$X_t = q_L l_{t-1,E-1}, \qquad l_{t,0} = I_t, \qquad l_{t,a+1} = q_L l_{t-1,a} \quad (a < E-1). \tag{S10}$$

Consequently $L_t = L_{t-1} + I_t - D_t - B_t^{\text{burn}} - X_t$. The discrete survival mass is $\sum_{a=0}^{E-1} q_L^a$, so bounded per-tick issuance gives bounded stock. In the continuous case the same argument integrates the survival kernel against issuance and bounds it by $\bar{I} m_L$. Legacy stock adds its surviving balance, bounded by opening units. With common fixed price p, exposure is pL; with age-dependent prices it is the integral against the vintage ledger. Expiry or cancellation is not a cash receipt.

For general cash events, no-interim-shortfall is checked at every recognition event, not merely at arbitrarily coarse reporting dates. The continuous-rate identity in the Core (Lemma 1) and the following tick convention are alternative representations of that same requirement. At tick end, with net cash f_t, the reserve is $R_t = R_0 + \sum_{u \le t} f_u$. The minimum is $\max(0, \max_t[\underline{R}_t - \sum_{u \le t} f_u])$. If the same floor is tested after outflows but before that tick's inflows, an additional condition is

$$R_0 \ge \max_t \left\{ \underline{R}_t - \sum_{u<t} f_u + O_t + C_t^{\text{red}} \right\}. \tag{S11}$$

The reported outflow-first sensitivity uses precisely this conservative ordering with the closing exposure floor. It does not change the final net cash total.

NPV and total cash are insufficient statistics for reserve need. For $c > 0$ and $0 < \delta < 1$, compare net flows $A = (0, 0, 0)$ with $B = (-c\delta, c(1 + \delta), -c)$ at dates $0, 1, 2$, using zero floors. Both total net cash and discounted net cash are zero. Yet A requires no initial reserve while B requires $c\delta$. A separate positive-NPV example is an outflow of 100 now and an inflow of 120 next period with $\delta = 0.9$: NPV is 8, but zero initial cash cannot pay the first outflow. These are algebraic illustrations of the inherited nonidentification argument, not new calibrated cases.

S6 Governance: Stopping, Load, and Reachability

The stopping advantage D_z is evaluated at the benchmark continuation value. It is not silently replaced by the solution of a provider Bellman equation. On the monotone safe-capable region, $D_z(\ell) = \varpi - b$ defines the crossing. Differentiability and strictly positive derivative give the two signs stated in the Core. The set $0 \le D_z(s) < \varpi - b$ consists of benchmark-withdrawal states retained by this rule. Setting $b = \varpi$ removes this stopping wedge but supplies neither a participation constraint nor the cash price of compensation.

For load monotonicity, couple paths until the benchmark stopping time. The retained path starts with the same accumulated active loads, then adds only nonnegative loads. Multiplication by the nonnegative renewal resolvent preserves a componentwise

Table S2: The stopping-only cases retain the same return matrix and institutional bridges. The last row is excluded from ordinary renewal accounting, not interpolated as a finite throughput.

w	Threshold	$\rho(Q)$	Rate limit	Binding screen
0.000	0.64	0.687	0.630	credit exposure
0.100	0.66	0.688	0.613	credit exposure
0.250	0.68	0.689	0.593	capacity
0.500	0.76	0.705	0.489	capacity
0.750	0.86	0.793	0.331	capacity
0.900	0.92	0.980	0.089	capacity
0.925	0.94	1.000	–	nonproper

increase in one-spell load vectors when P is fixed. The argument does not order calendar densities pointwise and does not cover all cases with endogenous changes in P. In particular, improving the exit state may lower subsequent return need while increasing current spell length.

The provided $Q_{0.925}$ is a 102-state matrix. Starting at phase R and grid index 3 ($\ell = 0.06$), its positive-entry graph has 94 reachable states, all active. The reachable submatrix has row sums one and forms a closed communicating class. There is no reachable zero withdrawal row. Hence this finite-grid process cannot exit through the recorded withdrawal map. This is an exact reachability property of the stored numerical graph; it is not a proof that a continuum model or every alternative grid has the same class.

A hard active-age cap H_g augments the state by age and removes active mass at that age. The active block is then nilpotent with spectral radius zero, and nonnegative per-episode load at most $\bar{y}$ gives total load at most $H_g\bar{y}$. The forced exit is recorded as its own event, not relabeled as a successful benchmark withdrawal. Finiteness does not prove durability or participant safety.

S7 Numerical Recipe and Additional Stress Checks

The companion inputs retain the baseline matrices, the three-class lifetime profiles, the return kernel, and the source calendar paths. The Core driver applies the declared bridges anew, rebuilds credit vintages, and independently recomputes reserves and the displayed stationary bounds. It does not re-solve the upstream individual control problem or recalibrate any behavioral parameter.

The selected law follows from target weights times eligibility times conditional uptake/access. The return matrix is shown in the Core; additional waits, measured after withdrawal in synthetic ticks, are

$$W^{\text{wait}} = \begin{pmatrix} 5 & 6 & 7 \\ 4 & 5 & 6 \\ 3 & 4 & 5 \end{pmatrix}. \tag{S12}$$

The start-to-start return kernel combines these waits with the current spell's exit-age distribution. Its total mass is P. Lifetime account arrays run through age 650 and were generated by renewal convolution; their totals are checked against the matrix resolvent.

Table S3: Testing the floor after outflows and before inflows increases required initial cash without changing terminal net cash. All amounts are synthetic currency units.

Schedule	Period-end	Outflows first	Increase
Front-loaded	22,081	22,231	150
Staged	5,709	5,972	263
Uniform	4,251	4,470	219

Exact within-spell tail identities and source tail residuals are retained. The 1,400-tick population reporting grid extends beyond admissions and the retained account profiles; padded zero capacity beyond the original service grid is checked to coincide with negligible demand rather than treated as a new active resource constraint.

Let h^A, h^U, h^S, h^W denote the four lifetime bases described in the Core. Physical-work and credit maps are precisely those printed there. Operating cash is $120h^D + 80h^M + 12h^P$. Credits redeem at 0.10 and cancel at 0.02 per tick, expire after 30 ticks, and settle at price 25. The resulting survival sum is 8.153322. At one base-opportunity unit per tick, steady issuance is 9.337714 and steady stock 76.133392. These are undiscounted units.

External funding is 150 each tick from 0 through 173 inclusive and zero thereafter, independently of rollout. Captured cash saving is 220 per first participant 50 ticks after admission. An optional separate valuation assigns 1,000 social-value units after 20 ticks and 60 external-damage units per participant, without treating either as fund cash. The cash-capture timing is assumed, not a measured consequence of withdrawal or governance. The baseline fund discount for its reported NPV is 0.999 per tick. It is not the individual model's $\beta = 0.94$.

Dual-route reconstruction. The Version 0.2 driver reconstructs the one-spell profiles μ_i^q by class from the frozen lifetime profiles through the deconvolution $\boldsymbol{\mu}^q = \boldsymbol{L}^q - \mathsf{R} * \boldsymbol{L}^q$, which is exact because $\boldsymbol{L}^q$ is the minimal solution of the renewal equation; the reconstructed one-spell totals agree with the stored totals to 10^{-9} and are nonnegative. Convolving the all-spell-start path with these one-spell profiles (Route B) reproduces the inherited first-spell-start route (Route A) on every account and schedule to within 5×10^{-8}.

Feedback bracket. The gate $e(B) = \exp(-B/\hat{B})$ is applied to re-enrollments starting at tick t, with B the platform backlog at the end of tick $t-1$ under constant platform capacity 4.75. Core Table 4 reports $\hat{B} = 10$ and $\hat{B} = 40$. For $\hat{B} = 40$, about eight ticks of capacity, the staged schedule's equilibrium has 50.7 spell starts, peak platform 6.05, peak exposure 1,549, and minimum reserve 3,636 (fixed design: 58.7, 6.77, 1,717, 5,709), the front-loaded equilibrium has minimum reserve 15,923, and the uniform schedule is unchanged; every screen outcome is the same as in the fixed design. The feedback path is computed by the forward recursion of Core Proposition 2(ii); the whole-path sweeps $\Phi^n(0)$ reproduce it to 10^{-9} within 12 sweeps in all six instances, and it lies between the lower and the fixed-design paths on every reported quantity. A direct forward construction that does not use whole-path iteration reproduces the same paths.

Coefficient sensitivity. Core Table 5 scales, one group at a time, the physical-work coefficients, the credit-issuance coefficients, the operating cash per resource unit, the settlement price, the redemption and cancellation rates, the expiry age, the captured saving per participant, the capture delay, the reserve-floor ratio, the resource capacities,

Table S4: Sensitivity of the stationary governance screen. The screening boundary decreases in the wedge under every rescaling; the wedge at which capacity becomes binding depends on the relative levels of the capacity and exposure limits. C, E, and L denote capacity, credit exposure, and cash flow.

Scaled input	Limit, $w = 0$	Limit, $w = 0.9$	Decreasing	Binding at $w = 0, .1, .25, .5, .75, .9$
Baseline	0.630	0.089	yes	E/E/C/C/C/C
Resource capacities ×0.5	0.324	0.045	yes	C/C/C/C/C/C
Resource capacities ×1.5	0.630	0.126	yes	E/E/E/E/E/L
Exposure ceiling ×0.5	0.315	0.066	yes	E/E/E/E/E/E
Exposure ceiling ×1.5	0.648	0.089	yes	C/C/C/C/C/C
Settlement price ×0.5	0.648	0.089	yes	C/C/C/C/C/C
Settlement price ×1.5	0.420	0.089	yes	E/E/E/E/E/E
Steady funding ×0.5	0.472	0.063	yes	L/L/L/L/L/L
Steady funding ×1.5	0.630	0.089	yes	E/E/C/C/C/C

and the exposure ceiling. Table S4 reports the corresponding rescalings for the stationary governance screen.

For the stationary governance screen, funding is instead 300 per tick, and each proper case has an operating-plus-redemption-minus-capture flow per target-opportunity unit. The cash-flow rate limit is 300 divided by this positive coefficient; capacity and exposure limits are computed separately and their minimum is reported. This stationary scenario must not be confused with the finite 26,100-funding rollout scenario or its initial-reserve requirement.

The common-shock stress multiplies scheduled arrivals by $M_t = \exp(0.15F_t - 0.15^2/2)$, where $F_t = 0.85F_{t-1} + \sqrt{1 - 0.85^2}\eta_t$, $F_0 \sim N(0, 1)$ and independent $\eta_t \sim N(0, 1)$. The multiplier is shared by entrants at a date and has mean one. A common factor does not vanish under population averaging: for $X_i = \mu+\lambda F+\varepsilon_i$ with independent mean-zero idiosyncratic terms, the average converges to $\mu+\lambda F$, not to μ. The numerical stress is therefore conditional on a shared-factor law rather than an idiosyncratic law-of-large-numbers approximation.

An earlier 5,000-path grid exploration selected scales 0.87 and 0.88 for a separate run. The held-out source experiment fixes these two candidates, seed 260915, and 10,000 paths. With x successes in n paths, each exact one-sided lower bound is $\mathrm{Beta}^{-1}(0.025; x, n - x + 1)$. Bonferroni gives at least 95% simultaneous coverage for the two bounds, without requiring independence between the candidates' outcomes. The counts are 9,611 and 9,431, giving proportions 0.9611 and 0.9431 and lower bounds 0.9571236 and 0.9383791. These results concern only the declared candidates under the synthetic model, not an optimum or real-world risk guarantee. The Core driver checks the stored counts and binomial bounds; it does not claim to have rerun the full held-out Monte Carlo in this compression step.

S8 Reproducibility, Missing Inputs, and Claim Boundary

The two PDF files contain all arguments needed for the retained Core claims. The computational package supplies frozen numerical inputs, a source-hash manifest, the reconstruction driver, the Version 0.2 extension driver for the feedback bracket and the sensitivity tables, plotted exhibit data, and checks. Its functions distinguish normalized process outputs, physical resource equivalents, actual modeled credit units, cash, and social values. Frozen inputs are not interpreted as observed data. The full development manuscript is retained separately as a research archive and is not a substitute for a missing proof in this supplement.

Missing empirical fields include population weights and access, episode-to-calendar duration, provider-hour conversion, actual credit issuance and redemption, funding commitments, captured saving, common shocks, and provider compensation. In addition, there is no measured map from legacy behavioral cases to the current MRP cases, no complete provider contracting solution, and no calibrated allocation of credit-redemption services to shared resource pools. These gaps limit external interpretation; they are not filled with zero values. Only the declared financial–operational model is screened numerically.

This compressed core deliberately omits the full legacy behavioral optimization, measurement-salience extension, broad welfare comparisons, and unestimated field protocols from the manuscript's main result set. They remain research material rather than being used to enlarge the contribution or to evade the supplement page budget. The new closely related service-returns, discharge, and loyalty-liability comparisons in the Core are bibliographic and positioning updates, not changes to the underlying numerical mechanism.