Populace dynamics: an open, scored longitudinal layer for policy microsimulation

Max Ghenis

July 2026

Abstract

Dynamic microsimulation — the longitudinal aging of a person-level population through earnings, family structure, disability, mortality, and program participation — underpins retirement and social-insurance policy analysis, yet the benchmark United States models — DYNASIM, MINT, and CBOLT — are closed: internal to government, tied to restricted administrative records, or accessible only through institutional relationships. This design paper specifies an open alternative built as an extension of Populace, PolicyEngine's open-source microdata stack, whose cross-sectional layer is the certified default United States microdata in PolicyEngine after outperforming its predecessor on held-out administrative targets. The design contributes four elements: a trajectory-weighted kernel in which multi-period calibration cannot silently destroy panel structure; a Dynamics operator that treats state transitions as conditional models, mixing deterministic demographic hazards with machine-learned earnings processes; an explicit domains-of-validity framework that refuses point forecasts where parameter uncertainty dominates — including the 75-year actuarial balance — and instead publishes sensitivity surfaces; and a scoring protocol under which every claim resolves against administrative publications, backtests with leakage control, or computes exactly from statute, with contributions merging only when they improve held-out scores. United States Social Security is the first validation domain; the layer itself is country-agnostic.

1 Introduction

Analysts who want to model taxes can run open, calibrated models directly: Tax-Calculator (Policy Simulation Library 2026) and PolicyEngine (PolicyEngine 2026) are openly callable on public data, alongside source-available models with restricted inputs (The Budget Lab at Yale 2026) and proprietary models used for outside-facing analysis (Tax Policy Center 2025; Institute on Taxation and Economic Policy 2025; Tax Foundation 2025; Penn Wharton Budget Model 2025). Analysts who want dynamic microsimulation — the longitudinal modeling that retirement and social-insurance policy requires — find the benchmark models closed: SSA projects retirement income and the distributional effects of policy proposals with MINT (Social Security Administration 2024), CBO produces its long-term Social Security projections with CBOLT (Congressional Budget Office 2018, 2024), the Urban Institute projects retirement income and long-term care with DYNASIM (Favreault et al. 2015; Urban Institute 2024), and Morningstar studies retirement adequacy with its Model of US Retirement Outcomes (Look and VanDerhei 2024). Outside users reach each only through an institutional relationship. The nearest open analogue, the Cato Social Security model (Chanwong 2026), simulates roughly 10,000 households from the 2007 CPS ASEC under the Social Security Administration's assumptions and reports trust-fund metrics and reform scores, without published validation against administrative benchmarks.

This paper specifies the design of an open longitudinal layer — *Populace dynamics* — as an extension of Populace, PolicyEngine's open-source microdata stack. It is a design paper: the cross-sectional

foundation is in production; this paper specifies the longitudinal layer, which the project builds in the open behind the scoring protocol of Section 5.

The design makes four contributions:

1. **A trajectory-weighted kernel with explicit alignment.** One weight per trajectory: multi-period calibration stacks constraint rows over the same weight vector, so that hitting cross-sectional totals in multiple periods cannot silently destroy panel structure — combined with event-selection alignment for period-by-period control, since weights alone cannot separate a correct life course from a correctly timed one (Section 4).
2. **Transitions as conditional models.** A single operator interface covers deterministic demographic hazards and machine-learned earnings processes, so candidate architectures compete under one evaluation standard (Section 4).
3. **Domains of validity as shipped metadata.** The model states which questions it will not answer with false precision — led by the 75-year actuarial balance, where input uncertainty dominates model fidelity — and publishes long-horizon results as sensitivity surfaces rather than point forecasts (Section 3).
4. **A scoring protocol in place of fidelity-only validation.** Claims resolve against administrative publications on an annual calendar, backtest against realized history with leakage control, or compute exactly from statute; contributions merge only when they improve held-out scores (Section 5).

United States Social Security is the first validation domain because benefit adequacy and reform incidence depend jointly on lifetime earnings, marriage and survivorship, disability, differential mortality, and claiming. The layer itself is country-agnostic and extends to other pension and benefit systems as PolicyEngine's country coverage grows.

2 Terminology and scope

This paper uses "dynamic" in the microsimulation field's standard sense, following Orcutt et al. (1961) and the tradition carried by DYNASIM, MINT, CBOLT, and SimPaths (Bronka et al. 2025): a longitudinal model that ages a person-level population through time (Li and O'Donoghue 2013). It does not mean "dynamic scoring" — the tax-policy usage denoting macroeconomic feedback in revenue estimation — and the design is not an overlapping-generations general-equilibrium model of the Auerbach–Kotlikoff kind, such as the Penn Wharton Budget Model operates (Penn Wharton Budget Model 2025). The model claims neither macroeconomic feedback nor equilibrium closure. Behavioral responses enter as labeled scenario inputs with documented ranges, not as point estimates carrying model authority.

3 Domains of validity

All models are wrong; a model earns its keep only where it improves predictions. The design therefore begins by bounding its own claims.

3.1 Parameter uncertainty dominates the long horizon

The 75-year actuarial balance of a pension system is a function of a small set of exogenous assumptions — fertility, mortality improvement, net immigration, real wage growth, interest rates — whose uncertainty dominates the microsimulation machinery that processes them. The public record makes this concrete for United States Social Security. The Trustees' own low- and high-cost scenarios bracket a range of 75-year balances wider than the intermediate deficit itself (Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds 2025). The Congressional Budget

Office's long-term projections differ from the Trustees' — chiefly for assumption reasons, though method differences such as CBO's micro-founded projection of the taxable share of earnings also contribute to the divergence (Congressional Budget Office 2024). Assumption variance dominates the headline; model structure matters for distribution and the near term, which is the division of labor the output tiers encode. Successive Technical Panels convened by the Social Security Advisory Board have recommended revising fertility and mortality-improvement assumptions as realized values ran persistently outside the projected path (Technical Panel on Assumptions and Methods 2023). Two institutions with administrative data and decades of refinement disagree with each other and have both missed realized demographic trends; a better microsimulation does not repair that, because the variance lives in the inputs.

3.2 Three output tiers

The model sorts its outputs into tiers, each carrying the strongest claim it can support.

Tier 1: distributional analysis under fixed assumptions. Reform analysis is a difference — outcome under reform minus outcome under baseline, holding the population and assumption path fixed — and much of the unforecastable demographic uncertainty is common to both arms and cancels in the difference. The required ingredients — a calibrated joint distribution of lifetime earnings, family structure, and differential mortality, plus an exact rules engine — are the components the scoring protocol validates directly. The design labels rather than buries the slice of a reform delta that does not cancel: reform-induced claiming responses, and interactions between the reform and uncertain dynamics — the mortality gradient enters many deltas as a covariance with benefit position, not a level, so it does not difference out. Deltas evaluated past trust-fund depletion also require an explicit scheduled-versus-payable baseline convention, which the model states with every such output. For claiming, scenario ranges anchor to the quasi-experimental record on retirement-age responses (Mastrobuoni 2009; Behaghel and Blau 2012), and the model publishes them as a scenario library rather than embedding them as point estimates.

Tier 2: near-term components that resolve. Over roughly a ten-year horizon, mechanics rather than demographic extrapolation dominate outputs: beneficiary counts by type, average benefits, covered earnings and taxable payroll, claiming-age distributions, disability incidence. These resolve against administrative publications each year, and the protocol scores them (Section 5).

Tier 3: the long horizon as a sensitivity surface. The model publishes long-horizon outputs as surfaces over documented assumption ranges — how the balance, cohort replacement rates, or distributional outcomes move as fertility, mortality improvement, and immigration vary — never as point forecasts. The distinction is between computing and blessing: the model computes 75-year balances and depletion dates *conditional on named assumption paths*, including the Trustees' intermediate path, so the numbers the policy debate runs on remain available as labeled conditional outputs; what the model declines is presenting any single path as its own forecast. Incumbent practice publishes the point estimate up front and the sensitivity analysis in an appendix; an open model can invert that and make the sensitivity the interface.

Every API response carries its tier, assumption path, and calibration history as metadata, so a downstream consumer — human or machine — weights the output by demonstrated reliability rather than by the producer's reputation.

4 Architecture

4.1 The cross-sectional foundation

Populace builds a calibrated synthetic population entirely from primary-source government data — the Current Population Survey ASEC, the IRS Public Use File, the Survey of Consumer Finances, SIPP, CPS outgoing-rotation groups, MEPS, and the ACS — synthesizing missing variables with weight-aware conditional models and calibrating to administrative aggregates treated as uncertainty-weighted facts. In June 2026 it replaced PolicyEngine's enhanced CPS as the certified default United States microdata in PolicyEngine, after a matched, symmetric-refit comparison on 41,314 households with a 739-target holdout (Table 1).

Table 1: Certification comparison from the Populace release manifest. The enhanced CPS wins more individual targets by small margins while its largest misses dominate the loss; the scoring protocol requires publishing the count that cuts against the headline.

Metric (lower is better)	Populace	enhanced CPS
Holdout loss (739 held-out targets)	0.038	0.317
Training loss	0.190	1.089
Full loss	0.228	1.405
Per-target wins	1,040	2,613 (51 ties)

4.2 Trajectory weights and population accounting

The longitudinal extension follows two kernel rules set in Populace's charter. First, **one weight per trajectory**: multi-period calibration targets stack as (target, period) constraint rows over a single trajectory-level weight vector, so that calibrating cross-sections independently — which severs the trajectory-level consistency a panel exists to provide — is a kernel-level error rather than a modeling temptation. Second, **population is not closed**: trajectories carry entry and exit markers (birth, death, immigration, emigration), and a trajectory's weight contributes to a period only while the person is present. Household and couple links are period-scoped, so family recomposition preserves per-period accounting identities; couples carry a shared unit weight derived from their trajectory weights so that spousal and survivor benefits have a well-defined representation.

Weights alone are one layer of alignment, not the whole answer: a weight cannot distinguish a correct life course from a correctly timed one, and reweighting trajectories to hit future cross-sectional cells risks selecting on entire correlated life courses. Period-by-period control therefore also operates through event selection — ranking individual transition probabilities and selecting the number of events an external control demands, the mechanism CBOLT and DYNASIM use — with trajectory weights reserved for base-year representation and slow-moving composition (Li and O'Donoghue 2013; Dekkers and Cumpston 2012). The kernel solves the stacked constraint system as uncertainty-weighted penalized least squares against target standard errors, so infeasible combinations resolve by SE-weighted compromise rather than silent failure, and projected demographic controls pin cohorts born after the base year rather than leaving them free.

4.3 The Dynamics operator

Dynamics is an operator from a population and a transition specification to a population with extended periods. A transition is a conditional distribution — the probability of next-period state given current state and covariates — which is the same interface Populace's synthesis models already implement. The shipped baseline for earnings is a regime-gated, sequentially chained, weighted quantile-regression-forest imputer (Meinshausen 2006) whose zero-inflation gate doubles as a nonemployment model; richer

architectures (zero-inflated neural distribution models, normalizing flows) are candidates that must beat the baseline on held-out longitudinal moments to merge.

The design is hybrid. Where the evidence base is tabular — mortality from official life tables with published income gradients, fertility from vital statistics, marriage and divorce from ACS- and CPS-based rates (federal collection of detailed marriage and divorce statistics ended in the 1990s), disability incidence from program statistics — transitions are deterministic hazards, auditable row by row. Marriage requires a matching model, not only a hazard: spousal and survivor benefits depend on assortative matching over lifetime earnings, which the design treats as a first-class estimation target rather than an afterthought. The design reserves machine learning for processes with rich conditional structure, led by earnings dynamics — where the process is not stationary: volatility and mobility vary by cohort and period in administrative data (Sabelhaus and Song 2010; Kopczuk et al. 2010), so transition models carry cohort and period conditioning rather than pooling across decades. Chained one-period models also understate long-spell persistence unless spell structure enters the model explicitly, and backcasting is a distinct conditional object from forward simulation rather than the same operator reversed; both enter the evaluation as targets in their own right, disciplined by held-out panel moments such as higher-order earnings-change distributions (Guvenen et al. 2021).

The design states one measurement caveat rather than hiding it: the most demanding earnings-dynamics moments come from administrative records that public panels understate, and survey- and administrative-based estimates disagree on volatility levels and trends. Where the project cannot recompute a published administrative moment on held-out public data, matching it is calibration, not validation, and the scorecard labels it as such.

4.4 Rules and delivery

Statute is the deterministic slice of any policy forecast. Core retirement benefit formulas — average indexed monthly earnings, primary insurance amounts, actuarial adjustments — and benefit taxation compute exactly today through PolicyEngine's rules engine via Populace's rules-adapter protocol, vectorized over person-periods. Auxiliary, spousal, and survivor benefit formulas are explicit build items in the validation program: the current engine carries them as calibrated aggregates rather than person-level formulas, and the scorecard treats "computes exactly" as a per-rule status each formula earns, not a blanket claim.

The rules adapter is engine-agnostic. PolicyEngine-US implements it today; Axiom — an open project that encodes statute declaratively and compiles it to Rust — is the next adapter, and the performance headroom matters when benefit formulas run over person-periods across the full trajectory panel and many reform scenarios. In that architecture, PolicyEngine is a composition: Axiom supplies the rules, Populace supplies the population, and behavioral responses enter as the labeled scenario layer.

Data governance is a design requirement, not an afterthought. Estimating transition models on restricted-use panels and publishing the estimated parameters is settled practice — open models such as OG-USA, and DYNASIM itself, estimate on the PSID and publish what they learn. Releasing a synthetic *microdata* artifact informed by such panels is a stricter problem, because donor-based samplers can emit observed training values. The design answers it structurally: released records originate from Populace's public-use cross-section, restricted panels train processes rather than donate records, samplers smooth or noise their draws so no verbatim donor value ships, and every release passes nearest-neighbor disclosure checks alongside its accuracy scorecard — and the project engages data producers directly where their terms of use warrant it. Uncertainty budgets therefore attach only to the components statute does not fix. The deliverable is a versioned artifact — a longitudinal population release with a manifest and scorecard, certified through the same path as the cross-sectional release — exposed through a Python library, a REST API, and a Model Context Protocol server so that AI agents can run baseline distributions and reform analyses with validity metadata attached.

5 Scoring and resolution

Validation by fidelity — does the model match published aggregates? — is necessary but weak: a model can reproduce the tables its authors fit it to. This project's standard: a claim counts as validated when it improves prediction of something that later resolves. Five scoring surfaces implement it.

1. **Annually resolving components.** Beneficiary counts by type, average and aggregate benefits, covered earnings and taxable payroll, cost-of-living adjustments, disability incidence, and claiming-age distributions resolve against administrative publications each year. Every published forecast cell carries a resolution rule naming the exact table and vintage that settles it.
2. **Forecasting the forecasters.** Official projections revise every year; predicting the next revision of headline quantities resolves in months rather than decades and is decision-relevant to anyone who acts on the official number. This is partly forecasting an assumptions process — panels advise, committees adopt with a lag — so each cell pre-specifies the naive baseline it must beat (an assumption random walk plus mechanical data update).
3. **Retrodiction with leakage control.** Retrodiction builds the model from data vintages available at a historical date and scores it against realized outcomes. Populace's versioned data registry pins vintages going forward; pre-registry history is a reconstruction problem — survey redesigns, revised administrative tables, re-released panels — so the protocol grades pre-registry backtests as pseudo-vintage, with a published log of deviations from true vintage, and no registry pins specification leakage: a "2005-vintage" model built today knows 2008 happened, and the protocol says so. Retrodictive calibration under the historical regime does not guarantee calibration under a new one, so backtests complement rather than substitute for live resolution.
4. **Statutory resolution.** Where an output is fixed by law, the rules engine computes it exactly, and enacted policy settles the corresponding conditional cells immediately.
5. **Held-out panel moments.** The protocol scores the population layer against moments it never fit: earnings-mobility matrices, autocorrelation and higher-order moments of earnings changes, cohort age-earnings profiles, and family-transition rates on held-out panel records.

Two governance rules complete the protocol. **Merge on score**: a contribution — a mortality module, a claiming model, an earnings architecture, from any contributor — merges if and only if it improves the population's score on held-out facts, the rule Populace already applies to its cross-sectional layer. **Publication discipline**: misses publish with the same prominence as hits; superseded methods keep their historical scorecards; and stage gates in the development roadmap are pre-specified score thresholds, not narrative judgments.

Openness supplies most of the refereeing. Resolution rules are pre-registered, scores recompute from public data, and anyone who distrusts a published scorecard can rerun it — or fork the project and publish a rival scorecard. Two pieces sit beyond reproduction. First, restricted-data checks: the most demanding earnings-history moments live in linked administrative records that the public pipeline never touches, so the reader must trust whoever runs the comparison — author or outsider — and a validator without a stake in the result adds evidence where rerunning is impossible. Second, judgment: gate thresholds, disputed resolution rules, and the assumptions library carry discretion that pre-registration narrows but does not remove. The project defines its validation procedures and gate thresholds before the components they gate, and seats an advisory board to review them — and disputes under them — in public. Independent scoring across models is a decision only a third party can make; the design encourages it — published projections from the closed models give any such body a comparison set without model access — but does not depend on it.

6 First validation domain: U.S. Social Security

Social Security is the first domain because eligibility and benefits depend on the highest thirty-five years of indexed earnings, marital and survivorship histories, disability pathways, differential mortality,

and claiming timing — jointly. A layer that scores well here earns reuse in adjacent domains — Supplemental Security Income interactions, long-term care, and retirement adequacy, the last of which is *harder*, not easier: wealth projection challenged even administrative-data models (Favreault and Smith 2016), and a wealth and pension roadmap is future work, not an assumed extension — and in other countries' pension systems.

The domain also makes the validity framework concrete. The canonical output of existing Social Security models — the 75-year balance and depletion date — is the quantity Section 3 declines to forecast. What the open layer offers instead is the combination the field lacks: tier-1 distributional incidence of reforms with reproducible assumptions; tier-2 near-term components with a public resolution record; and tier-3 sensitivity surfaces that make the assumption-dependence of long-horizon claims the interface rather than the appendix. No existing model, closed or open, publishes that combination. The closed benchmarks do publish validation and cohort tables; what outsiders cannot do is rerun the pipeline, vary its assumptions, or audit the intermediate states behind the published numbers.

7 Gate 1 in practice: the first pre-registered runs

The protocol of Section 5 stopped being a design in July 2026. This section reports what it produced: a locked gate, five registered and scored model runs, and the findings the failures purchased. Every number below recomputes from committed artifacts in the project repository; the run log lives in the repository's pull requests and the candidate registry in issue #42.

7.1 The lock

Gate 1 — earnings-history credibility — locked on 2026-07-05. Thresholds derive from committed noise-floor artifacts measured at the scale the protocol scores at (two disjoint 20%-of-persons samples of the PSID family earnings panel scored against each other), with every derivation stated as floor mean plus a named multiple of the floor's seed standard deviation and enforced by a test: a floor rebuild that shifts any artifact fails continuous integration rather than silently orphaning the rationale. Before ratification the proposed thresholds went through three rounds of adversarial review, published in full on the ratification pull request. Round one found that the proposed numbers did not follow their own stated rule and were calibrated against a floor at five times deployment scale; round two found the same wrong-scale defect reintroduced on a second view and demonstrated that no window-2 statistic can catch a chained one-period model, forcing the persistence guard onto the battery's long-horizon autocorrelation bands; round three verified the amendments and ratified. The merge of the ratification pull request is the lock event; the thresholds have not changed since, and any change requires a public amendment and a fresh review round.

7.2 Five runs, four failures, and what they isolated

Each candidate registers its complete specification — every modeling degree of freedom pinned — before its single scored run. The log autocorrelation ladder at 2, 4, and 10 years (locked bands 0.730±0.05, 0.657±0.06, 0.539±0.07) tells most of the story:

candidate	2yr	4yr	10yr	verdict
chained weighted QRF (baseline)	0.726	0.573	0.333	fail
+ person effect as feature	0.726	0.688	0.649	fail

candidate	2yr	4yr	10yr	verdict
+ persistence-aware decomposition	0.722	0.695	0.647	fail
structural three-component assembly	0.464	0.401	0.354	fail
donor splicing (single-donor)	0.779	0.704	0.616	fail
donor splicing (segments)	0.720	0.631	0.490	fail
Gaussian-copula rank dynamics	0.716	0.653	0.507	fail
empirical rank kernel	0.692	0.548	0.381	fail
k-NN rank bootstrap	0.719	0.636	0.459	fail
permanent-rank matching	0.791	0.733	0.670	fail
calibrated blend + Q0 regime	0.757	0.677	0.514	fail
inner-validated composition	0.773	0.695	0.510	fail
inner-validated composition (re-registered)	0.773	0.695	0.510	**pass**

The baseline failed exactly where the pre-lock review predicted a one-step chain must: window-2 geometry passed on every seed while the 10-year autocorrelation collapsed toward the Markov value. The second and third candidates added a latent person effect as a conditioning feature — first from a naive variance decomposition, then from a persistence-aware one that halved the drawn variance — and produced statistically identical ladders. That invariance is the program's sharpest finding to date: a quantile forest rescales its response to a person-effect feature inversely to the feature's scale, so conditional-draw generation transmits the raw within-panel person-mean variance share (0.647 on the training splits, inflated by transitory persistence over few biennial observations) no matter how the feature is constructed. No feature-side dial can set person-level variance. The fourth candidate set the variance structurally — assembling log earnings from an age profile, a person effect drawn at the decomposed permanent variance, a chained transitory component, and an observation-noise layer — and the variance landed as designed while the marginal distribution broke: parametric lognormal assembly inflated levels where the earlier candidates had inherited the data's marginal from empirical conditional draws by construction.

The symmetry across the four failures defines the remaining design problem. Conditional-draw candidates preserve the marginal but cannot set person-level variance; the structural candidate set the variance and lost the marginal. The data demand both at once.

The generative track then ran the composition the taxonomy below suggests. A Gaussian-copula rank model — empirical quantile marginals, a calibrated permanent-plus-transitory latent, simulated-moment calibration — put the autocorrelation ladder inside its bands on every seed, the first generative candidate to do so, and its calibrated shares independently reproduced the variance decomposition estimated three candidates earlier; but Gaussian innovations churned earnings quintiles far too fast

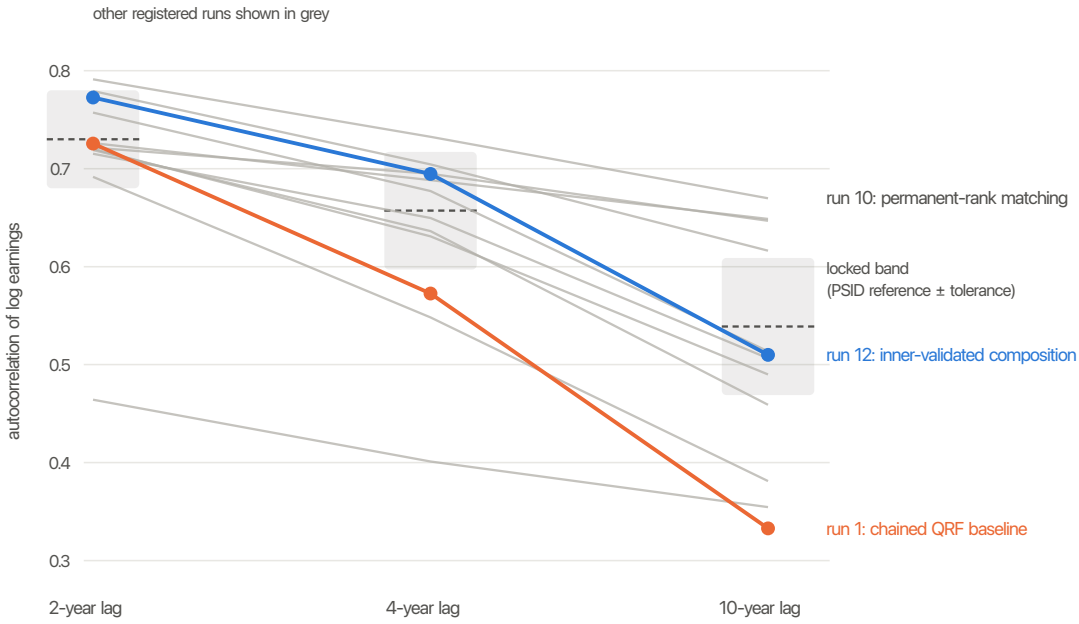

Figure 1: The autocorrelation ladder across the twelve distinct registered model runs. Shaded bands are the locked tolerances around the committed PSID reference at each horizon; run 13 re-registered the run-12 specification and reproduced it bit-exactly, so its curve coincides with run 12's. Every value is computed from the committed run artifacts at figure-build time (`scripts/build_paper_figures.py`).

and the classifier read the synthetic joint easily. Replacing the Gaussian law with an empirical rank-transition kernel fixed the mobility matrix and halved the classifier gap while its one-step memory collapsed the ladder's tail; adding two-step-plus-anchor memory through nearest-neighbor conditional draws brought the strongest scorecard of the sequence — the pairs-view classifier under its threshold on all five seeds, two seeds clearing the entire gate — with the window-3 classifier and the ten-year rung each missing on three seeds by thousandths. A final variant that matched donors on their estimated permanent rank rather than their observed anchor converted anchor noise into permanent signal and overshot every rung: the tenth registration, and the second time the program measured the same lesson from the opposite side.

7.3 How existing models solve the joint problem

Every serious earnings-history model confronts the same triple — the cross-sectional marginal, person-level persistence, and measurement noise — and the field has four working answers, each buying one thing at a stated price.

Reuse observed careers. MINT splices segments of donor workers' administrative earnings records onto targets matched on demographics and recent earnings, adding a person-specific fixed effect in its regression projections (Social Security Administration 2024). Reuse dissolves the joint problem rather than solving it: the marginal is exact because the values are real, and person-level dependence is exact because whole careers come from one person. The price is donor support — no pattern appears that no donor exhibited, cohort drift needs ad hoc adjustment, and conditioning is only as rich as the match cells.

Generate parametrically, then align. DYNASIM carries an individual-specific error term in its

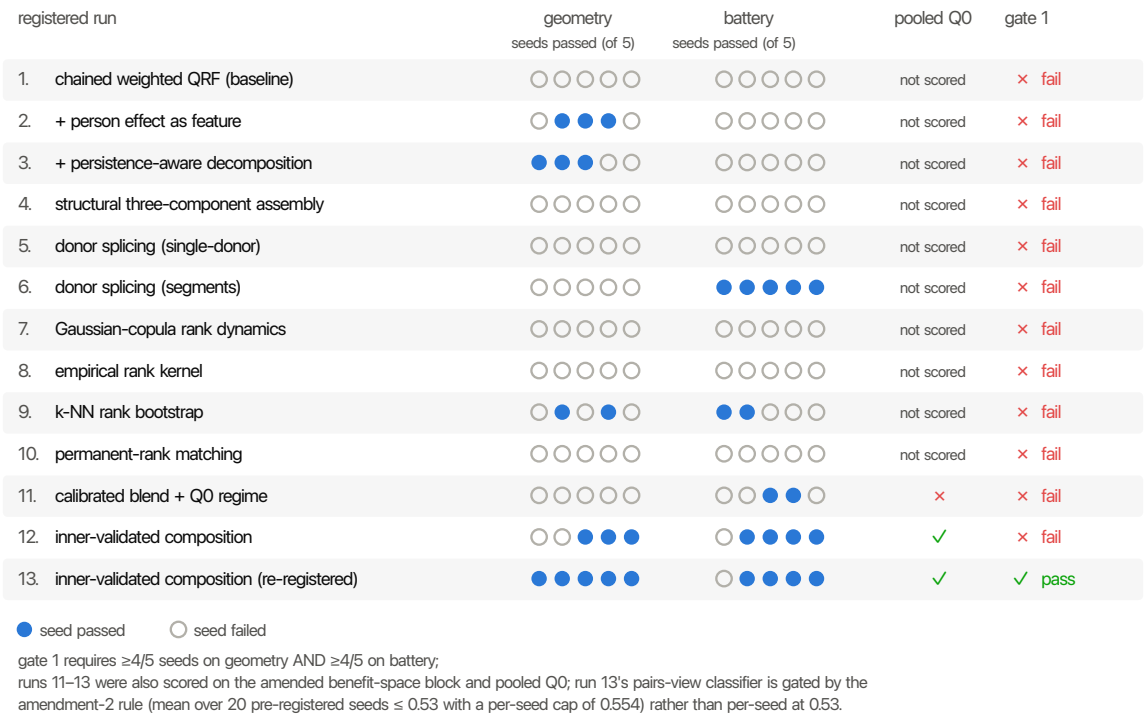

Figure 2: Per-seed gate conjunction for the thirteen registered runs, from the committed run artifacts. Each cell shows the five locked holdout seeds; gate 1 requires at least four of five on geometry and on battery jointly. Runs 11–13 were additionally scored on the amended benefit-space block and the pooled zero-anchor gate, and run 13's pairs-view classifier is gated by the ratified amendment-2 rule — the mean over twenty pre-registered seeds at the unchanged 0.53 line with a 0.554 per-seed cap — rather than per-seed.

earnings equations (Favreault et al. 2015; Urban Institute 2024); CBOLT decomposes each worker's earnings into a permanent shock — the long-run gap from the group mean — and a transitory shock, estimated on administrative panel records (Congressional Budget Office 2018). Both then align: simulated outputs are recalibrated, typically rank-preservingly, to external distributional and aggregate targets (Li and O'Donoghue 2013). Alignment is the field's standing admission that generated marginals drift; it re-imposes them after the fact and guarantees the published tables. The documented cost is that alignment can distort the relationships between variables and hits its targets whether or not the underlying process is right (Li and O'Donoghue 2013) — the class of silent correction this project's gate exists to expose rather than absorb: alignment would have masked the fourth candidate's marginal failure entirely.

Fit the simulator, not the one-step step. The administrative-data literature estimates rich nonlinear processes by simulated method of moments — parameters chosen so that simulated trajectories reproduce the target moments jointly (Guvenen et al. 2021). This answers the composition problem directly: one-step conditionals estimated from noisy data need not compose into correct long-horizon dynamics, so the generator is scored as a whole, at the horizon that matters. The price is a parametric process and total dependence on the chosen moment set.

Put the dynamics in rank space. Nonlinear panel frameworks place persistence on a latent quantile position and read earnings off the empirical quantile function (Arellano et al. 2017). The marginal becomes untouchable by construction and all modeling effort concentrates on the rank process, where the permanent–transitory structure lives — at the price of heavier estimation machinery than any off-the-shelf learner supplies.

The fifth run tested the first strategy as a benchmark, in its simplest registered form: one donor per target, whole careers spliced by age with a level adjustment at the anchor. It failed — informatively. The marginal came through exactly as reuse promises (every distributional and tail metric passed comfortably), but the omnibus classifier still separated spliced from held-out trajectories on every seed, and the battery found whole-career reuse too persistent: mobility, exit rates, and zero-spell lengths all sit outside their bands, and the autocorrelation ladder overshoots at every horizon. The run's own diagnostics name the artifacts — donor and target cohorts occupy offset points of the biennial age grid, so half the spliced values come from an adjacent age; scaling a whole career to a noisy anchor observation converts that noise into a permanent component; and one donor per career is strictly more persistent than the segment splicing MINT actually performs. The result bounds naive splicing rather than refuting the strategy. The registered segment variant — three-period segments from multiple donors, spliced by calendar period and level-adjusted at each segment boundary — then became the first candidate to clear the entire battery: all five seeds pass every dynamics tolerance, with the autocorrelation ladder, mobility, exit rates, and spell lengths inside their bands at once. What remains is a single metric family: the omnibus classifier still separates spliced from held-out trajectories by 0.017 and 0.027 above its two thresholds, with every other geometry metric passing on every seed. Six runs in, the gate has narrowed from "the dynamics are wrong" to "a classifier can still tell," which is the precise question the generative track now has to answer. The generative track's next design composes the third and fourth strategies with the project's existing machinery: empirical quantile marginals, persistent dynamics in rank space, and innovation dispersion calibrated so iterated — not one-step — dynamics match the data's ladder (Arellano et al. 2017; Guvenen et al. 2021).

Two auxiliary results harden the target. Excluding every PSID observation carrying an income-assignment flag (8.7% of positive-earnings person-periods overall, rising to 14–18% in the 2020 and 2022 waves) moves the 10-year reference by +0.012 — survey imputation does not explain the gap the candidates must close. And the statutory-resolution surface of Section 5 produced its first artifact: the project's Python benefit oracle and the Axiom rules engine compute identical primary insurance amounts — 240 of 240 synthetic careers exact to the cent, both code paths pinned by revision — after the cross-engine comparison surfaced and fixed a statutory off-by-one in the elapsed-year count and a

rounding-direction error in a test fixture that no single-engine test had caught.

7.4 Recalibrating the gate toward decision relevance

Nine runs in, the maintainer asked the governance question the protocol exists to make askable: is the bar too high? The program answered with analyses rather than judgment. Classifier forensics showed the two best candidates' identical residuals were distinct defects, not a shared wall. A downstream-relevance analysis then pushed generated and real careers through the project's own statutory benefit calculator: the best candidate's benefit distributions were statistically indistinguishable from real ones — the distributional distance inside the real-versus-real noise floor, central gaps under two percent against a five-percent criterion — while the same analysis exposed a defect no locked metric had isolated: careers generated for people observed with zero earnings at their anchor overstated their benefits by nine percent, exactly the population a progressive benefit formula weights most. The locked gate was strict on an axis that does not matter for benefits and silent on one that does.

The response was the contract's own amendment mechanism, exercised in full for the first time: a proposal committed as an inert object changing nothing; a fresh adversarial referee round (which caught a misattributed criterion citation and forced disclosure that the proposed demotion flips four historical geometry verdicts — none of which changes an overall outcome); fixes; a verification pass; maintainer ratification by merge; and a follow-up flipping the ratified content into the locked block. The amended gate demotes the benefit-immaterial window-3 classifier to reported status and adds a gated benefit-space block — distributional and decile bands at the five-percent criterion, a distance bound derived from a committed real-versus-real anchor, and a pooled band on the zero-anchor subgroup where reality measures under three percent and the best candidate nine. Recalibration added strictness where the evidence says it matters: the best candidate still fails the amended gate, and reality still passes it.

7.5 Nested validation and the forecast discipline

Two practices matured in the runs that followed the amendment. The eleventh candidate composed the two best-understood mechanisms and failed through two couplings its registration had not foreseen — including a swing of the zero-anchor benefit gap from +9 to −18 percent through an interaction between its memory coordinate and the very subgroup it was fixing. The response was methodological rather than another guess: the one-shot rule protects the outer holdout, so model development may iterate freely on inner splits carved from each seed's training complement. An inner-validation harness now mirrors the amended gate at inner scale, and a design sweep raced the candidate mechanisms head to head on it — establishing with no outer-holdout contact that the zero-anchor participation refit alone closes the benefit gap, that the drawn persistent-state coordinate fixes the dynamics battery completely while wrecking the joint completely, and that one composition stood near-clearing everything at once. The twelfth candidate froze that composition, every constant selected by nested validation.

Each of the last two runs also carried a pre-registered forecast — component probabilities logged on the public registry before the run, graded against the outcome after. The twelfth run graded almost exactly: forecast 0.42 with modal failure named as the pairs-view classifier one seed short, and that is what happened — the battery passed (the ten-year rung in-band on every seed for the first time), the zero-anchor gap closed to +0.04 percent, the per-seed benefit metrics passed everywhere, and two seeds clipped the pairs-view classifier by 0.0015 and 0.0030. Twelve registered runs in, the program's entire remaining distance to its own gate is a classifier residual of thousandths on one view.

A twenty-seed extension of that measurement — reported, not gated — placed the residual against the matched real-vs-real floor (Figure 3). The twenty-seed mean sits 0.0066 below the locked 0.53 line; the five locked gate seeds alone average 0.0021 below it, and the two seeds that clip the line are the two largest candidate scores across all twenty. Whether that pattern is seed noise or a seed-set property

became the second amendment proposal, and the adversarial referee round earned its keep: the referee found the locked seeds non-exchangeable with the fresh ones in exactly the damaging direction (a random five-subset has a mean that high with probability 0.009), found the proposal's headline margin describing the twenty-seed mean where the gate scored the locked-five mean — one standard error below the line, not five — and named the self-rescue: unlike the first amendment, whose trigger still failed after ratification, this one's triggering candidate would have flipped to a pass.

7.6 The first pass

The reworked amendment survived a verification round and was ratified as an estimator change that refuses its own trigger. The pairs-view classifier now gates on the mean over twenty pre-registered seeds at the unchanged 0.53 line — an operating characteristic *stricter* above the line than the per-seed rule it replaces — with a per-seed catastrophe cap re-derived from a classifier-version-matched floor, and two standing rules: no candidate's committed verdict changes under a rule proposed after its own run, and floor derivation and candidate scoring must share a classifier version. Run 12's verdict stands as a fail permanently.

The thirteenth registered run was therefore a fresh registration of the identical specification, with the protocol's plainest disclosure yet: its pairs-classifier outcome was already public before the run, so the registered forecast was 0.97 with the residual entirely on execution error. The run reproduced every committed baseline bit-exactly — all twenty pairs scores, the locked-seed battery, benefit, and geometry blocks, to the last digit — and passed every block of the amended gate: geometry five of five, battery four of five, pooled zero-anchor gap +0.04 percent, twenty-seed classifier mean 0.5234 against 0.53 with a maximum seed at 0.5330 against the 0.554 cap (Figure 2). The first pass of the program arrived, in other words, not from a better model but from a better-measured gate — and the record distinguishes those two things explicitly, which is the point of keeping one.

A registered reform-delta diagnostic followed the pass (reported, not gated): two opposite-incidence mechanical reforms — the first PIA factor raised from 90 to 95 percent, and the taxable maximum removed from the AIME step — computed on real versus generated histories under the locked holdout protocol, each gap measured against a real-vs-real half-split floor. The aggregates a reform score leads with land inside the floor on both reforms: mean benefit change (gap $0.29 against a $0.48 floor bottom-loaded; $4.02 against $7.77 top-loaded), winners share, and the zero-anchor subgroup. The fine incidence curve is not fully reproduced: the bottom-loaded reform misses only the concave first-bend decile (1.9 times its floor, about one percent of the flat $51 monthly delta), while cap removal shifts roughly a quarter of the top compared decile's gains downward — real +$75.76 a month at the ninth decile against generated +$55.33 — with six of seven compared deciles individually outside their floors. The named mechanism is that the generator under-concentrates cap-riding careers, diluting the extreme top of the earnings distribution, consistent with the run-12 microtexture forensics; it is the program's next target, found by the protocol's own diagnostics rather than by a downstream user.

7.7 What failure buys

Twenty-nine registered runs across two gates — thirteen at gate 1, sixteen at gate 2 — twenty-seven published failures against two passes under a rule that refused to rescue its own trigger, four gates locked through their own adversarial rounds with four ratified amendments among them — two at gate 1, and at gate 2 an estimator aligned to its operating characteristic and then a tranche structure made explicit, the household-composition and marriage-by-earnings tranches both locked as-is without candidate ladders yet — four registered forensics rounds and a reform-delta diagnostic feeding the next registration, a cross-anchor cost-ordering synthesis and a same-frame revenue pseudo-projection that between them isolated the compression the representative-frame transport must remove, six external-anchor replications reported at the same standing as any gated run, and twenty-five forecasts registered and all twenty-five graded is the protocol working as designed. Each failure published with

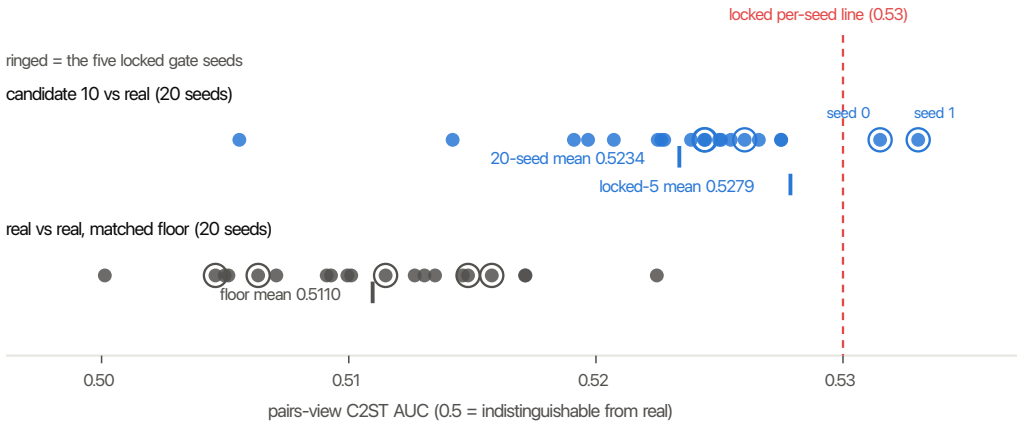

Figure 3: The twenty-seed extension of the run-12 pairs-view classifier measurement (reported, not gated), from the committed diagnostics artifact. Blue: candidate 10 scored against real holdouts; grey: the matched real-vs-real floor. Ringed points are the five locked gate seeds; both the twenty-seed and locked-five candidate means are marked.

the same prominence as a pass would, each narrowed the design space with a finding that transfers beyond this project, and none required trusting the authors: the registrations predate the runs, the artifacts recompute, and the thresholds cannot drift to accommodate a result — when the gate itself was recalibrated, it moved only through a public, refereed, ratified amendment whose every verdict change is disclosed in the contract's own history. The contrast with validation by fidelity — where a model meets the tables it was fit to and the first hard test arrives after adoption — is the point.

8 The demographic layer and the replication anchors

The tally above compresses several developments the same protocol governed once gate 1 produced its first pass; this section details the first two and the sections that follow the rest: a second stage gate, opened and locked on the family-transition layer, which its sixteenth registered candidate passed on the marital-transition and fertility tranche, and the reform-scoring surface, tested against external anchors on real microdata. Neither program is closed — the gate's household-composition tranche and forward projection remain — and both are on the public record with the discipline the earnings-history gate established.

8.1 The second gate's lock ceremony

Gate 2 — family and benefit outputs — governs the demographic transitions survivor, spousal, and caregiver reforms score on: first-marriage, divorce, widowhood, and remarriage hazards, cohort nuptiality, fertility, and the dissolved-state stock shares eligibility rides on. It locked on 2026-07-08, through the ceremony gate 1 established: draft floors, an adversarial round, fixes, verification, and ratification by merge.

The draft built the reference moments and a person-disjoint half-split noise floor on five split seeds, with each per-cell tolerance the floor mean plus four standard deviations across forty gate-eligible cells. The adversarial round returned *amend before lock*, and its objection was quantitative: on five seeds the floor's standard deviation is a five-draw estimate of a half-normal, so the committed tolerances realized anywhere from 0.9 to 5.6 times their own measured noise, and a faithful candidate — one drawing from the reference process itself — cleared the four-of-five gate with probability 0.023. The

referee also showed that the draft's floor scale, pass statistic, and seed rule described three different experiments, and that a verbatim copy of a training half passed at the noise floor, which no moment gate can prevent — the memorization defense is procedural (registration, holdout exclusion, and the no-self-rescue rule), not a property of the cell set.

The fixes rebuilt the floor on one hundred split seeds, stabilizing the estimator to roughly 3.2 times each cell's own noise; added a power cap that gates a cell only when its stabilized tolerance is at most $\ln(1.5)$ — a 1.5-times rate error — demoting under-powered per-age cells to report-only and recovering their coverage through pre-registered aggregates; and added the sequence and stock statistics a banded-marginal gate misses: origin-split remarriage, cohort ever-married-by-40, and dissolved-state stock shares by age and sex. Under the same rule on the rebuilt floor, the faithful-candidate operating characteristic is 0.969 over the resulting forty-six gated cells. Verification confirmed the amendments, and maintainer ratification by merge was the lock event. Every figure recomputes from the committed floor artifact (`runs/gate2_floors_v2.json`).

The gate's external anchors are shape reports, not level gates, and the ceremony made the reason explicit. The raw recent PSID marriage and divorce rates run about 2.4 times the national vital-statistics rates — but the PSID counts persons transitioning where the vital-statistics series counts couples (a factor of two exactly) against a person-year denominator for the age-15-plus population (a further 1.22). Attributing those two concept factors leaves residuals of 1.036 for marriage and 0.989 for divorce: the anchor is a near-bullseye once the population concepts are aligned, and the period-matched fertility series sits at a median PSID/NCHS ratio of 0.93.

8.2 The gate-2 candidate ladder

Sixteen candidates have run against the locked family-transition gate, each registered on the campaign registry (issue #42) before its single scored run and graded against a pre-registered forecast after. The ladder narrows the failure from fifteen distinct cells to a single pass, and the record of what each step cost is the evidence the gate exists to produce. The base composition — stratified empirical hazards with a mortality-composed widowhood component — fails all five seeds across fifteen distinct cells. Adding an age-by-sex first-marriage interaction and a decade-period widowhood mortality makes the composed widowhood backfire, widening the failure to nineteen cells. Pooled-rate shrinkage un-explodes that cascade back to eight; a parametric per-sex mortality trend then fixes the 65–74 female widowed stock but lets the residual go diffuse. Replacing the trend's source — PSID's own male mortality slope was unstable — with an external NCHS life-table trend, and adding single-year-kernel fertility and an origin-split remarriage table, narrows the failure to six cells. The sharpest fix is the sixth: source-aligning the spouse-death level to the surviving spouse's own marriage-history widowhood incidence removes a sex-asymmetric wedge in the earlier death-record level — which understated female widowhood by 1.9 to 2.8 times while overstating male — lifting the 75-plus female widow stock from one seed to four and clearing one full seed. Its residuals are the untargeted male lifetime-marriage sequence cell, two seeds over its 0.047 tolerance; a three-cell miss on a third seed; and the fertility clip inherited from candidate 5 on a fourth — the first candidate to clear a seed.

Candidates 7 through 9, still under the gate's original single-draw estimator, worked the two chronic marriage-count cells. Order-conditioned remarriage (candidate 7) fired the registered change-what-works risk — seed 0 regressed from forty-six passing cells to forty-four — and fixed the sign, since the generator under-produces lifetime marriages; an aggregate-preserving order split (candidate 8) isolated the deficit as compositional rather than a remarriage-level error, reconciled to a zero residual; and observed undatable-marriage initial states (candidate 9) cured the male count while the observed residual overshot the female one. By candidate 9 every remaining failing cell's twenty-draw-mean tilt measured sub-tolerance while single draws still decided verdicts: the binding constraint had become the gate's estimator, not the model, and the ladder paused for a refereed amendment.

The seven candidates that followed ran under the amended mean-over-draws estimator against one

named level target — the elderly-widow stock — that four forensics rounds resolved in turn. Candidate 10 confirmed the stock cell as an unambiguous level under-production rather than absorbed draw noise; candidate 11's elderly-remarriage split exposed two compensating errors in the pooled band; candidate 12 cleared the cell on all five seeds with observed already-widowed initial states, while falsifying a co-registered spousal-gap-draw delta on the run's own inertness test; candidates 13 and 14 re-banded the young and the oldest surviving-spouse widowhood, trading the stock against aging-in; and candidate 15 removed the deployment-time NCHS mortality trend the gate's untrended reference does not carry, reaching three of five seeds — the ladder's best short of a pass. Candidate 16 conditioned the widowhood hazard on one observed covariate and passed.

Table 2: The gate-2 candidate ladder, from the committed run artifacts (`runs/gate2_hazard_v1.json` through `v16.json`). "Distinct failing cells" counts gated cells missing on at least one of the five locked seeds; the gate requires at least four of five seeds with every one of the forty-six gated cells inside its locked tolerance, which candidate 16 is the first to meet. Candidates 10–16 are scored under the amendment-1 mean-over-draws estimator; candidates 1–9's committed verdicts stand.

candidate	distinct failing cells	seeds passing	headline lesson
1 — stratified hazards, composed widowhood	15	0 / 5	the base composition; young first-marriage and female widowed-stock cells drift
2 — + age×sex first marriage, period widowhood mortality	19	0 / 5	the composed widowhood backfires — the widest failure
3 — + pooled-rate mortality shrinkage	8	0 / 5	shrinkage un-explodes the cascade; the widowed stock persists
4 — + parametric per-sex mortality trend	8	0 / 5	fixes the 65–74 female widowed stock; the residual goes diffuse
5 — + external NCHS mortality trend, kernel fertility	6	0 / 5	an external trend replaces PSID's unstable male slope
6 — + source-aligned spouse-death level	5	1 / 5	removes the sex-asymmetric widowhood wedge (1.9–2.8×); first seed cleared
7 — + marriage-order remarriage, marital-status fertility	7	0 / 5	the change-what-works risk fires — seed 0 regresses 46→44; the generator under-produces lifetime marriages
8 — + order-split remarriage, aggregate counts preserved	6	0 / 5	the marriage-count deficit is compositional, not remarriage-level (reconciled to zero residual)

candidate	distinct failing cells	seeds passing	headline lesson
9 — + observed undatable-marriage initial state, low-parity fertility	7	0 / 5	the count fix cures males but overshoots females; sub-tolerance tilts expose the single-draw estimator as binding
10 — first run under the amended estimator; + age-band remarriage	4	1 / 5	under the mean estimator the 75+ widow stock is a level miss, not draw noise
11 — + 50–64 / 65–74 / 75+ remarriage split	4	1 / 5	the pooled elderly band hid two compensating errors; splitting trades outflow for aging-in
12 — + entry-widowed initial states, age-conditioned gap draws	3	2 / 5	observed already-widowed states clear the 75+ stock 5/5; the gap-draw delta is proven inert
13 — + young surviving-spouse widowhood bands (18–34, 35–44)	2	2 / 5	removes an order-of-magnitude young-widowhood rate error; both counts clear, but less aging-in lowers the stock
14 — + split the 75+ widowhood band (75–84, 85+)	2	2 / 5	re-banding recovers incidence (0.93→0.95) but a fixed aggregate cannot lift the stock
15 — − NCHS mortality trend inside the gate	2	3 / 5	the gate's reference is the untrended panel; removal reaches three seeds but leaves a survival-to-75 stock leak
16 — + widowhood support-composition stratum	1	4 / 5	**PASS** — conditioning on the observed support window closes the yield leak; the sole miss is an RNG-isolated fertility split artifact

8.3 The estimator amendment

Candidate 9's grading named a constraint the model could not move. Every failing cell's mean over twenty pre-registered simulation draws sat inside its tolerance, yet single draws — one frozen replicate per seed — were deciding the verdicts. The gate had ratified its operating characteristic on a draw-noise-free basis: a faithful candidate — one drawing from the reference process itself — modeled

per cell as a half-normal with the floor's own standard deviation and no simulation-draw term, which gives a per-seed pass probability of 0.9404 and a four-of-five gate pass of 0.9685. But the single-draw estimator the gate shipped with injected a per-cell draw-noise term the tolerances never budgeted for, dropping that same faithful candidate to roughly 0.885 and 0.896 on the six measured cells: the pass probability the gate was ratified to deliver was not merely hard but unachievable under its own estimator.

The fix — proposed as an inert object, carried through an adversarial round and verification, and flipped live in a follow-up merge, the ceremony gate 1's second amendment established — scores each cell on the mean cell rate over twenty pre-registered draws rather than one, leaving the tolerances and the forty-six-cell four-of-five conjunction untouched. Averaging over twenty draws shrinks the injected noise toward the draw-noise-free rate the tolerance was measured against, restoring the faithful candidate to roughly 0.939 and 0.967 — back to the numbers the gate was locked with, now achievable because the estimator finally shares their derivation basis. It is an estimator aligned to its own operating characteristic, not a loosened error budget: the four noise-dominated cells collapse toward zero clip probability while the one systematically mistuned marriage-count cell fails harder under the mean, not softer, and a candidate with gross level errors still fails by orders of magnitude at any number of draws. The amendment is prospective only — candidates 1 through 9 stand as committed failures under the no-self-rescue rule, and candidate 10 was registered as the first fresh run under it. The operating characteristic recomputes from the committed floor artifact (`runs/gate2_floors_v2.json`), and the record keeps the better-measured gate and the better model distinct exactly as gate 1 did.

8.4 Four forensics rounds

Between the scored runs the ladder registered four diagnostic rounds on the campaign registry, each reported, never gated, published whatever it found, and cited by the candidate it licensed. They are the method's throughline: candidates 9, 11, 12, and 16 — the pass among them — each registered only after, and citing, the round that preceded it.

The first round decomposed the two chronic marriage-count cells on the training halves and found the male deficit was not a rate error but a measurement-concept residual — the reference carries lifetime-marriage counts from episodes with undatable start or dissolution years that no hazard model can generate, so the fix is an observed initial state, not a dial. Its stability sub-question, twenty redraws of the same specification, measured every remaining clip's mean tilt sub-tolerance, the evidence that motivated the estimator amendment. The second round decomposed the elderly-widow stock gap and the female count residual and found one mechanism behind both: a pooled fifty-plus current-age remarriage band applying a rate roughly nine times too high to 75-plus widows, depleting the stock from the outflow side while carrying the female count over-production. The third round audited how the reference constructs the 75-plus widowed stock and found that 12.1 percent of it is carried from spouse deaths predating the person's panel support — structurally unreachable by any transition rate, and entirely fixable by an observed initial state — while the simulated young-widowed pool ran to 3.16 times reference at ages 15 to 49, fed by the youngest surviving-spouse band edge. The fourth round split its two remaining cells: seed 2's fertility clip decomposed as a split artifact — a systematic deficit comfortably inside tolerance, wrapped in a maximum-of-five reference draw and a minimum-of-five simulation draw, isolated from the widowhood delta's random stream and so failing regardless, which fixed the pass path at seeds 0, 1, 3, and 4 — and the elderly-stock leak decomposed as a survival-to-75 yield gap: real widowhoods correlate with long observed support, which a uniform within-band hazard cannot see, so simulated 50-to-64-onset widowhoods reached age-75 windows at 39 percent against the reference's 57 percent. Candidate 16 conditioned on exactly that window.

8.5 The first gate-2 pass

Candidate 16 added one covariate to the surviving-spouse widowhood hazard: whether a person's observed support window reaches age 75, a binary stratum known in advance from the panel attributes alone — the same observed-data class as the initial-state fixes — true of 3,147 of 41,409 persons, a 12.7 percent exposure-weighted share. Within each age band and sex the two strata are train-estimated and recombine to candidate 15's band aggregate by an exposure-weighted identity, so aggregate widowhood incidence is preserved (recombination residual 1.7e-18, reconciled to zero; 75-plus incidence moves from 1.060 to 1.061 of reference) while the event composition matches the reference's window correlation, and all four gated widowhood-incidence cells hold on every seed.

The delta closed the yield leak the fourth forensics round had sized. The 50-to-64-onset survival-to-75 yield rose from 0.581 to 0.816 of reference as the share of those widowhoods whose window reaches 75 moved from 0.391 to 0.551 toward the reference's 0.572, lifting the 75-plus female widowed stock from 0.841 to 0.914 of reference. The chronic stock cell cleared all five seeds — scores 0.057 to 0.111 against a 0.185 tolerance, with the two seeds candidate 15 had failed both flipping — and the marriage counts the recomposed exposure put at risk held on both sexes, minimum margins plus 0.008 and plus 0.011. The gate passed four of five.

The scope is on the record with the pass. It is a model delta under a locked gate — thresholds, protocol, and the ratified estimator amendment all fixed before the run was registered — not an amendment rescuing its own trigger; candidates 1 through 15 stand as committed failures. It certifies the marital-transition and fertility tranche the locked cells score, and no more: the household-composition tranche and the marriage-by-earnings joint the gate's own scope note names remain unscored, and the sole failing cell anywhere in the passing run is seed 2's completed-fertility clip — the split artifact the fourth forensics round diagnosed, byte-identical to candidate 15 and isolated by construction from the delta that carried the gate.

8.6 The replication anchors

While the generative track worked the demographic gate, the reform-scoring surface — the statutory AIME/PIA chain and the survivor, spousal, and caregiver benefit plumbing, run on real PSID careers — was tested against six external DYNASIM anchors. Each is reported, not gated: registered on the campaign registry before the run and published regardless of outcome. Four score real careers directly; the price-indexing anchor additionally routes generated careers through the same calculator. A sixth, added once the gate-2a pass validated the marital histories it needs, recomputes Mermin's exact shared-earnings quintile concept on real couples and closes the own-versus-shared concept delta the price-indexing anchor had carried (Section 9).

Table 3: The six external-anchor replications, reported not gated, from the committed diagnostic artifacts; each registered on issue #42 before its run. The shared-earnings row (Section 9) was run after the gate-2a pass validated the marital histories it requires.

provision (anchor)	achieved replication	artifact / registration
progressive price indexing — Mermin (2005)	the price-indexing scalar lands at 66.73% of scheduled against DYNASIM's 67.8%; the progressive-price-indexing quintile pattern reproduces — real and generated careers decline together from 100% to 84% of scheduled across quintiles, matching DYNASIM's monotone gradient — with real-versus-generated gaps inside the floor at four of five quintiles	`replication_ppi_mermin_v1.json`; #42 c4907444903
normal-retirement-age increase to 70 — Mermin (2005)	79.75% of scheduled overall, cross-quintile spread 0.045pp, each quintile within 0.03–0.33pp of Mermin's 79.4–79.9% row	`replication_mermin_rows_v1.json`; #42 c4911609804
COLA reduced 0.4pp — Mermin (2005)	99.2% of scheduled at ages 62–67 (anchor 98.9%) and 93.0% at 80–85 (anchor 92.4%)	`replication_mermin_rows_v1.json`; #42 c4911609804
earnings sharing — Favreault and Steuerle (2007)	the best-populated cell — married women, the large gainers — lands on DYNASIM: gain 20% 49 vs 44, gain 5% 63 vs 60, lose 5% 24 vs 22; all four registered directional calls hold	`replication_r7_sharing_v1.json`; #42 c4911171806
caregiver credit — Smith et al. (2020)	all four candidate plans place 58–71% of aggregate gains in the bottom lifetime-earnings quintile, in or above the anchor's 52–62% band; the concentration and reach rankings match (Spearman 0.8)	`replication_caregiver_v1.json`; #42 c4911453454

provision (anchor)	achieved replication	artifact / registration
shared-earnings progressive price indexing — Mermin (2005)	Mermin's exact shared-lifetime-earnings quintile concept on 3,131 real couples (own-record benefits, shared ranking): shared Q1 lands at 99.49% of scheduled against the anchor's 98.7 where the own-record analogue read 100.00, Q1–Q3 each move closer to the anchor, the distribution stays monotone (99.49 → 86.62), and the price-indexing wedge is flat at 66.73% against 67.8; 45.8% of the persons change quintile between the two rankings; all four registered calls hold	`replication_ppi_shared_v1.json`; #42 c4931009783

The discipline that makes those comparisons legible is that every level difference from the DYNASIM projections is named, not absorbed. The populations differ — observed PSID retirees eligible 2005–2019, born 1943–57, against DYNASIM's projected 1960–80 cohorts evaluated in 2049 — as do the windows (a biennial PIA-proxy convention, the national average wage projected forward to each indexing year) and the earnings concept (the earnings-sharing exercise scores real couples with both spouses computable, a shared-versus-individual distinction the incidence turns on). Where the older PSID cohorts hold more single-earner couples, married men lose harder than DYNASIM's 2049 projection, and the direction, not the level, is what the test certifies. The same discipline caught an error in an anchor: Favreault and Steuerle's Table 3 carries a package-1c married-women column summing to 109.1, a 9.1 in a "no change" cell where every comparable earnings-sharing cell reads 0.0 — a printed typo, flagged and carried verbatim, feeding no scored result.

Two things are true at once, and the record keeps them distinct. Distributional scoring on validated real microdata is anchor-grade across the cataloged provision classes — price indexing, retirement-age and cost-of-living reductions, earnings sharing, and caregiver credits — each reproducing its DYNASIM incidence pattern within the noise floor or a named delta. The generated-population demographic track reached its own gate at the sixteenth candidate, on the marital-transition and fertility tranche, and the tranches it does not yet score — household composition and the marriage-by-earnings joint — stay named and unscored rather than assumed. The paper reports all of it because the protocol scores all of it, and a reader who wants to know which claims are load-bearing today can read it off the record rather than the prose.

9 The reform-scoring surface, tested against itself

The gate-2a pass validated the marital histories the anchor replications need, and three registered diagnostics followed — each reported, not gated, each graded against a forecast filed before the run. One closed the concept delta the price-indexing anchor had carried; the other two scored the committed encodings against each other and added the revenue side the projection roadmap requires, and both surfaced the same distortion a representative-frame transport exists to remove.

9.1 The shared-earnings concept, closed

Section 8's progressive-price-indexing anchor reproduced Mermin's quintile gradient in shape but not in grouping: Mermin (2005) ranks retired workers by *shared* lifetime earnings — own earnings when single, half the couple's when married — and the Phase-A run could only rank by own record. With the gate-2a marital histories in hand the exact concept becomes computable on real couples, and the sixth anchor in Table 3 runs it, holding the benefit reform at the Phase-A encoding verbatim and changing only the ranking variable. On the 3,131 couples with both spouses computable, 45.8 percent of persons change quintile between the own and shared rankings, and the regrouping moves the bottom onto the anchor: shared Q1 lands at 99.49 percent of scheduled against Mermin's 98.7, where the own-record analogue read a flat 100.00, and Q1 through Q3 each sit closer to the anchor than the own-record version. The distribution stays monotone and the price-indexing wedge is flat at 66.73 percent against Mermin's flat 67.8. The one carried delta is stated rather than chased: the upper quintiles move slightly *away* from the anchor under shared ranking — the arithmetic complement of pulling the low-earning spouses of high earners up out of the top groups — and the top stays compressed against Mermin's 71.7 by the truncated observation window. All four registered calls held; the own-versus-shared concept delta the earlier anchor named is closed on real data, and the generated-couples version of the same concept waits on the marriage-by-earnings tranche.

9.2 Cost ordering and a revenue projection: two compression fingerprints

The anchor replications each scored one provision family in isolation. A cost-ordering synthesis then scored every committed encoding once on a single common frame — 1,549 sex-resolvable careers under the Phase-A 2050 transport — and tested the aggregate cost deltas *ordinally* against the anchors' published cost columns, since observed completed careers are not a Trustees projection and the levels differ by construction. The signs all agreed, eight of eight — four Mermin provisions score as savings, four caregiver plans as costs — and the caregiver ordering matched the anchor at rank correlation 0.913, its one inversion a pair the anchor itself ties. But the Mermin quartet's cost ordering broke on a single adjacent swap: the frame ranks the retirement-age increase above progressive price indexing, where DYNASIM ranks them the other way. The mechanism is the compressed support the earlier anchors already named — on these observed careers most indexed earnings sit below the thirtieth-percentile bend that progressive price indexing protects, so it bites lightly (−11.0 percent) while the uniform retirement-age reduction cuts across the board (−20.2 percent). The forecast of a perfect ordering failed, and the miss is a diagnosis: progressive price indexing is exactly the provision whose relative magnitude depends on getting the earnings distribution right.

The same distortion reappeared on the revenue side. The first roadmap milestone added a taxable-payroll aggregation, a present-value balance analogue, and an endogenous trust-fund-exhaustion ledger to that common frame, calibrated so the baseline exhausts in the 2034 year Smith (2015) reads from its own DYNASIM run, and scored the five solvency provisions Smith projects — levels frame-relative and ungraded, signs and orderings the content. Every sign held, fourteen of fourteen, and a registered *disagreement* held in its stated direction: raising the full retirement age to 72 exhausts Smith's fund by under a year because his increase phases in against a fixed horizon, but on completed careers with no phase-in the cut never lets the calibrated fund exhaust at all, so it ranks first where Smith ranks it last. The revenue ordering then broke on one adjacent swap, as the benefit ordering had — the frame puts a two-point payroll-rate rise above full removal of the taxable maximum, where Smith has removal first, because only about 12.7 percent of the frame's taxable payroll sits above the wage base, under the 16.1 percent break-even where removal's revenue would overtake the rate rise and under the roughly 17–18 percent the administrative data carries. It is the taxable-maximum analogue of the bend-point swap: the same truncated-career compression, once at the benefit bends and once at the contribution ceiling. The record commits both as before-and-after test cases for the representative-frame transport, the milestone whose whole job is to make the frame's distribution representative enough that both orderings come out right.

10 Two more gates and the projection roadmap

Alongside the surface work, an external architecture review reset the engineering standard for deployment, a third and a fourth stage gate locked — on household composition and on the marriage-by-earnings joint — a disability module landed, and the program published the capability path from here to a full projection. None of it moved a locked threshold.

10.1 Hardening the ledger for deployment

An external architecture review of the candidate chain through its fourteenth member returned a one-line thesis — preserve the research ledger, stop extending its execution architecture, which had grown too implicit for deployment on the production population — and its ten findings dispatched into the contract's own machinery. A blocker, that gate 2 overclaimed which tranche it certified, became the ratified tranche amendment that split the family gate into its marital-transition, household-composition, and marriage-by-earnings tranches and promoted the rule binding each to its scored surface. Three priority-zero fixes landed as a legacy digest manifest with a no-overwrite guard, per-marker test tiers, and a seam-ownership record. The priority-one item became a flattened component registry that ports the passing sixteenth candidate into immutable, injected components with no runner imports, carrying a compatibility certificate that reproduces the candidate's committed 20-by-46-by-5 rate cube bit-for-bit — 4,600 of 4,600 values equal to the IEEE-754 bit, signed zeros included — and its gate-2a pass verdict. A referee certified the certificate real rather than vacuous by re-deriving the cube independently and confirming that every mutation it injected was caught.

10.2 The household-composition gate

The gate-2 pass certified the marital-transition tranche and named household composition as unscored; it now carries its own locked gate. Gate 2b — coresidence of spouses, children, parents, grandchildren, multigenerational households, and household size, read from the PSID relationship matrix — locked on 2026-07-10 through the ceremony gate 2 established, and the ceremony is the point, because the draft floor did not survive its adversarial round. The referee returned *amend before lock* on nine findings, three of them blockers, and earned the standing by reproducing the floor twice: a full rerun bit-identical to the committed artifact, and a fully independent recompute from its own parsers that matched all 81 reference moments, all 405 per-seed cell values, and all five holdout hashes to nine decimals. The three blockers were the discipline working, not failing. The draft protocol had reintroduced the single-draw estimator gate 2's own amendment had already retired, making its stated operating characteristic unachievable. The multigenerational family carried a code-frame bug that mapped first-year cohabitors to a spurious extra generation, which the referee's independent recompute showed flipped 13,400 of 91,261 multigenerational person-waves — 14.7 percent, concentrated in exactly the young cells the family gates. And the headline estimand, "waves 1969–2023," was not the scored surface: a weight rescaling leaves the first 28 waves carrying 0.19 percent of the estimand's weight, so the gate effectively certifies 1997–2023, which the standing rule that a tranche describe exactly its scored surface — promoted a day earlier by the tranche amendment above — forbids it to hide.

The eight required fixes adopted the ratified mean-over-draws estimator, corrected the cohabitor code and pinned the multigenerational concept to three distinct generations, restated the estimand as effectively 1997–2023, and — the review's own precondition for any level anchor — bundled a concept-decomposed Census anchor before the flip rather than after. That anchor carries the discipline the marriage-and-divorce anchor established one tranche earlier: the PSID family unit is not the Census household, so each ratio carries a named concept factor, and after the partner-inclusion bridge all fourteen coresident-spouse cells land between 0.75 and 1.14 of their Census counterparts, the widowhood-asymmetric 0.84 at the oldest women stated honestly below one. Verification returned *lock as-is*, the maintainer ratified by merge, and the flip that inserted the thresholds — 46 gated cells against 47 report-only, the faithful-candidate operating characteristic recomputing to 0.9397 per

seed and 0.9678 at four of five — passed its own flip-fidelity referee, who recomputed every number, re-fetched the ceremony comments, and mutated the new bindings to confirm they bite, clearing all eight verification sections before the merge. What the lock does not claim is a pass: no candidate has run against gate 2b, and its ladder is future work. The gate is built; the model has not yet cleared it.

10.3 The marriage-by-earnings gate

The gate-2 pass certified the marital-transition tranche and named the marriage-by-earnings joint — the surface spousal and survivor benefit *levels* key on — as a separate later tranche; it too now carries its own locked gate. Gate 2c — the own-by-spouse assortative-mating contingency, the earnings-conditional first-marriage and remarriage hazards, the around-event earnings dynamics, and the couple's shared-earnings distribution, all on a per-year indexed-earnings axis over the selected universe of PSID couples with a computable earnings history for both partners (7,994 directed couples over 14,952 earnings-supply persons) — locked on 2026-07-10 through the same ceremony, and once again the ceremony is the point, because the draft floor did not survive its adversarial round. The referee returned *amend before lock* on eight findings, three of them blockers, and earned the standing the household referee had, by reproducing the floor twice: a full rerun bit-identical to the committed artifact, and a fully independent recompute from its own fixed-width PSID parser, AIME chain, and splitter that matched all forty-nine reference moments, all two hundred forty-five per-seed cell values, and all five holdout hashes to nine decimals.

The three blockers were the same three classes the earlier tranches had surfaced, each caught before the lock. The first was the floor: the draft measured its noise between two person-disjoint halves, but a couple surface is not person-disjoint — under the person split half the couples straddled the two sides (50.4 percent of mirrored pairs on one seed), so the floor priced person-level noise where the tranche gates couples, and the disclosed faithful-candidate operating characteristic of 0.95/0.977 was 0.83/0.80 at the couple-level noise a faithful model actually faces. The second was the estimand: the artifact described its earnings window as 1993–2022 while the committed panel spans 1968–2022 — a twenty-five-year misstatement of its own observation surface, with 69.0 percent of the earnings-supply persons carrying a pre-1993 positive year, and a candidate faithfully implementing the *described* window failing five of thirteen gated cells on the description alone; the standing rule that a tranche describe exactly its scored surface forbids it. The third was construct validity, and it is the assortative-mating sibling of the compression fingerprints the reform-scoring surface had already turned up: the career-summed earnings proxy the draft gated correlated only 0.12 across spouses, but that number is the observation process, not earnings sorting — cross-sex tercile pooling and the anti-correlated observed-year counts of single-earner-era couples together halve it, and the proxy's 0.977 split-half reliability proves the weak signal is what the career sum measures, not sampling noise. The within-couple rank on per-year indexed earnings is 0.49, literature-scale; gated on the proxy the joint would have certified the observation mechanics and inverted the tranche's purpose, failing a candidate with true earnings sorting and passing one that merely reproduced PSID's observation cadence.

The eight required fixes rebuilt the floor couple-disjoint by connected component of the couple graph (13,275 components, none larger than four persons, zero couples straddling), restated the true 1968–2022 estimand everywhere, re-specified the earnings axis onto per-year indexed earnings, pinned the candidate's directed both-orientation couple emission (a single-orientation emission was measured non-conformant, off by up to 4.9 times), and detrended the around-event windows against a placebo drift deflator of 1.2019 — roughly three-quarters of each raw window's magnitude was generic nominal life-cycle drift any window on this panel shows, so a real-terms candidate is no longer failed by construction. Re-priced on the rebuilt floor, twenty-seven cells gate against twenty-two report-only, and the faithful-candidate operating characteristic recomputes to 0.9641 per seed and 0.988 at four of five. The review's precondition for the flip, as at the household gate, was an external anchor bundled before it rather than after: a concept-bridged assortative-mating anchor built from the CPS spouses'-earnings correlation of Schwartz (2010) and the educational assortative-mating

series of Greenwood et al. (2014), reported and never gated. Its per-year rank of 0.49 sits above Schwartz's 0.23 dual-earner annual-earnings correlation and its earnings-contingency diagonal delta of 1.37 below Greenwood's 1.6-to-2.0 educational delta — both directions the named concept deltas predict — and it moves no floor value. Verification returned *lock as-is*, the maintainer ratified by merge — the ceremony branch re-homed across three pull requests as its merge reference settled, the earlier two closed unmerged on the same floor — and the flip that inserted the thresholds passed its own flip-fidelity referee, who recomputed every tolerance from the frozen floor, re-fetched the ceremony comments, re-derived the anchor against the archived Schwartz and Greenwood sources, and confirmed the new bindings bite across all eight verification sections. What the lock does not claim, again, is a pass: no candidate has run against gate 2c, and its ladder is future work.

With that, all three tranches of the family gate are locked: the marital-transition surface passed at the sixteenth candidate, and household composition and the marriage-by-earnings joint locked but not yet cleared, their candidate ladders the open threads. Across the three the adversarial round kept returning the same three failure classes — a floor or estimator that priced the wrong noise, an estimand that misdescribed its own surface, and a construct whose signal was the measurement rather than the thing measured — and the ceremony quantified and fixed each before the threshold locked, none after. That is the design working rather than failing, and it is why a locked gate here is worth what it claims: the hard test comes before adoption, in public, against a referee who recomputes.

10.4 Disability and the capability roadmap

The disability module added the last major transition process in reported-not-calibrated form: DI incidence and recovery hazards from PSID self-reported work limitation, and the statutory conversion at full retirement age, where a disabled worker's benefit becomes a retirement benefit at the same primary amount — a factor of exactly one — so that the claiming sampler's excluded conversion mass reconstitutes the full administrative entrant mix. Its validation follows the anchor pattern: the PSID disabled-to-retired transition, read against the archived conversion column of the 2023 SSA supplement, runs at 0.267 of the administrative share for women and 0.322 for men, reported as a ratio far below one and never calibrated toward it, because seven named concept deltas separate a self-reported labor-force status from an adjudicated award — definition, insured population, severity threshold, recovery churn, conversion denominator, biennial timing, and secular period. A wanted-tables list names the exact administrative series a future level-anchored disability gate will need, pinning the gate before its components as the contract requires.

These modules sit on a published roadmap of eight capability milestones that carries the program from replicating DYNASIM's incidence patterns to a projection in its class. The layer certified or anchor-validated today — earnings, marital and fertility dynamics, mortality, claiming, and the statutory formula with its auxiliary benefits — is the first milestone; the same-frame pseudo-projection and the disability module are two more; ahead lie the household and assortative-mating tranches — both gates now locked above but not yet passed — the transport of the certified generators from their PSID estimation basis onto the representative population — the scientific crux and the funded build — a year-by-year projection engine, full revenue and trust-fund accounting, and per-year execution of the statutory rules on the projected panel. The architectural bet under all of it is a single object: one person-period panel — the cross-sectional population extended through time, with future births and immigrants entering as new persons — of which every product, from a one-year tax-benefit microsimulation to a local-area cut to the seventy-five-year projection, is a row, column, or time slice, rather than the separately funded and mutually unreconcilable models a projection of this kind has historically required.

11 Related work

Dynamic microsimulation originates with Orcutt et al. (1961); Li and O'Donoghue (2013) survey the field's alignment and calibration practice. This design borrows the central structural lesson of DYNASIM (Favreault et al. 2015; Urban Institute 2024), MINT (Social Security Administration 2024), and CBOLT (Congressional Budget Office 2018) — an annual state engine with family links and external alignment — while departing on openness, trajectory-level weighting, and scoring; Dekkers and Cumpston (2012) treat the weighting question directly. The Cato model (Chanwong 2026) is the nearest open system.

The pattern repeats internationally — Pensim2 at the United Kingdom's Department for Work and Pensions, MOSART at Statistics Norway, MIDAS at Belgium's Federal Planning Bureau (Li and O'Donoghue 2013) — with exceptions. INSEE has published the source of Destinie 2 (Blanchet et al. 2010), the pension model behind France's official projection exercises (github.com/InseeFr/Destinie-2). SimPaths, from the University of Essex's Centre for Microsimulation and Policy Analysis, is an open-source life-course model estimated for the United Kingdom and being adapted to several other European countries (Bronka et al. 2025). And open frameworks exist without open populations: LIAM2 — built at the same Federal Planning Bureau that runs MIDAS (Menten et al. 2014) — and OpenM++, an open reimplementation of Statistics Canada's Modgen platform (OpenM++ development team 2026), supply generic simulation engines but ship no calibrated population and no rules stack. On OpenM++, WIFO's microWELT models welfare transfers comparatively across Austria, Spain, Finland, and the United Kingdom, with a United States variant for labor-force projection (Spielauer et al. 2020) — one open engine carrying a multi-country dynamic model. None of these combines an open codebase with a certified calibrated microdata baseline and a public scoring protocol under which claims resolve. SimPaths is closest on openness; the difference here is inheritance — a certified cross-sectional population, one rules platform across countries, and credibility staked on resolved scores rather than publication.

On the measurement side, administrative-data studies of earnings dynamics discipline the earnings process: lifecycle moments (Guvenen et al. 2021), long-run mobility (Kopczuk et al. 2010), non-stationary volatility (Sabelhaus and Song 2010), and the attenuating link between current and lifetime earnings (Haider and Solon 2006) — with nonlinear panel frameworks (Arellano et al. 2017) the academic cousin of the quantile machinery used here. The quasi-experimental record on claiming responses (Mastrobuoni 2009; Behaghel and Blau 2012) anchors the behavioral scenario library, and the assumption-failure record compiled by successive Technical Panels (Technical Panel on Assumptions and Methods 2023) motivates the domains-of-validity framework.

12 Status and roadmap

The cross-sectional foundation is in production. Populace's charter specifies the longitudinal kernel rules. This paper is the project's front door; the supplementary design appendices — operational chapters on earnings-history construction, family and auxiliary benefits, disability and claiming, mortality and projection drift, calibration targets, the Social Security validation program, and a source-based DYNASIM dossier — live in the project repository at github.com/PolicyEngine/populace-dynamics. Development proceeds through stage gates defined as score thresholds — earnings-history credibility first, family and benefit outputs second, forward projection third, productization last — and every gate's evidence publishes whether it passes or fails.

The layer is open source under the MIT license. The project invites corrections, particularly to the benchmark characterizations, and contributions merge on the same standard as the authors': improve the held-out score.

References

Arellano, Manuel, Richard Blundell, and Stéphane Bonhomme. 2017. "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework." *Econometrica* 85 (3): 693–734.

Behaghel, Luc, and David M Blau. 2012. "Framing Social Security Reform: Behavioral Responses to Changes in the Full Retirement Age." *American Economic Journal: Economic Policy* 4 (4): 41–67.

Blanchet, Didier, Sophie Buffeteau, Emmanuelle Crenner, and Sylvie Le Minez. 2010. *The New Destinie 2 Microsimulation Model: Main Characteristics and Illustrative Results*. Document de Travail Nos. G2010-13. INSEE.

Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds. 2025. *The 2025 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds*. Social Security Administration. https://www.ssa.gov/oact/TR/2025/tr2025.pdf.

Bronka, Patryk, Justin van de Ven, Daniel Kopasker, S. Vittal Katikireddi, and Matteo Richiardi. 2025. "SimPaths: An Open-Source Microsimulation Model for Life Course Analysis." *International Journal of Microsimulation* 18 (1): 95–133. https://doi.org/10.34196/ijm.00318.

Chanwong, Krit. 2026. *Social Security Cato Model*. https://github.com/kchanwong/social_security_cato_model.

Congressional Budget Office. 2018. *An Overview of CBOLT: The Congressional Budget Office Long-Term Model*. Congressional Budget Office. https://www.cbo.gov/publication/53667.

Congressional Budget Office. 2024. *CBO's 2024 Long-Term Projections for Social Security*. Congressional Budget Office. https://www.cbo.gov/publication/60392.

Dekkers, Gijs, and Richard Cumpston. 2012. "On Weights in Dynamic-Ageing Microsimulation Models." *International Journal of Microsimulation* 5 (2): 59–65.

Favreault, Melissa M, Karen E Smith, and Richard W Johnson. 2015. *The Dynamic Simulation of Income Model (DYNASIM): An Overview*. The Urban Institute. https://www.urban.org/research/publication/dynamic-simulation-income-model-dynasim.

Favreault, Melissa M., and C. Eugene Steuerle. 2007. *Social Security Spouse and Survivor Benefits for the Modern Family*. Retirement Project Discussion Paper Nos. 07-01. The Urban Institute.

Favreault, Melissa, and Karen E Smith. 2016. *The Accuracy of MINT Wealth Projections*. Urban Institute.

Greenwood, Jeremy, Nezih Guner, Georgi Kocharkov, and Cezar Santos. 2014. "Marry Your Like: Assortative Mating and Income Inequality." *American Economic Review: Papers & Proceedings* 104 (5): 348–53.

Guvenen, Fatih, Fatih Karahan, Serdar Ozkan, and Jae Song. 2021. "What Do Data on Millions of U.S. Workers Reveal about Lifecycle Earnings Dynamics?" *Econometrica* 89 (5): 2303–39.

Haider, Steven, and Gary Solon. 2006. "Life Cycle Variation in the Association Between Current and Lifetime Earnings." *American Economic Review* 96 (4): 1308–20.

Institute on Taxation and Economic Policy. 2025. *ITEP Tax Microsimulation Model Overview*. https://itep.org/itep-tax-model.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2010. "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937." *The Quarterly Journal of Economics* 125 (1): 91–128.

Li, Jinjing, and Cathal O'Donoghue. 2013. "Alignment and Calibration of a Dynamic Microsimulation Model." *Journal of Artificial Societies and Social Simulation* 16 (3): 1–15.

Look, Spencer U., and Jack VanDerhei. 2024. *Beyond the Retirement Crisis Headlines: Why Employer-Sponsored Plans Are the Key to Retirement Adequacy for Today's Workers*. Morningstar Center for Retirement & Policy Studies. https://www.morningstar.com/content/cs-assets/v3/assets/blt9415ea4cc4157833/bltd4bb26598046aed4/66a1535de91a178e5c15872a/Introducing_the_Morningstar_Model_of_US_Retirement_Outcomes_-_July_2024_-_final.pdf.

Mastrobuoni, Giovanni. 2009. "Labor Supply Effects of the Recent Social Security Benefit Cuts: Empirical Estimates Using Cohort Discontinuities." *Journal of Public Economics* 93 (11-12): 1224–33.

Meinshausen, Nicolai. 2006. "Quantile Regression Forests." *Journal of Machine Learning Research* 7: 983–99. https://jmlr.org/papers/v7/meinshausen06a.html.

Menten, Gaëtan de, Gijs Dekkers, Geert Bryon, Philippe Liégeois, and Cathal O'Donoghue. 2014. "LIAM2: A New Open Source Development Tool for Discrete-Time Dynamic Microsimulation Models." *Journal of Artificial Societies and Social Simulation* 17 (3): 9. https://doi.org/10.18564/jasss.2574.

Mermin, Gordon B. T. 2005. *The Effect of Benefit Reductions on the Distribution of Social Security Benefits*. Report No. 411260. The Urban Institute.

OpenM++ development team. 2026. *OpenM++: Open Source Microsimulation Platform*. https://openmpp.org.

Orcutt, Guy H, Martin Greenberger, John Korbel, and Alice M Rivlin. 1961. "Simulation of Economic Systems." *The American Economic Review* 51 (5): 893–907.

Penn Wharton Budget Model. 2025. *Penn Wharton Budget Model: Microsimulation*. https://budgetmodel.wharton.upenn.edu/model/microsimulation/.

Policy Simulation Library. 2026. *Tax-Calculator*. https://taxcalc.pslmodels.org/.

PolicyEngine. 2026. *PolicyEngine: Open-Source Tax-Benefit Microsimulation*. https://policyengine.org.

Sabelhaus, John, and Jae Song. 2010. "The Great Moderation in Micro Labor Earnings." *Journal of Monetary Economics* 57 (4): 391–403.

Schwartz, Christine R. 2010. "Earnings Inequality and the Changing Association Between Spouses' Earnings." *American Journal of Sociology* 115 (5): 1524–57. https://doi.org/10.1086/651373.

Smith, Karen E. 2015. *Can Social Security Be Solvent?* Program on Retirement Policy Report No. 72196. The Urban Institute.

Smith, Karen E., Richard W. Johnson, and Melissa M. Favreault. 2020. *Five Democratic Approaches to Social Security Reform: Estimated Impact of Plans by 2020 Presidential Candidates*. Report No. 103050. The Urban Institute.

Social Security Administration. 2024. *Projection Methodology: Modeling Income in the Near Term, Version 8 (MINT8)*. https://www.ssa.gov/policy/docs/projections/methodology.html.

Spielauer, Martin, Thomas Horvath, and Marian Fink. 2020. *microWELT: A Dynamic Microsimulation Model for the Study of Welfare Transfer Flows in Ageing Societies from a Comparative Welfare State Perspective*. WIFO Working Papers 609/2020. Austrian Institute of Economic Research (WIFO).

Tax Foundation. 2025. *The Tax Foundation's Taxes and Growth Model*. https://taxfoundation.org/research/all/federal/overview-tax-foundations-taxes-growth-model/.

Tax Policy Center. 2025. *Microsimulation Model FAQ*. https://taxpolicycenter.org/resources/tax-model-resources/tpcs-microsimulation-model-faq.

Technical Panel on Assumptions and Methods. 2023. *2023 Technical Panel on Assumptions and Methods: Report to the Social Security Advisory Board*. Social Security Advisory Board. https://www.ssab.gov.

The Budget Lab at Yale. 2026. *Tax-Simulator: Microsimulation Model of US Federal Tax System*. https://github.com/Budget-Lab-Yale/Tax-Simulator.

Urban Institute. 2024. *Urban's Dynamic Simulation of Income Model 4 (DYNASIM4)*. https://www.urban.org/research/publication/urbans-dynamic-simulation-income-model-4.