Administrative quality control data as an oracle: verifying encoded SNAP rules and simulating payment error under cost sharing

Max Ghenis

2026-08-09

Beginning in fiscal year 2028, each state's Supplemental Nutrition Assistance Program (SNAP) payment error rate sets its share of benefit costs, in steps of five percentage points at rates of 6, 8, and 10 percent. A measurement system built for oversight now moves large sums, which raises two questions this paper treats together: whether the rules that produce benefit amounts can be executed correctly, and how the errors that remain should be measured and simulated. I use the USDA Quality Control (QC) public-use files as a verification oracle for independently encoded SNAP rules: across seven states, the encodings reproduce the file's benefit-computation chain, from file-recorded intermediates, exactly — all 6,081 in-scope cases of the 6,194-case official universe (113 documented exclusions), six computation stages per case, zero dollar tolerance — a process that surfaced and fixed two defects in the encodings and surfaced errata in the federal technical documentation. Replaying reconstructed pre-edit case facts through the verified engine explains 78.4% of Colorado's above-threshold official error cases — and 91.4% of sub-threshold deviations — as correct arithmetic on wrong facts. I also fit and validate a distributional model of case-level payment deviations, and report its validation failures alongside its results. The central finding, from an open Monte Carlo simulator resampling each state's own QC cases, is about measurement: at regulatory sample sizes, several states face nearly even odds — 46 to 50 percent — of being assigned a different cost tier than their official point rate implies. The fiscal 2025 rates, published in June 2026, land where that mechanism says realized years should: 18 of 53 jurisdictions changed cost-sharing tiers in one year, and only seven of those changes exceed the 95% band that two years of sampling noise — estimated from fiscal 2024 microdata alone — implies.

1 Introduction

Under 7 U.S.C. § 2013(a)(2), as amended by the One Big Beautiful Bill Act (Pub. L. No. 119-21, 2025; OBBBA), a state whose SNAP payment error rate reaches 6% will pay 5% of benefit costs beginning in fiscal year 2028; at 8%, the share is 10%; at 10%, it is 15%. The rate that determines the FY2028 share is the state's fiscal 2025 or 2026 rate, at its election; later years use the third preceding year. Implementation is delayed for the highest-error states: a state whose fiscal 2025 rate times 1.5 reaches 20% — that is, a rate above 13.33% (13.34% at the published two-decimal precision) — starts in FY2029 instead, and a state whose fiscal 2026 rate crosses the same test starts in FY2030.[1] For a mid-sized state, one tier is worth tens of millions of dollars a year; for the largest states, more than a hundred million.

The error rate is estimated by the Quality Control system: each state re-reviews a probability sample of its own active cases — by formula in 7 C.F.R. § 275.11, between 300 and 2,400 annually, or 300 to 1,020 under an optional reduced schedule that every state elected in fiscal 2024 (U.S. Department of Agriculture, Food and Nutrition Service 2025b) — and federal reviewers re-review a subsample, from which the Food and Nutrition Service (FNS; renamed the Food and Nutrition Administration in 2026 — the fiscal 2025 rate publication carries the new letterhead (U.S. Department of Agriculture, Food and Nutrition Administration 2026), and this paper keeps FNS, the name on every document the fiscal 2024 data cite) estimates official state rates with adjustments including federal re-review integration. The public-use files that result record, for roughly 45,000 cases a year, what the agency issued and what the review concluded the case should have received, with coded findings for every discrepancy.

This paper treats those files as a research instrument with three uses, the first in the software-testing sense of an oracle: an external source of ground-truth outputs for given inputs.

First, as a verification target for rules as code. I test independently encoded SNAP rules — statute, regulations, and state policy manuals encoded as executable rule modules — against the QC file's recorded benefit computations in seven states, holding the bar at exact reproduction: zero dollar tolerance across six computation stages per case (Section 3). The target's precise character matters and is stated in Section 3.1: the file's benefit chain is the FNS QC Minimodel's computation on edited, internally consistent case records, so parity certifies agreement with the agency's own computational canon, not independent adjudication. Reaching it surfaced defects on three sides: two in the encodings under test, errata in the federal technical documentation, and a residue attributable to state issuance systems.

[1] 7 U.S.C. § 2013(a)(2)(B)(iii). The test applies mechanically to each year's rate and is independent of the state's election. The fiscal 2025 rates published in June 2026 settle the first leg: seven jurisdictions — Alaska (23.15%), the District of Columbia, Delaware, Georgia, Illinois, New Mexico, and Oregon — crossed on fiscal 2025 and owe nothing in FY2028, with no bill before FY2029 — keyed to the fiscal 2026 rate unless that rate also crosses, which pushes the start to FY2030 and is nearly certain for Alaska. Fiscal 2024 rates would have qualified ten, including New York (14.09%); New York's fiscal 2025 rate of 13.18% misses the 13.33% threshold, so its delay now rides on the fiscal 2026 measurement. Any projection of FY2028 outlays must model the delay, and the simulator does: delayed years enter expected bills as zero.

Second, as raw material for decomposing measured error. QC's own cause coding, a finding-nature classification, and a replay of reconstructed pre-edit case facts through the verified engine give three nested estimates of how much of the error rate reflects computation failure versus wrong facts correctly processed (Section 4).

Third, as training and validation data for a case-level model of the error process (Section 5) and the basis for a distributional simulator of measured rates under the cost-sharing formula (Section 6), public at snap-qc-sim.vercel.app. The simulator's central output is a measurement result: near tier boundaries, sampling noise alone gives several states nearly even odds of a different cost tier than their point rate implies.

The analysis pipeline, simulator, and this manuscript live in one public repository; a fact catalog maps every quantitative claim to a committed artifact (Section 10). Because the claims depend on it, the project's verification discipline applies to itself: the pipeline is deterministic (independent clean-room reruns reproduce every analysis-pipeline artifact byte-for-byte, most recently on 2026-08-07, and the headline simulation artifact was independently reproduced from the raw file to four decimals), and this revision incorporates four rounds of adversarial review — nineteen referee reports, archived unedited in the repository — whose findings included two direction-of-effect errors in an earlier draft of this manuscript and a data-definition defect in the deployed simulator. An author verification pass between the second and third rounds additionally caught validation claims that had gone stale when the model pipeline was refit after publication; the third round then found two further stale siblings from the same refit. All are corrected here, with the superseded figures preserved in the fact catalog.

2 The QC system and the money attached to it

QC reviews produce, for each sampled active case, the issued allotment (RAWBEN), a corrected allotment constructed from the reviewer's findings (BENFIX), and — when they differ by more than a tolerance threshold ($56 in fiscal 2024, indexed from a $37 statutory base under 7 U.S.C. § 2025(c); the published sequence $48, $54, $56, $57, $58 for fiscal 2022–26 is uniquely reproduced by flooring the indexed value, which puts the fiscal 2027 threshold at an estimated $59 — $59.58 unrounded on the May 2026 Thrifty Food Plan cost of $1,018.20, with $60 requiring a June figure of at least $1,025.40) — a payment error; findings are coded for sub-threshold deviations as well. The official state rate combines over- and underpayment rates (U.S. Department of Agriculture, Food and Nutrition Service 2025a): Colorado's 9.97% is 7.91 over plus 2.06 under, and Maryland's 13.64 includes the nation's highest underpayment rate, 4.79 — so cost sharing partly bills states for underpaying their own residents.

Four features of the system frame everything that follows.

The public file is not the full review universe. The edited public-use file excludes cases with ineligible findings — for which the official methodology scores the entire issued benefit as error dollars — cases whose overissuance equals or exceeds the issued benefit, incomplete

reviews, and internal-consistency drops (U.S. Department of Agriculture, Food and Nutrition Service 2025b). In fiscal 2024 this removed 1,037 ineligible-finding and 406 full-overissuance cases nationally. The file is therefore truncated at the top of the error-magnitude distribution. Every file-based estimate in this paper inherits that truncation: it is part of why Colorado's file-derived rate (8.88%) sits below its official 9.97%. Restoring the excluded cases would add a large between-component variance term — their per-case error contributions are the full issued benefit — so the truncation biases simulated sampling variability downward; the i.i.d. approximation to states' month-stratified sampling designs plausibly biases it upward, so the tier-noise results of Section 6 are likely, though not certainly, understated.

The measured rate is a property of the measurement system. Between roughly 2009 and 2016, several states used consultant-driven error-review practices that biased QC findings downward; the Department of Justice recovered more than $67 million in False Claims Act settlements from eight states over 2017–2021 (U.S. Department of Justice 2021), FNS issued formal anti-bias guidance, and national rates for fiscal 2015–16 were not published (Congressional Research Service 2018). This history is why the present analysis trains no earlier than fiscal 2017 — and why the fiscal 2025–27 rates now being measured, which carry orders of magnitude more money than the settlements-era rates did, cannot be treated as exogenous to the incentive they price.

Only active cases are priced. The QC system also reviews negative actions — denials, terminations, suspensions — feeding a separate case-and-procedure error rate that OBBBA does not price. A state can therefore lower its priced error rate by tightening the front door, shifting error into the unpriced metric. This paper analyzes the priced, active-case system; the asymmetry belongs on any list of the formula's incentive properties.

Precision was contractually waived. A state electing the reduced sampling schedule agrees not to dispute later error-rate findings on the basis of the precision of the estimates (U.S. Department of Agriculture, Food and Nutrition Service 2025b). All 53 jurisdictions had so elected as of fiscal 2024 — before precision determined tier placement worth tens of millions of dollars. Whether states revert to the standard schedule, regaining both precision and dispute rights, is a live sampling-plan question that Section 6 prices.

3 The QC file as a verification oracle

Rules as code — encoding statutes and regulations as executable, testable logic — has a recognized validation problem: against what do you test the encoding? Unit tests encode the encoder's own reading of the law, and cross-model comparisons inherit both models' assumptions; the literature has approached validation by compiling legal text alongside experts (Mérigoux, Chataing, et al. 2021; Mohun and Roberts 2020) and, closest to this paper, by testing an independent encoding of the French tax code against the tax authority's own published test cases (Mérigoux, Monat, et al. 2021). Administrative QC data extends that

strategy to a benefits program at case level: tens of thousands of real cases per year, each carrying inputs and the agency's own computed outputs.

I tested encodings of federal SNAP law and seven states' policy manuals (Arizona, California, Colorado, Georgia, Maryland, New York, Texas), maintained in the public rulespec-us repository and executed by a deterministic rules engine, against the fiscal 2024 QC file. The file contains 6,194 cases in the official universe for those states; 6,081 are in scope for replay — all 113 exclusions (1.8%) are SSI-CAP (Combined Application Project) standardized-benefit units, whose allotments follow a demonstration schedule outside the computed chain, logged per case by the harness. Table 1 reports the counts. For each in-scope case the harness asserts six values — gross income, standard deduction, excess shelter deduction, net income, maximum allotment, and benefit — against the file's recorded values at zero dollar tolerance.

All 6,081 in-scope cases match on all six asserted values — 36,486 exact cell comparisons, no mismatches.

Table 1: Fiscal 2024 verification coverage by state. All exclusions are SSI-CAP standardized-benefit units, logged per case by the harness.

State	In-scope cases replayed exactly	Official-universe cases	Excluded
Arizona	922	925	3
California	883	883	0
Colorado	856	856	0
Georgia	945	945	0
Maryland	722	745	23
New York	847	885	38
Texas	906	955	49
Total	6,081	6,194	113

3.1 What the parity target is

Exactness claims invite overreading, so the scope deserves a precise statement — more precise than an earlier draft of this paper gave it.

The compared benefit chain is the one recorded in the edited public-use file, whose benefit variable (FSBEN) is computed by the FNS QC Minimodel — the agency's own benefit-calculation software — and whose editing process reconciles case records until internal identities hold, adjusting deduction fields in a documented sequence until the calculated benefit matches the raw benefit within $5 (U.S. Department of Agriculture, Food and Nutrition Service 2025b). The comparison also supplies several intermediates directly from the file rather than deriving them from raw facts: the QC-calculated medical and child-support deduction amounts, the

recorded utility allowance, and the recorded categorical-eligibility status; eligibility screens receive passing defaults, since every filed case was enrolled.

Parity therefore certifies the following: that the encoded computation chain — income aggregation, deduction sequencing, net-income arithmetic, allotment lookup, rounding — agrees exactly with the agency's own computational canon, over the full diversity of 6,081 real case configurations, parameter values, and boundary conditions. It is an agreement between two implementations of the same rules, closer in kind to the cross-model comparisons this method extends than to independent adjudication of case outcomes; what distinguishes it from ordinary cross-model comparison is the counterparty (the agency's own canon), the case realism, and the zero-tolerance bar. It does not certify independent derivation of every state eligibility and deduction pathway from raw facts. Extending the harness to feed raw facts — recorded medical expenses rather than the computed deduction, energy-assistance facts rather than the utility tier — is the necessary next step before the engine can compute counterfactual intermediates under alternative policy (Section 7). The reviewer-adjudicated quantities in the file (RAWBEN, BENFIX, error amounts and findings) are not the parity target; they are the subject of Section 4.

Holding tolerance at zero is what made the exercise informative. Early runs did not match, and each mismatch class was a finding about somebody's rules: two defects in the encodings under test (a regulatory dollar literal superseded by statute, and a missing whole-dollar rounding step — both fixed in the public repository, with the oracle as the detector), errata in the technical documentation's variable descriptions (reported upstream), and — once the encodings reproduced the file's chain exactly — a residue in issued amounts attributable to the issuing side, which Section 4 takes up. The corrections were global rule fixes, not case-specific accommodations; still, the exactness result is post-correction on the same case set, a reuse the certification protocol mitigates but does not eliminate. A confirmatory run on the fiscal 2025 file at its release is the committed next test.

The toolchain is certified for reproduction: a pinned engine build and encoding-repository commit reproduce the Colorado suite — 856 cases, 5,136 cells including the 856 benefit cells — under recorded hashes for every input, with an independent audit rebuilding the engine to a byte-identical binary and re-running the comparison to byte-identical evidence. Throughput is 230 cases per second end to end on a laptop, so verification at this scale is not compute-bound for this workload; details and timing calibration are in the repository's certification report.

4 Decomposing measured error: computation versus facts

If a verified engine issued every benefit, how much measured error would disappear? The QC file supports three nested answers for Colorado fiscal 2024, summarized in Table 2: 305 of 856 sampled cases carry recorded payment deviations (RAWBEN BENFIX) — but only 110 of the 305 are official errors above the $56 threshold; the other 195 are sub-threshold deviations the official rate excludes. Weighted, the 305 carry $112.6M per year in deviation dollars on

$1.268B of issuance, of which $18.5M (16%) is sub-threshold: the deviation-based 8.88% file rate therefore combines an official-definition file rate of 7.42% with sub-threshold dollars, while the official 9.97% additionally incorporates the federal re-review adjustment and the excluded ineligible-case error of Section 2. All three layers exclude, by construction of the file, that ineligible-case error.

Table 2: Three nested decompositions of Colorado fiscal 2024 measured error. Layers 1–2 are shares of weighted error dollars; layer 3 is a share of filtered cases. Cause and nature coding reflect state reviewer judgment under handbook guidance (U.S. Department of Agriculture, Food and Nutrition Service 2023) and carry cross-state coding variation; fiscal 2024 also brought minor revisions to the finding-code sets.

Layer	Definition	Computation-side share	Universe and units
1. Cause codes	QC agency-cause codes naming the computing apparatus (programming, arithmetic, mass change)	7.3% strict; 10.5% adding policy-misapplied, budgeted-wrong, computer-user	Colorado error dollars; nationally 3.9% strict ($320M/yr), 18.4% broad ($1.5B of $8.1B)
2. Finding natures	Nature codes inherently describing computation (rounding, conversion, wrong standard, benefit miscomputed), cause-disambiguated	3.3% pure-computation; 6.6% system-caused input; 4.2% mixed; 86.0% other input	Colorado error dollars
3. Engine replay	Reconstructed pre-edit facts replayed through the verified engine and compared to issued amounts	Above-threshold official errors: 21.6% of cases unexplained (upper bound), 78.4% explained as correct arithmetic on wrong facts. Sub-threshold deviations: 8.6% unexplained, 91.4% explained. Blended: 13.1% / 86.9%	283 filtered Colorado deviation cases (97 above threshold, 186 below); case shares, not dollar shares

The strictest layer reconstructs each error case's pre-edit original values using the public reconstruction solver of Giannella and Molin (2026) — which exploits the file's disclosure-

protection perturbation structure to recover original values — adapted to fiscal 2024, then replays those originals through the verified engine. Of 283 deviation cases surviving the solver's consistency filters (22 of 305 are excluded), 246 reproduce the issued amount at the file's own $5 editing tolerance: the agency's arithmetic was correct, applied to wrong facts. The rate differs by threshold status, and the distinction matters for how the headline reads: among the 97 above-threshold official error cases, 76 are explained (78.4%, binomial standard error about 4 points; the 21.6% residual is the computation-side upper bound for official errors), while among the 186 sub-threshold deviations 170 are explained (91.4%). The solver and the engine partition the 283 cases identically — two implementations with distinct codebases and logic reaching the same classification, though both consume the same file, and 33 of the 246 explained cases have deviations of $5 or less, mechanically within the comparison tolerance. The residual class includes solver limitations as well as genuine multi-variable or computational discrepancies. The complete per-case replay output (283 rows), the layer-1 and layer-2 classification outputs, and the reconstruction scripts are committed with this paper.

The three layers agree on the structure: most measured SNAP payment error among eligible cases is information failure, not calculation failure — consistent with the administrative-burden literature's account of where compliance costs bind (Herd and Moynihan 2018). Correct-computation guarantees address the smaller share directly — roughly 3 to 10% of error dollars by layers 1–2, and bounded above by the replay's 21.6% case share among official errors — plus whatever share of information failures better tooling prevents indirectly through verification prompts and documentation-requirement computation at intake. Ineligibility errors, excluded from the file, are predominantly facts-driven, which would push the information-failure share higher still; that inference, unlike the in-file shares, is not directly computed here. The indirect channel is where policy design enters: rules determine what a caseworker must collect and verify, which is the mechanism the modeling section takes up.

5 Modeling the error process

Pricing decisions — what a documentation requirement costs in error, what an audit-volume change buys — require a model of the error-generating process at case level. I fit one on the QC files for 2017–2019 and 2022–2023 (pandemic years excluded), evaluating on fiscal 2024: 217,656 training and 44,800 evaluation cases in the official universe, all estimates weighted by the file's case weights, with models estimated in scikit-learn (Pedregosa et al. 2011). Fiscal 2024 informed pipeline development across the audit-and-correction rounds and is not a pristine holdout; a frozen-pipeline confirmation on fiscal 2025 is committed as the next test.

Targets come from the file's adjudication fields. The official error label is reviewer status (over- or underpayment) with an error amount above the year's threshold. The signed deviation is $D = \text{RAWBEN} - \text{BENFIX}$, which reconciles with the recorded error amount for 99.997% of weighted fiscal 2024 cases; the recomputed formula benefit FSBEN — an earlier, incorrect

target choice that an adversarial audit caught — matches only 83.64%, because it conflates legitimate adjustments such as proration with error.

Discrimination. A gradient-boosted classifier on household covariates plus a formula-benefit anchor reaches a weighted area under the receiver operating characteristic curve (ROC AUC) of 0.761 on fiscal 2024; adding policy-burden intermediates extracted from the file (medical-expense claiming and documentation-requirement proxies, self-employment records, utility-claim structure, deduction counts) moves it to 0.767 (+0.006 AUC, +0.003 PR-AUC). Burden intermediates add little discrimination on this label. Operationally the model is nonetheless useful: at a 5% weighted review budget it reaches 47.8% precision against a 13.4% weighted base rate — case targeting at three and a half times chance, though QC-sampled cases are not the operational targeting population and the estimate carries no design-based standard error.

The deviation distribution. For simulation I estimate a per-case distribution of the signed deviation with a hurdle structure: a deviation probability $P(|D| > 0.5)$ and an above-threshold probability, each estimated by gradient boosting with nested out-of-fold isotonic calibration; a sign model (calibrated AUC 0.700 in the shipped frozen configuration); nine conditional quantiles of $\log|D|$ fitted by quantile-loss gradient boosting (Koenker and Bassett 1978), made monotone per case by rearrangement (Chernozhukov et al. 2010); and an exponential tail in logs beyond the 99th percentile. Conditional magnitudes retransform with an out-of-fold Duan smear of 1.173 (Duan 1983); predicted mean conditional magnitude is $186 against $189 observed. Model figures throughout this section quote the current committed artifacts, regenerated when the seventeen additive features described at the end of this section entered every stage; the pre-feature values (sign AUC 0.686, tail log scale 0.271, the earlier calibration table) are preserved in analysis/FEATURES_REPORT.md with the full before/after roll.

Validation, including failures. Table 3 summarizes. An adversarial statistical review of the first deployed model found the tail fitted at one depth but attached at another (overstating extreme-tail mass) and cross-state level gaps propagating into simulated spreads. The fix round repaired both — refitting the tail at its attachment depth (log scale 0.252 in the current artifact, from 0.467 pre-fix), capping each case's magnitude at a physical maximum, and freezing fiscal-2023-fit state dollar factors — and also moved the coverage evaluation to the configuration that actually ships: the frozen model trained through fiscal 2022 (62,984 training deviators, against the through-2023 primary model's 79,919), with fiscal 2023 held out for factor fitting. That evaluation-frame change, not the three fixes — none of which enters the coverage computation, which compares observed log-magnitudes with the raw conditional quantiles — is what moved the coverage figures. An earlier version of this section reported the through-2023 primary model's coverage: within 3 points of nominal at seven of nine levels, gaps −0.3 to −3.5 points; the fact catalog's E3 row records the supersession. The shipped frozen configuration does not attain it: within 3 points at only two of nine levels, all nine gaps negative (−0.38 to −7.30 points, worst at the 75th–90th percentiles). This is one-sided under-coverage — predicted quantiles sit too low, understating mid-to-upper magnitudes, coherent with the $186-versus-$189 comparison. A benchmark against a quantile regression forest on the identical protocol, with the decision rule fixed in the protocol code before results

existed, shows the failure is not specific to the gradient-boosted quantile stack: run against the pre-feature stack (the expanded feature set has not been re-benchmarked), the forest improves mean absolute coverage (4.10 against the pre-feature 4.38 points) but shifts under-coverage from the tail to the central levels with a worse maximum gap, degrades probability-integral-transform calibration, and loses the factor-adjusted state dollar-rate comparison (0.951 against 0.928 points — a different object from Table 3's official-error-rate MAE), so the gradient-boosted stack is retained; causes the two estimators share — the feature set, the tree-ensemble family, the temporal gap from the frozen training window to fiscal 2024 — remain candidates. The deployed simulator serves this model in exactly one place — a standard-medical-deduction scenario whose flipped per-case parameters, paired-bootstrap intervals, and level gate ship in an export pinned to the exact model file by hash, with the seven jurisdictions outside the [0.7, 1.4] factor-adjusted level-ratio range disabled — and no simulation result in this paper uses model draws.

Table 3: Fiscal 2024 state-level calibration of predicted official-error rates, matched frozen-model comparisons, from the current committed artifact (post-feature-round; the pre-feature table — 1.81/1.65pp unfactored, 0.885/0.785pp factored — is preserved in analysis/FEATURES_REPORT.md). Factors are fit on out-of-sample fiscal 2023 predictions and frozen before touching fiscal 2024.

Quantity	Unfactored (frozen model)	With frozen FY2023 factors
State MAE, equal-weighted	1.73pp	0.875pp
State MAE, issuance-weighted	1.57pp	0.808pp
Cross-state correlation (equal-weighted)	0.56	0.91

State-level calibration uses a strictly temporal protocol: models fit through 2022 predict 2023 out of sample; state observed-to-predicted ratios, shrunk toward one by an empirical-Bayes precision rule $\tilde{f}_s = 1 + (f_s - 1)\,\tau^2/(\tau^2 + v_s)$, with v_s the delta-method ratio variance on Kish effective sample size, are frozen and applied to fiscal 2024. Most cross-state alignment comes from the factors, whose interpretation — administration, omitted policy, model error — the data do not identify.

Policy contrasts are descriptive. Under an event definition requiring a medical finding with payment impact, elderly/disabled claimants in states without a standard medical deduction — where itemized documentation is required — show a 2.60% medical-event rate, against 2.97% in a mixed comparison cell (standard-deduction-state claimants pooled with below-floor claimants everywhere); the cross section runs against a simple burden-increases-error reading, and adoption is endogenous. Calendar-aligned adoption contrasts against never-adopting states are small and mixed — Arizona +2.07 points claimant-conditioned but +0.64 on the stable all-elderly/disabled denominator; Kentucky −0.42 claimant-conditioned; California +0.11 claimant-conditioned but −1.40 on the stable denominator; Louisiana and Michigan within 0.2 points of zero, though those two adopters have zero treated-period events in the

windows, so their contrasts are control-trend arithmetic only — with single-digit event counts throughout. Two cautions transfer beyond this paper: an earlier pipeline's dramatic Kentucky estimate (−5.3 points) dissolved under a corrected event definition and calendar windows; and standard-medical-deduction adoption is typically bundled with utility-allowance offsets for cost neutrality, so these are contrasts of policy bundles. Causal identification awaits engine-computed counterfactuals (Section 7).

A subsequent feature round, prompted by practitioner review, added seventeen case-level features in three families: certification timing (months since certification and period length from the file's LASTCERT and CERTMTH, a final-two-months indicator, and its interaction with elderly/disabled composition), state broad-based categorical eligibility from the FNS State Options Reports (44 adopting agencies in fiscal 2024, mapped to each training year's report edition; the file's own CAT_ELIG stays excluded because its codes conflate BBCE with traditional categorical eligibility), and Medicare Part B premium bands — reconstructing gross medical expense as the file's excess amount plus the $35 deduction floor and comparing it with the sampled month's published premium. The additive gains are modest and reported as such: classifier ROC AUC 0.7666 to 0.7679, hurdle stage-1 AUC 0.8356 to 0.8397, equal-state calibration MAE 1.83 to 1.71 percentage points, while distributional coverage worsens slightly (mean absolute gap 4.38 to 4.64 points) and is retained. The band feature's descriptive validation is exact where it should be: all 36 elderly Colorado cases censored at the state's $165 standard medical deduction map, after the floor reconstruction, to $200 gross — inside the just-above-premium band a practitioner predicted would matter — where a naive comparison on the excess scale misclassifies six of them.

6 Simulating measured rates under cost sharing

The public simulator resamples each state's own sampled cases with replacement at the realized or a hypothetical sample size and recomputes the weighted error rate (Efron 1979), centered on the official rate. The error definition is the official one — adjudicated status with the recorded amount above the threshold; an adversarial review found the deployed version initially gated on the banned benefit-difference definition, now corrected. Three findings, all on fiscal 2024 data, all inheriting the file's truncation of ineligible-case error (which understates the variability shown) and an i.i.d. approximation to each state's actual sampling design; the official rate's federal re-review adjustment enters only as a fixed level. All figures quote the committed simulation artifact.

Tier assignment approaches a coin flip for boundary states (Figure 1, Table 4). With no policy change at all, the probability that a fresh QC-style sample lands in a different cost-share tier than the official point rate implies is 46–50% for the five states nearest boundaries.

The formula maps a point estimate to a tier with no adjustment for sampling uncertainty; near a boundary, tier assignment turns as much on the sampling draw as on the underlying rate.

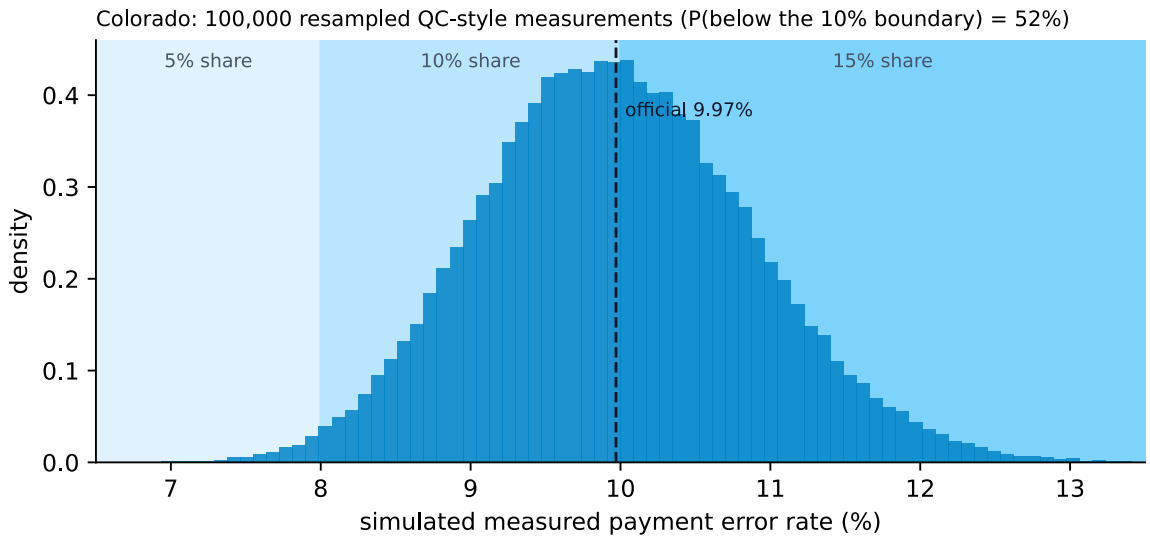

Figure 1: Colorado's simulated measured-rate distribution (100,000 resampled QC-style measurements, official error definition, centered on the official 9.97% rate) against the cost-share tier bands. The state's official rate sits 0.03 points below the 15%-share boundary; 52% of simulated measurements fall below it and 48% above. Generated by the committed paper/generate_figures.py (seed 11).

This is the accountability-measurement problem documented for school accountability ratings (Kane and Staiger 2002), now carrying direct fiscal prices.

Table 4: Boundary states under the fiscal 2028 tier schedule applied to fiscal 2024 issuance — an illustration of stakes, not a forecast. Colorado's step from the 10% to the 15% tier is $63.4M per year.

State	Official rate	P(different tier)	Expected annual cost share	SD of the bill
North Dakota	7.91%	50%	$7.9M	$3.5M
Washington	6.06%	48%	$50.2M	$48.5M
Colorado	9.97%	49%	$156.4M	$32.8M
Kansas	9.98%	47%	$46.9M	$9.5M
Nevada	5.94%	46%	$23.7M	$26.0M

Audit volume is a two-sided instrument. Increasing the QC sample shrinks sampling variance around the state's underlying rate, holding the error process fixed (Section 8; the corrective-feedback channel from reviewing more cases is not modeled, and these figures are gross of review costs — which rose for states when OBBBA cut the federal administrative match from 50 to 25% beginning fiscal 2027). In expectation, 500 additional reviews save money for states just below a boundary — Missouri +$3.9M, Tennessee +$3.7M, Indiana +$3.3M annually — and cost money for states just above one, for whom the chance of a below-boundary draw had positive expected value: California −$30.3M, Pennsylvania −$13.1M, Texas −$13.1M. Added volume almost always reduces the standard deviation of the annual bill (California by $44M), so risk-averse states may rationally buy audits against their expectation. The formula thus embeds an asymmetry: expected cost share falls with additional sample volume for states just below a boundary and rises for states just above one — a property worth reading alongside the no-dispute election of Section 2.

Category-suppression bounds are large. Removing 50% (100%) of the observed error dollars attached to the finding elements that four policy options standardize — the standard medical deduction, standard self-employment deduction, heat-and-eat utility standardization, and broad-based categorical eligibility resource exemptions — corresponds to roughly $609M ($1,310M) per year nationally in expected cost share. These are accounting bounds, not causal estimates: they hold the error process fixed and answer "what if these categories' observed error dollars shrank," an upper-bound identity rather than a behavioral prediction — the descriptive adoption contrasts of Section 5 are the reason for the caution. They speak only to measured error and cost share; the options themselves change benefit amounts, eligibility, and program cost, none of which is modeled here. Two further scope notes: OBBBA itself narrowed the heat-and-eat channel, restricting the LIHEAP-triggered utility allowance to households with an elderly or disabled member (§ 10103), so that lever's fiscal 2024 error mix overstates

its remaining reach; and the expected cost-share effect of a marginal error reduction is near zero for states deep within a tier (New York at 14.09%, Alaska at 24.66%) and largest near boundaries.

6.1 The realized fiscal 2025 rates

The simulation's claim — that tier assignment at current sample sizes is substantially a draw from sampling noise — is built from fiscal 2024 microdata alone. On June 24, 2026, the renamed agency published the fiscal 2025 rates (U.S. Department of Agriculture, Food and Nutrition Administration 2026); this repository postdates that publication, so what follows is a check of input independence, not a registered forecast — the fiscal 2025 rates enter the analysis only as the realized outcomes being classified, and the noise scale that classifies them uses no fiscal 2025 information. Eighteen of 53 jurisdictions changed cost-sharing tiers. Whether any single change reflects program performance is exactly the question sampling error poses: under the committed simulator's FY2024 sampling standard deviations, with both years carrying independent noise of equal scale, a year-over-year move is distinguishable from noise at 95% confidence only beyond 2.77 of a single year's standard deviations. Ten jurisdictions cleared that bar (Delaware, Florida, Hawaii, Illinois, Kentucky, Minnesota, North Carolina, New Jersey, Ohio, and West Virginia) — but only seven of those ten changed tiers, so 11 of the 18 tier changes sit inside the noise band; the median jurisdiction moved 1.38 standard deviations. Movement figures quote the committed artifact (analysis/fy2025_movement.json, deterministic at seed 11); the tier-step and election dollars below come from the deployed simulator's committed inputs (app/public/data.json) and its election engine.

The two tails of that distribution make the point more sharply than the counts. Colorado moved from 9.97% to 10.09% — 0.12 percentage points, 0.13 standard deviations, indistinguishable from re-measuring the same program — and that move crossed the 10% boundary into the 15% tier, repricing its FY2028 exposure by roughly $63M a year. Real movement exists alongside it: Hawaii went from 6.68% to 10.92% (11.2 standard deviations, two tiers up), New Jersey from 14.33% to 6.86% (−11.2, from the top tier to the 5% tier), Kentucky from 9.11% to 4.70% (−7.2, into the 0% tier). The formula cannot tell these apart; it prices Colorado's noise and New Jersey's real movement — whatever mix of program change and review practice produced it — with the same schedule. The national rate fell from 10.93% to 10.62%.

The realized year also calibrates how much true year-over-year process movement the sampling-only simulator omits. Across states, the variance of the one-year change decomposes as twice the mean squared sampling standard deviation plus a process-drift variance τ^2; the method-of-moments estimate is $\tau = 1.62$ percentage points (95% CI 0.81–2.25 from a paired bootstrap over jurisdiction movement–SD rows), and a median/MAD version robust to the large realized moves gives $\tau = 1.07$ (CI 0–1.74). Process drift on this transition is likely real — the classical interval excludes zero; the robust one does not — and comparable in scale to sampling noise for a typical state, though a single transition cannot separate program drift from any change in QC design or practice between the two years. The simulator's draws remain sampling-only —

a disclosed lower bound on total year-over-year uncertainty — with the drift estimate carried as a candidate calibration rather than silently added.

Churn reaches the delay clause too. Had fiscal 2024 rates carried the delay test, ten jurisdictions would have met it; at fiscal 2025 — the first year the test binds — seven do. Florida, Massachusetts, Maryland, New Jersey, and New York left that hypothetical roster and Delaware and Illinois joined it: the assignment that zeroes a state's FY2028 bill churns under the same measurement noise that moves tiers.

The fiscal 2025 publication also started the clock the statute set: the FY2028 share keys to each state's fiscal 2025 or fiscal 2026 rate, at its election, and fiscal 2026 closes September 30, 2026. The deployed simulator now prices that election — for Colorado, a 52.5% probability that a QC-sized fiscal 2026 measurement lands below the locked 10.09% rate, worth about $31M a year in expectation from the election alone; the dollars come from the near-even odds the measurement lands below the 10% tier boundary, repricing the $63.4M step — and models the delay clause as it binds: delayed years enter expected bills as zero, and a fiscal 2026 draw that crosses 13.33% pushes the start to FY2030, zeroing FY2029 as well.

7 Toward engine-computed counterfactuals

The counterfactual gap has begun to close. For one state and one option — Colorado and the standard medical deduction — the pieces now run end to end, and the run itself produced findings. The first is structural: the certified parity chain never consumes Colorado's standard-medical-deduction rule, because QC editing bakes the deduction into the stored expense fields and the federal computation path reproduces every case exactly without it — a sharper instance of the Minimodel-canon scope point of Section 3.1. (A separate raw-facts leg, reconstructing inputs rather than supplying the file's intermediates, agrees on 845 of 856 cases, its eleven divergences diagnosed to missing public TANF facts and one utility-allowance tier conflict; only the supplied-intermediates certified chain is exact.) A simulation-only composition adapter that binds the state rule into the federal path, validated by exact baseline invariance on all 856 cases, reprices the option's removal only as bounds, because the public file censors 46 medical-expense records at exactly $165: between −$1.46 and $0 per case-month. The second is a sign disagreement: feeding the engine-recomputed intermediates through the error model's direct crossing classifier, applied as a level shift, implies a cost-share change of −$1.55M to −$2.20M per year for removing the deduction. That figure is a range across the three censoring variants, not a confidence interval; the paired-bootstrap intervals are wider — the point variant's is [−$4.1M, +$0.3M] — and span zero in two of the three. The accounting-reverse bound for the same flip is +$7.1M, and the robust form of the disagreement survives the intervals: all three exclude it, with a maximum upper bound of +$0.6M. Neither sign is validated — the artifact itself warns that the fitted models implying lower error without the deduction is not evidence that documentation requirements reduce errors. The argument for retiring the accounting levers from the public simulator is accordingly not that the model is right

but that the two mechanisms answer different questions — one is an upper-bound identity scaling observed error dollars by finding category, the other a model-implied association of re-adjudication risk with the flipped feature — and their divergence on a real case shows the substitution was never innocuous. The simulator now serves only the model scenario, level-gated and labeled as an association. What remains open is breadth: engine legs for the remaining options (self-employment is not identified from public data, which records net rather than gross income) and the other verified states, feeding the harness raw facts instead of file-derived intermediates (Section 3.1), and the cost-sharing formula itself as an encodable target: the sampling formulas of 7 C.F.R. § 275.11(b) and the tier schedule of 7 U.S.C. § 2013(a)(2), with its election and delay clauses, can be encoded and tested at their published boundary values, replacing hand-transcribed constants in simulators like this one.

8 Limitations

The parity result certifies agreement with the edited file's Minimodel-computed chain from supplied intermediates — not independent derivation from raw facts, and not the reviewer-adjudicated outcome fields (Section 3.1); it is also post-correction on the same case set, pending a fiscal 2025 confirmatory run. The public file excludes ineligible-case and full-overissuance error, truncating the error distribution every file-based estimate inherits. The replay decomposition depends on a reconstruction solver whose filters exclude 22 of 305 error cases, and its residual class is an upper bound in case shares, not dollar shares. The error models predict an adjudicated, thresholded label — a property of the measurement system, with a documented manipulation history (Section 2). Cross-state model transport is imperfect: the model's raw level underpredicts high-error states, all nine coverage gaps remain one-sided negative after the tail and level-factor fixes, and the deployed simulator accordingly serves the model only for its level-gated standard-medical-deduction scenario, with seven jurisdictions disabled outright; every simulation result in this paper uses the observed-resample engine. That mode resamples i.i.d. from a probability sample whose actual design varies by state, treats the official rate's federal re-review adjustment as a fixed level with no variance contribution, and holds the error process fixed — no behavioral response of agencies or households to cost sharing, no corrective-feedback channel from audit volume, both live questions the fiscal 2025–27 files will begin to answer under the new incentives. Adoption contrasts are descriptive contrasts of policy bundles; nothing here identifies causal effects of burden policies on error.

9 Conclusion

A measurement system designed for oversight now sets prices. That makes the QC files three things at once: a case-level verification oracle against the agency's own computational canon — where the bar can be exact, and holding it exact is what made the residuals informative; the dataset showing that most measured error among eligible cases is information failure rather

than calculation failure; and a sampling frame whose noise the new formula converts into near-even-odds tier assignment for boundary states. The modeling bar cannot be exact, and this paper went through four rounds of adversarial review, archived unedited in the repository, which found — among other things — two direction-of-effect errors, one wrong external figure, and one misdescribed epistemic status (a realized-year check first framed as a committed prediction) in earlier drafts. The system this points toward is one in which the rules that compute benefits, the requirements they impose on households, and the errors they produce are executable, testable, and priced from the same verified source.

10 Data and code availability

The repository PolicyEngine/snap-qc-sim contains the analysis pipeline (deterministic; independent clean-clone reruns reproduce all five pipeline artifacts byte-for-byte on the pinned interpreter version, recorded in .python-version and each artifact's provenance block), the simulator, frozen simulation, decomposition, and certification artifacts under paper/snapshot/ (including the complete per-case replay output, the layer-1 and layer-2 classification outputs, and the reconstruction scripts), the seventeen adversarial review reports and round-1 editorial synthesis under paper/reviews/, the fact catalog (paper/FACTS.md), and this manuscript. The verification harness and per-case suite reports are public in TheAxiomFoundation/axiom-oracles; the encodings in TheAxiomFoundation/rulespec-us; the engine in TheAxiomFoundation/axiom-rules-engine. QC public-use files are available from USDA FNS (U.S. Department of Agriculture, Food and Nutrition Service 2025b), and the training files are tracked in the public Giannella and Molin (2026) repository. Reproduction commands are in the repository README.

11 Disclosure

The author leads PolicyEngine, which builds the simulator described here, and contributed to the axiom-oracles verification harness and the rulespec-us encodings whose parity this paper reports. The adversarial review protocol — seventeen independent reviews across three rounds, archived unedited in the repository — is the mitigation offered for that entanglement.

12 Acknowledgments

I thank Eric Giannella and Ben Molin for the open reconstruction and registry code this work builds on. All errors are my own.

Chernozhukov, Victor, Iván Fernández-Val, and Alfred Galichon. 2010. "Quantile and Probability Curves Without Crossing." Econometrica 78 (3): 1093–125. https://doi.org/10.3982/ECTA7880.

Congressional Research Service. 2018. Errors and Fraud in the Supplemental Nutrition Assistance Program (SNAP). No. R45147. https://www.congress.gov/crs_external_products/R/PDF/R45147/R45147.4.pdf.

Duan, Naihua. 1983. "Smearing Estimate: A Nonparametric Retransformation Method." Journal of the American Statistical Association 78 (383): 605–10. https://doi.org/10.1080/01621459.1983.10478017.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." The Annals of Statistics 7 (1): 1–26. https://doi.org/10.1214/aos/1176344552.

Giannella, Eric, and Ben Molin. 2026. Snap_qc: Reconstruction and Analysis Code for the SNAP Quality Control Public-Use Files. https://github.com/giannella/snap_qc.

Herd, Pamela, and Donald P. Moynihan. 2018. Administrative Burden: Policymaking by Other Means. Russell Sage Foundation.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." Journal of Economic Perspectives 16 (4): 91–114. https://doi.org/10.1257/089533002320950993.

Koenker, Roger, and Gilbert Bassett. 1978. "Regression Quantiles." Econometrica 46 (1): 33–50. https://doi.org/10.2307/1913643.

Mérigoux, Denis, Nicolas Chataing, and Jonathan Protzenko. 2021. "Catala: A Programming Language for the Law." Proceedings of the ACM on Programming Languages 5 (ICFP): 1–29. https://doi.org/10.1145/3473582.

Mérigoux, Denis, Raphaël Monat, and Jonathan Protzenko. 2021. "A Modern Compiler for the French Tax Code." Proceedings of the 30th ACM SIGPLAN International Conference on Compiler Construction (CC '21), 71–82. https://doi.org/10.1145/3446804.3446850.

Mohun, James, and Alex Roberts. 2020. Cracking the Code: Rulemaking for Humans and Machines. OECD Working Papers on Public Governance No. 42. OECD Observatory of Public Sector Innovation. https://doi.org/10.1787/3afe6ba5-en.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research 12: 2825–30. https://jmlr.org/papers/v12/pedregosa11a.html.

U.S. Department of Agriculture, Food and Nutrition Administration. 2026. Supplemental Nutrition Assistance Program Payment Error Rates, Fiscal Year 2025. https://www.fns.usda.gov/snap/qc/per.

U.S. Department of Agriculture, Food and Nutrition Service. 2023. FNS Handbook 310: SNAP Quality Control Review Handbook. https://www.fns.usda.gov/snap/fns-handbook-310.

U.S. Department of Agriculture, Food and Nutrition Service. 2025a. Supplemental Nutrition Assistance Program Payment Error Rates, Fiscal Year 2024. https://www.fns.usda.gov/snap/qc/per.

U.S. Department of Agriculture, Food and Nutrition Service. 2025b. Technical Documentation for the Fiscal Year 2024 Supplemental Nutrition Assistance Program Quality Control Database and the QC Minimodel. https://snapqcdata.net/datafiles.

U.S. Department of Justice. 2021. False Claims Act Settlements Concerning Bias in SNAP Quality Control Processes (Virginia, Wisconsin, Alaska, Texas, Louisiana, Mississippi, Florida, Tennessee), Cumulative Recoveries Exceeding $67 Million. https://www.justice.gov/archives/opa/pr/tennessee-department-human-services-agrees-pay-68-million-resolve-false-claims-act-liability.