Validating Generated Latent Quantities for Survey Inference

Hyun Hak Kim*

Abstract

Survey analyses use imputed, predicted, or synthetic latent quantities in downstream estimators. Completed-value accuracy or point-estimate recovery alone cannot validate task means, finite-replica covariance, refitting uncertainty, root behaviour, or the probability law supporting inference. We organise these requirements as a four-gate diagnostic evaluated in three public surveys. Across 1,000 household-remasking repetitions, Survey of Income and Program Participation coverage is 0.934–0.959. A Health and Retirement Study experiment computes 12,000 full-refit finite-M roots, yielding coverage of 0.938–0.953, standard-error-to-standard-deviation ratios of 1.003–1.015, and variance slopes near -1. Yet joint-versus-independent path covariance reverses between the first two systems. A planned-missing application shows negligible integration error at $M = 500$, whereas a weak-information Current Population Survey design exhibits root failures and deficient conditional coverage. The diagnostic locates failure in the task model, path law, replica budget, refitting calculation, or claim law. Results concern conditional validation, not national survey coverage or universal method dominance.

Keywords: finite-replica covariance; generated regressors; inferential validation; remasking; stochastic completion; survey imputation

*Professor, Department of Economics, Kookmin University, Seoul, Korea, hyunhak.kim@kookmin.ac.kr

1 Introduction

Survey production and secondary analysis increasingly insert a stochastic processing stage between the recorded answers and the statistic ultimately reported. A sampled person may not answer an item; an agency or analyst then fits a response model, an outcome model, or both; missing or protected values are generated; and the completed records are passed to a weighted mean, regression, or estimating equation. When generation is repeated, the analyst averages M completed-data paths and attaches a standard error to the resulting estimate. This workflow occurs in multiple imputation, synthetic public-use files, predicted latent scores, stochastic disclosure protection, and modern estimators that use learned nuisance functions. In each case, the reported quantity is produced by a pipeline rather than read directly from the survey record.

Validation often stops too early in that pipeline. A data producer may compare marginal distributions of observed and generated values, while an analyst may check only whether a point estimate from one completed file resembles a benchmark. Those checks are useful, but they do not establish that the generated paths preserve the particular conditional mean used by a downstream estimator, that averaging a finite number of paths has the advertised Monte Carlo variance, or that a reported confidence interval includes uncertainty from refitting the response and outcome learners. These distinctions matter especially in surveys, where design weights, household clustering, repeated observations, heterogeneous response propensities, and nonlinear target statistics can turn apparently small completion errors into consequential inferential errors.

Consider a concrete version of the problem. Starting with an observed survey file, the analyst estimates nuisance functions, uses them to create two stochastic components of a completed-data score, averages those components across M replicas, and solves for a target parameter. The two components can each have the correct mean yet be dependent because they use the same fitted learner, household, random seed, or completed record. Treating them as independent then removes a covariance term from the finite-M variance. Increasing M reduces replica-specific noise but does not remove variation shared by all replicas. Likewise, a standard error computed with the fitted learner held fixed answers a

different repeated-sampling question from one in which the learner, masks, paths, and estimating-equation root are all recomputed. Finally, success under researcher-generated masks is conditional on the completed records and masking law; it is not automatically a claim about the original survey design, the actual nonresponse mechanism, or a national population.

This paper studies how these failures can be detected before a generated quantity is used for substantive inference. We ask three empirical questions. First, does task calibration survive full learner refitting under a mask for which a benchmark is observable? Second, conditional on task-mean recovery, do path dependence and the chosen number of draws M yield the predicted variance slope, calibrated standard errors, and nominal coverage? Third, which conclusions remain defensible when the experiment moves from a researcher mask to an actual-missingness application or to a weaker-information design?

We make three contributions. First, we translate analysis-specific validation into a sequential survey workflow. The workflow separates recovery of the downstream task mean from recovery of the finite-replica covariance and from repeated-sampling inference, so that a pass at one level cannot silently be promoted to a stronger claim. Second, we make dependence between generated score components an observable diagnostic object. An exact finite-M decomposition shows where the component covariance enters, and empirical joint-path and independent-path constructions demonstrate that its sign and relative magnitude are features of the target and data system, not universal properties of an architecture. Third, we require an outer, full-refit experiment for inferential claims: each repetition reconstructs the mask, nuisance learners, stochastic paths, and final root. This connects algebraic $1/M$ predictions to actual root distributions, convergence, standard-error calibration, and coverage, while recording the probability law under which each statement is justified. The contribution is an operational synthesis and empirical validation architecture, not a claim that the underlying ideas of multiple imputation, generated-regressor correction, posterior predictive checking, or resampling are new.

The empirical design uses three public survey systems with deliberately different roles. The 2023 Survey of Income and Program Participation (SIPP) supplies two nonlinear

household-balance-sheet targets, strong household-level dependence, and a complete $M = 1, 5, 20, 100, 500, \infty$ grid. The 2006–2018 Health and Retirement Study (HRS) supplies two person-wave logistic projection targets, a new 1,000-mask full-refit finite-M experiment, and a planned-missing application. The January 2025 Current Population Survey (CPS) supplies a weak-information stress design in which lowering the observation probability from .50 to .25 exposes root and coverage failures. Thus, SIPP is the controlled finite-replica benchmark, HRS is the principal full-refit and applied demonstration, and CPS marks a boundary at which formal identities no longer guarantee usable inference.

The results caution against ranking path architectures in the abstract. In 1,000 full-refit SIPP household-remasking repetitions, conditional 95% coverage is 0.934–0.959. The HRS experiment computes 12,000 actual finite-M roots across two targets and two path laws: coverage is 0.938–0.953, standard-error-to-standard-deviation ratios are 1.003–1.015, and estimated variance slopes range from -1.024 to -0.996. Yet the covariance ordering reverses across systems. Independent-to-joint covariance ratios are 1.19–1.54 for the SIPP targets but 0.86–0.89 for the HRS targets. The HRS planned-missing application indicates that the prespecified $M = 500$ one-step approximation is close to numerical integration, whereas the CPS low-observation design loses root convergence and conditional coverage. Moreover, the HRS utility comparisons do not establish universal dominance over the prespecified alternative estimators.

We organise these checks as four sequential gates:

G1: verify the downstream task mean;

G2: verify path covariance and the finite-replica budget;

G3: propagate seeds, refitting, and root uncertainty;

G4: state the probability law under which the result was validated.

A lower-gate pass remains informative when a higher gate fails, but it cannot be promoted to the higher claim.

2 Related literature and contribution

2.1 From plausible completed data to analysis-specific utility

Multiple imputation provides the canonical framework for propagating uncertainty from missing survey values to downstream estimates [Rubin, 1987]. Its repeated-imputation rules combine average complete-data variance with between-imputation variation, but their justification depends on how the imputation and analysis procedures relate. Meng [1994] formalizes congeniality and shows why an imputer's model and an analyst's procedure can supply incompatible inputs. Robins and Wang [2000] consequently develop variance estimation that remains valid for a broader class of imputation estimators when imputation and analysis models are misspecified or incompatible. This literature establishes a point central to our setting: a completed file is not an inferential end product, and its validity cannot be assessed without reference to the analysis that consumes it.

Early imputation diagnostics often concentrate on the generated records themselves. Abayomi et al. [2008] propose examining completed-data displays, comparing observed and imputed distributions, and checking observed data against the imputation model. Bondarenko and Raghunathan [2016] develop graphical and numerical comparisons conditional on response propensities, including tools that an analyst can use without full knowledge of the producer's imputation model. Such diagnostics can expose implausible tails, conditional distributions, or response strata. They remain valuable in our workflow, but a generated distribution can pass them and still distort a nonlinear target or an estimating-equation component.

Analysis-specific checking addresses that limitation. He and Zaslavsky [2012] apply target analyses to posterior replicates of completed data and compare the resulting estimands rather than relying only on generic discrepancies. In the synthetic-data literature, Snoke et al. [2018] draw a related distinction between general utility—similarity of the original and synthetic joint distributions—and specific utility, defined by agreement for analyses of interest. Partially and fully synthetic releases also require inferential rules that differ from ordinary missing-data MI because what is conditioned on and generated is

different [Raghunathan et al., 2003, Reiter, 2003, Reiter and Raghunathan, 2007]. These results motivate our first gate. We begin with the task mean of the score actually used downstream; distributional fidelity or prediction accuracy is supporting evidence, not a substitute for that check.

Our focus differs from a broad synthetic-data utility ranking. We do not seek a scalar quality score that orders generators across unspecified future analyses. Instead, we ask whether a stated generator supports a stated survey target under a stated masking and refitting law. This deliberately narrower formulation makes a failure actionable: it points first to the task model or estimand, rather than prematurely changing the replica budget or standard-error formula.

Specific utility itself is not sufficient for the inferential claim we study. Agreement of a generated-data point estimate with its confidential or complete-data counterpart can occur even when repeated generation contributes the wrong amount of variation, or when the comparison conditions on a nuisance fit that would change in a new sample. Conversely, a generator can add visible Monte Carlo noise at small M while remaining correctly centred and yielding calibrated uncertainty once that noise is included. We therefore separate point-target fidelity from finite-replica and repeated-sampling calibration. This separation also clarifies reporting: an analysis may legitimately pass the task-mean gate while failing the coverage gate, without treating the former evidence as worthless or the latter failure as a contradiction.

2.2 Finite replicas and Monte Carlo reproducibility

The number of imputations or synthetic releases has long been recognized as a computational and inferential choice. Graham et al. [2007] show that more than the traditional handful of imputations may be required for practically equivalent inference, and Bodner [2008] demonstrates that p-values, confidence-interval widths, and estimated fractions of missing information can be unstable across repeated MI runs when M is small. Practical guidance for chained-equation imputation therefore treats convergence and the number of imputations as substantive parts of the analysis [White et al., 2011]. von Hippel [2020]

sharpens the point by distinguishing efficient point estimates from reproducible standard-error estimates: the latter can require a number of imputations that grows quadratically with the fraction of missing information. Small-sample reference distributions add another finite-M consideration [Barnard and Rubin, 1999].

This work typically studies Monte Carlo error in a pooled estimate or its estimated standard error. Our setting exposes an additional layer because each replica contributes multiple stochastic score components. The components may share a nuisance fit, household draw, or random path, so their covariance contributes to the $1/M$ term. The resulting issue is not resolved by citing asymptotic efficiency or by reporting seed stability for the final scalar alone. One must retain the component structure, estimate its covariance under the actual path law, and verify that the observed variance across several values of M follows the predicted slope.

We therefore treat M as a design parameter chosen only after task-mean recovery and path dependence have been established. The exact decomposition used below is elementary, and we do not present the $1/M$ rate itself as novel. The contribution is to convert it into an auditable diagnostic: report the two component variances and their covariance, compare joint and independent path laws without presuming an ordering, and confirm the algebra with actual finite-M roots rather than projections alone.

2.3 Generated nuisance functions and full-refit inference

The need to propagate first-stage estimation is older than modern imputation software. In regressions with generated regressors, Pagan [1984] shows that treating an estimated input as observed can invalidate conventional uncertainty calculations. Murphy and Topel [1985] derive corrections for two-step estimators in which an auxiliary model supplies an unobserved regressor, emphasizing that second-stage covariance calculations must include first-stage sampling error. Modern survey workflows reproduce this structure whenever estimated response propensities, imputed outcomes, or predicted latent scores enter a final estimating equation.

Orthogonal estimation reduces, but does not erase, the need for careful validation.

Chernozhukov et al. [2018] combine Neyman-orthogonal scores with cross-fitting so that regularization and overfitting in high-dimensional nuisance learners have a smaller first-order effect on a low-dimensional target. Double robustness in survey nonresponse problems similarly combines response-propensity and outcome-regression models so that consistency can survive misspecification of one component [Kim and Haziza, 2014]. These protections concern sensitivity to nuisance error under specified regularity conditions. They do not guarantee that a finite generated-path implementation has regular roots, that its component covariance has been estimated correctly, or that a fixed-learner bootstrap represents the repeated use of the full procedure.

The ordering of resampling and generation is particularly consequential under uncongeniality. Bartlett and Hughes [2020] show that imputation followed by bootstrapping is generally not valid under uncongeniality or misspecification, whereas appropriate bootstrap-then-impute procedures can be. von Hippel and Bartlett [2021] likewise use bootstrap samples followed by repeated imputations to separate sampling and imputation variation. Our outer experiment follows this full-pipeline logic. In every outer repetition we regenerate the researcher mask, refit nuisance learners, reconstruct all stochastic paths, and solve the final estimating equation. The comparison between analytic standard errors and the empirical distribution of those roots is therefore a check on the implemented inferential procedure, not only on a frozen prediction model.

2.4 Survey nonresponse and dependent generation

Survey weighting supplies the substantive context for the nuisance functions used here. Nonresponse weighting is not merely a mechanical exchange of lower bias for higher variance. Little and Vartivarian [2005] show that an adjustment variable's association with the survey outcome is central and that weighting can sometimes reduce variance as well as bias. The broader weighting process also includes base weights, nonresponse adjustment, calibration to known totals, and sometimes trimming or smoothing [Haziza and Beaumont, 2017]. We therefore distinguish the public surveys' design and production weights from the response propensities and experimental weights constructed for our validation tasks.

A good response classifier alone is not the objective; its usefulness depends on the target outcome and on overlap.

Recent SRM studies illustrate the empirical depth required to evaluate missing-data and learned survey procedures. Axenfeld et al. [2024] combine planned missingness and item nonresponse in real German Internet Panel data and evaluate imputation over a large Monte Carlo design. Lipps and Kuhn [2023] use Swiss Household Panel data and an external register-linked source to construct realistic longitudinal nonresponse mechanisms, assessing cross-sectional statistics, longitudinal dependence, and applied regression targets. Walraad et al. [2024] compare machine-learning and regression nuisance models in linked survey, sensor, and register data, separately examining prediction, point estimation, variance, and stability across years. These papers reinforce two choices in our design: real survey records provide the dependence and leverage structure that stylized simulation may miss, and evaluation should include the final survey estimand rather than stop at nuisance prediction.

Dependence can also be created by the generation design itself. Nested and two-stage multiple-imputation methods distinguish variability arising at different levels of an expensive and inexpensive generation process [Rubin, 2003, Harel, 2007]. More generally, Xie and Meng [2017] view data collection, preprocessing, imputation, and analysis as multiple inferential phases that carry different information and assumptions. This perspective explains why independently generated rows, jointly generated household paths, and replicas that share a fitted learner need not have the same covariance structure even when their marginal task means agree.

The unresolved practical gap is thus not the absence of any one diagnostic. The literature provides distributional checks, target-analysis checks, finite-imputation guidance, generated-regressor corrections, orthogonal scores, bootstrap ordering results, survey weighting theory, and nested-generation rules. What is often missing in an applied survey analysis is a stopping rule that orders these checks and prevents evidence from one probability law or inferential layer from being used for another. Our four-gate workflow fills that operational gap. It links task-specific utility to component covariance, finite-M

behaviour, full-refit root inference, and an explicit claim law; it then tests the links in three survey systems chosen to reveal both successful calibration and failure boundaries.

Section 3 defines the four objects and minimal identities. Section 4 summarises the three survey systems and the unified benchmark; detailed data construction is in Supplementary Sections S2–S4. Section 4.3 defines estimation and performance criteria. Section 5 reports SIPP, HRS, and CPS results. Section 6 concludes.

3 Inferential framework and validation gates

3.1 Survey clusters, latent components, and estimands

Let $h = 1, \ldots, H$ index survey households or longitudinal household clusters and let $i = 1, \ldots, n_h$ index eligible records within cluster h, with $N = \sum_h n_h$. The always-observed field is W_{hi}, the downstream outcome is Y_{hi}, and Z_{hi} is a component needed by the analysis but unavailable on some records. The indicator $R_{hi} = 1$ means that Z_{hi} is observed. In the controlled SIPP, HRS, and CPS experiments, R_{hi} is induced by a researcher-generated household arm. In the observed HRS application, $R_{hi} = S_{hi}J_{hi}$, where S_{hi} records planned half-sample assignment and J_{hi} records whether a loneliness score can be constructed within the assigned arm; questionnaire eligibility and return are reported separately in the sample flow. This distinction matters because the first design has a known conditional probability law, whereas the second requires a maintained response model.

Write $\psi_{hi}(\theta; Z_{hi})$ for the complete-record estimating contribution and

$$G_F(\theta) = N^{-1} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \psi_{hi}(\theta; Z_{hi}).$$

For a fixed completed validation panel $\mathcal{D}^\dagger$, the reference estimand θ_F is the local solution of $G_F(\theta_F) = 0$. It is therefore a finite-panel coefficient or functional, not automatically a population parameter. SIPP's two θ_F's are completed-panel housing functionals; the HRS and CPS values are coefficients of completed-panel logistic projections. In an

actual-missingness application, the analogous target is instead defined under the stated model-based law. We keep these targets distinct even when the same estimating equation is used.

A fitted generator produces replica $Z_{hi}^{*(b)}$, $b = 1, \ldots, M$. What matters for inference is not the record alone but its image under the downstream task. For a scalar functional that image may be $q(Z_{hi}^{*(b)}, W_{hi})$; for a coefficient it is the score $\psi_{hi}(\theta; Z_{hi}^{*(b)})$ and, where needed, its derivative. Conditional on the information $\mathcal{D}$ used to fit the generator, define the generic task objects

$$T_b = q(Z^{*(b)}; \mathcal{D}), \qquad m_{\mathcal{D}} = \mathbb{E}(T_b \mid \mathcal{D}), \qquad K_{\mathcal{D}} = \mathrm{Var}(T_b \mid \mathcal{D}).$$

The mean locates the generated task, while $K_{\mathcal{D}}$ governs fresh finite-replica noise. Both depend on the target, masking design, cluster aggregation, and path law. They are not generic measures of how realistic a completed file appears.

This distinction is concrete in the applications. In SIPP, a path model for log home value and log primary-residence mortgage-and-loan balance is evaluated through a threshold indicator and a capped ratio; accurate marginal levels do not imply an accurate mean for either nonlinear transformation. In HRS and CPS, the generated component enters a logistic score, so the relevant conditional object changes with the coefficient at which the score is evaluated. Task calibration consequently means matching the conditional functional that appears in the estimating equation, not declaring the entire fitted conditional distribution correct. It can support the named analysis while leaving other summaries of the generated records unvalidated.

3.2 Oracle and feasible estimating equations

The finite-M conditional task field is

$$\widehat{m}_{hi,M}(\theta) = M^{-1} \sum_{b=1}^{M} \psi_{hi}\{\theta; Z_{hi}^{*(b)}\}, \qquad \widehat{m}_{hi,\infty}(\theta) = \mathbb{E}_{\widehat{Q}}\{\psi_{hi}(\theta; Z_{hi}^{*}) \mid W_{hi}, Y_{hi}\}.$$

Here $\widehat{Q}$ is fitted only on training clusters. A schematic orthogonal contribution is

$$\widehat{\phi}_{hi,M}(\theta) = \widehat{m}_{hi,M}(\theta) + \frac{R_{hi}}{\widehat{\pi}_{hi}}\{\psi_{hi}(\theta; Z_{hi}) - \widehat{m}_{hi,M}(\theta)\}, \tag{1}$$

and the feasible estimate solves $N^{-1}\sum_{h,i}\widehat{\phi}_{hi,M}(\theta) = 0$. Under researcher masking, $\widehat{\pi}_{hi}$ is replaced by the known household visibility probability. In the HRS application it is a cross-fitted observation propensity under the maintained score-CAR interpretation. Equation (1) is an architectural template: it explains the common comparison without claiming that the three surveys share a population or sampling design.

The benchmarks isolate different possible failures. The complete-panel root uses G_F. Complete-case or known-IPW estimation uses only observed complete scores, with or without known inverse visibility weights. Deterministic substitution evaluates a nonlinear score at a predicted conditional mean of Z, which generally differs from integrating the score. A nonorthogonal generated-score estimator averages observed and generated scores without the augmentation in (1). The orthogonal estimator combines the conditional task field with the observed residual correction. Agreement among these methods is an empirical result, not a defining property of generated integration.

Cross-fitting is performed at the household level: no household appears in both the training and evaluation part of a fold. Task, path, and propensity learners are estimated without the hidden target component in the evaluation fold. More importantly, every outer remasking repetition refits the declared pipeline rather than reusing a single favourable learner. This full-refit convention incorporates the interaction between a new mask, nuisance estimates, and the local estimating root. It does not prove that nuisance estimation is negligible; it makes that source of variation visible to the validation experiment.

The term orthogonal refers to local insensitivity of the population estimating equation to first-order perturbations of the task and propensity nuisances under the maintained conditions. It is not a synonym for unbiasedness, robustness to every learner, or efficiency. Cross-fitting limits own-observation reuse, while the augmentation in (1) can cancel particular first-order nuisance errors. Neither device guarantees adequate overlap, a unique

regular root, or correct response-model scope. Those properties are checked separately at G3 and G4. This separation also explains why deterministic and nonorthogonal comparators remain informative: they reveal whether an apparent gain comes from integrating the nonlinear task, from the residual correction, or merely from holding fitted quantities fixed.

3.3 Paired uncertainty and the finite-replica component

In a completed-panel remasking experiment, feasible and complete scores are evaluated on the same fixed records. Let

$$d_h(\theta) = \sum_{i=1}^{n_h} \{\widehat{\phi}_{hi,M}(\theta) - \psi_{hi}(\theta; Z_{hi})\}$$

be the paired household difference and let $\widehat{J}_M$ be the derivative of the feasible mean score at its root. With $\bar{d} = H^{-1} \sum_h d_h$, the paired household sandwich has the form

$$\widehat{V}_{\text{pair},M} = \widehat{J}_M^{-1} \left\{ \frac{H}{H-1} \frac{1}{N^2} \sum_{h=1}^{H} (d_h - \bar{d})(d_h - \bar{d})' \right\} \widehat{J}_M^{-\prime}. \tag{2}$$

Pairing removes completed-panel score variation common to the feasible and reference roots; household aggregation preserves dependence within a SIPP household, an HRS longitudinal household, or a CPS household. The resulting interval is validated against repeated researcher masks conditional on $\mathcal{D}^\dagger$. It is not a substitute for a survey-design variance estimator. In the actual HRS application, where an unobserved complete-panel reference is unavailable, inference uses the feasible household score under the maintained independent-household model rather than the paired validation difference.

The law of total variance clarifies the role of M. Conditional on the fixed panel, write A for the outer household mask and include in $\widehat{\theta}_\infty(A)$ every nuisance refit and root calculation. Let K_A denote the covariance of one fresh, household-aggregated path-score perturbation on the mean-score scale for that fitted mask. For locally regular roots and

fresh conditionally independent paths,

$$\mathrm{Var}\{\widehat{\theta}_M - \theta_F \mid \mathcal{D}^\dagger\} \approx \mathrm{Var}_A\{\widehat{\theta}_\infty(A) - \theta_F\} + \mathbb{E}_A\left[J_A^{-1}\frac{K_A}{M}J_A^{-\prime}\right]. \tag{3}$$

The first term contains mask dispersion, learner refitting, and their covariance; it should not be decomposed into independent pieces without evidence. The second is incremental integration noise at the task-score scale. In SIPP and HRS, joint and independent paths are calibrated to the same task mean, so their finite-M contrast isolates different K_A's. In CPS, the reported finite-M comparison is limited to a one-mask projection and a small exact-root check, not a second outer experiment. Increasing M reduces the second term in (3); it cannot repair an unstable Jacobian, poor overlap, task-mean bias, or a misspecified response law.

At the household level, K_A contains both recordwise variances and cross-record score products. A shared household draw can make those products positive or negative after multiplication by task-specific score weights. Consequently, preserving within-household dependence need not always increase or always decrease finite-replica variance. The opposite SIPP and HRS orderings are therefore evidence that K is a generator–design–task object, not a global quality ranking of joint and independent generators. Practical choice of M should compare the resulting root-scale K/M contribution with the first term of (3), rather than count generated records as if they enlarged N.

3.4 Four gates and supported conclusions

The gates are sequential because each addresses a different non-implication. A lower-gate pass remains useful when a later gate fails, but it supports only the statement in its own row.

G1 therefore precedes any discussion of covariance: a precise simulation around the wrong task mean is not useful. G2 is task-specific, and its path ordering can reverse, as it does between the SIPP housing targets and the HRS loneliness projections. G3 asks whether the favourable calculation survives the pipeline an analyst would actually rerun;

Table 1: Failure, diagnostic, repair, and claim at each validation gate

Gate	Failure mode	Primary diagnostic	Repair after failure	Supported conclusion after a pass
G1: task mean	Generated records look plausible but miss the downstream score, threshold, cap, or derivative	Held-out task and derivative loss; infinite-M bias against a completed reference	Recalibrate the task field, revise predictors or support, or narrow the estimand	Task-mean recovery for the named estimand under the active law
G2: path covariance and M	Equal means conceal different within-cluster covariance; M is too small or paths share a shock	Joint–independent K; $M\,\mathrm{Var}(\widehat{\theta}_M - \widehat{\theta}_\infty)$; log-variance slope	Change the path coupling, remove unintended shared seeds, or increase M to a declared budget	Finite-replica precision for the tested task and path law
G3: refit and root	A frozen learner understates variation, or overlap and the local Jacobian fail	Full-refit dispersion, paired SE/SD and coverage, convergence, condition number, seed audit	Refit the full pipeline, stabilize the score or nuisance fit, change support, or report failure	Conditional full-pipeline inference under the repeated experiment
G4: probability law	A fixed-panel result is promoted to actual nonresponse, survey sampling, or a new population	Explicit statement of what is fixed and repeated; response and sensitivity analysis where applicable	Add the required response/design experiment or restrict the wording	Validity only under the law that was actually evaluated

the CPS weak-information cell shows why positivity alone is insufficient. G4 is a reporting restriction as much as a statistical check. Researcher remasking of a completed panel, a model-based actual-missingness analysis, and repeated survey sampling answer different questions.

3.5 Minimal covariance identities

Proposition 1 (Independent-replica identity)**.** *If $T_1, \ldots, T_M$ are conditionally independent and identically distributed given $\mathcal{D}$, with finite conditional covariance $K_{\mathcal{D}}$, then*

$$\mathrm{Var}\left(\frac{1}{M}\sum_{b=1}^{M} T_b \,\middle|\, \mathcal{D}\right) = \frac{K_{\mathcal{D}}}{M}.$$

This exact identity explains why the one-path covariance must be attached to a task and path law before M is chosen. It is not a new asymptotic result.

Example 1 (Same mean, different covariance)**.** *Let $T_b^{(A)} = m_{\mathcal{D}} + \varepsilon_b$ and $T_b^{(B)} = m_{\mathcal{D}} + A\varepsilon_b$, with $\mathbb{E}(\varepsilon_b \mid \mathcal{D}) = 0$. Both laws pass the same G1 target, while their G2 objects are $K_{\mathcal{D}}$ and*

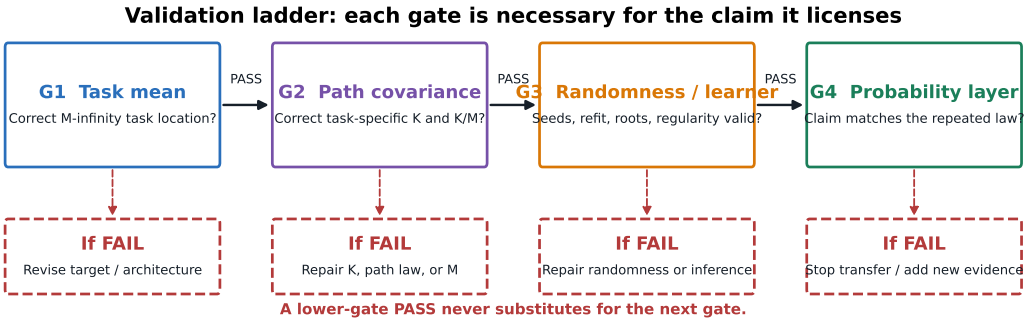

Figure 1: Four-gate workflow for generated latent quantities. A lower-gate pass supports only its own conclusion. Each failure leads to a different repair rather than to a single aggregate score.
Alt text: Flow diagram of four sequential validation gates. The gates assess the task mean, path covariance and replica count, refitting and root uncertainty, and the probability law; each failed gate points to a distinct repair.

$AK_{\mathcal{D}}A'$.

Proposition 2 (Shared-shock plateau)**.** *Suppose* $T_b = m_{\mathcal{D}} + U + \varepsilon_b$*, where* U *is shared by every replica and the* ε_b*'s are conditionally independent of one another and of* U*. With* $A_{\mathcal{D}} = \mathrm{Var}(U \mid \mathcal{D})$ *and* $B_{\mathcal{D}} = \mathrm{Var}(\varepsilon_b \mid \mathcal{D})$*,*

$$\mathrm{Var}(\bar{T}_M \mid \mathcal{D}) = A_{\mathcal{D}} + B_{\mathcal{D}}/M.$$

If $A_{\mathcal{D}} \neq 0$*, increasing* M *cannot remove the shared component.*

Supplementary Section S1 gives proofs. Supplementary Sections S2–S4 document SIPP, HRS, and CPS, respectively; Supplementary Section S5 reports implementation diagnostics and HRS sensitivity analyses. The identities are diagnostic bookkeeping, not claims of priority for finite-M Monte Carlo error.

4 Survey data, masking designs, and estimation

The empirical benchmark uses three survey systems with deliberately different roles. SIPP supplies a nonlinear household-balance-sheet problem with two latent quantities that must be generated together. HRS supplies both a completed-responder remasking experiment and an application to a genuine planned-missing psychosocial measure. CPS

supplies a weak-information stress test in which the same algorithm is exposed to less observed information. The systems are not pooled, and their survey weights, missing-data mechanisms, and generated-path laws are never treated as interchangeable. Instead, each system answers a different part of the validation question: whether task means recover nonlinear targets, whether within-cluster path covariance affects finite-replica uncertainty, whether a selected finite M works after complete refitting, and where the procedure fails.

Three probability mechanisms must be distinguished from the outset. First, each agency drew its public-use survey sample under its own complex design. Second, HRS psychosocial assignment is part of the survey design, while constructibility of an observed loneliness score among assigned respondents forms an additional response process. Third, the SIPP, HRS completed-panel, and CPS earnings validation experiments impose new researcher-generated masks on completed records. The repeated-mask results are conditional on those fixed completed arrays. Consequently, a 95% inclusion rate over researcher masks is neither a claim of national survey coverage nor evidence that an agency's actual allocation or nonresponse mechanism is ignorable.

4.1 Data overview

The analyses use the 2023 SIPP public-use file [U.S. Census Bureau, 2024a,b], the 2006–2018 HRS core and psychosocial products [Health and Retirement Study, 2026], and the January 2025 Basic CPS file [U.S. Census Bureau, 2025, 2019]. Table 2 summarises the source, analysis unit, fixed sample, downstream target, and empirical role of each system. SIPP provides nonlinear household housing functionals; HRS provides completed-responder remasking and an observed planned-missing application; and CPS provides a completed direct-earnings benchmark and a weak-information stress design. Detailed source-file construction, field definitions, exclusions, sample-flow reconciliations, scale coding, adjustment fields, and survey-specific generator and masking specifications are given in Supplementary Sections S2–S4. Supplementary Section S5 records numerical implementation and sensitivity checks.

Table 2: Survey data, analysis samples, and downstream targets

System	Data and unit	Analysis sample	Downstream target	Generated latent quantity and empirical role
SIPP	2023 public-use file; December household copy	5,219 direct-report mortgaged, non-mobile owner households	High leverage, $\Pr(D/V > .80)$; capped debt-to-value, $E\{\min(D/V, 2)\}$	Joint home value and balance on the first three mortgages and loans; finite-M validation
HRS remasking	2006–2018 core and psychosocial person-waves; longitudinal-household clusters	Work: 26,485 waves; health: 26,473; 10,697 households	Raw-loneliness coefficient in logistic projections of current paid work and poor/fair health	Three-item loneliness scale; conditional full-refit validation
HRS application	Same source panel; designed assignment and observed score constructibility	Work: 72,890 risk-set waves in 14,700 households; health: 72,843 in 14,696	Same descriptive logistic projections	Planned-missing loneliness under maintained score-CAR
CPS stress test	January 2025 Basic CPS; household-clustered workers	7,085 workers; 4,801 households	Raw coefficient on log(1+ weekly earnings) in a multiple-jobholding logit	Researcher-hidden direct capped earnings; regular and weak-information conditions

Notes: V is reported primary-residence value and D is the reported principal owed on the first three mortgages and loans against that residence. Sample sizes are fixed-panel denominators, not population totals. All coefficients and housing targets are descriptive.

4.2 Masking experiments, estimators, and comparisons

Table 3: Unified benchmark design and what is repeated

System	Mask and repeated unit	Learner and refit policy	Replications and M	Primary comparison
SIPP	Bernoulli(.5) block visibility per household	Five household folds; task and path learners fully refit at each mask	B = 1,000; $M = 1, 5, 20, 100, 500, \infty$	Joint versus independent residual paths
HRS remasking	Bernoulli(.5) visibility for all waves in a longitudinal household	Five household folds; task/path learner and root fully refit	B = 1,000 per target; actual $M = 20, 100, 500, \infty$	Shared-household versus row-independent paths
HRS application	Actual $R = SJ$ observation pattern	Cross-fitted task, path, and score-observation-propensity learners	$M = \infty$ root; M = 500 one-step approximation	Complete case, IPW, plug-in, generated, orthogonal
CPS stress test	Researcher visibility $p = .50, .25$ by household	Five outer and five inner household folds; full refit	B = 1,000 retained masks per p; $M = \infty$ inference	Regular versus weak-information roots

Notes: Each generated path is a numerical-integration draw, not an additional observation. The HRS completed-panel experiment uses one nested path batch at every outer mask, with prefixes at $M = 20, 100, 500$.

The common experiment fixes the completed panel and its oracle target, draws a cluster-level visibility indicator, refits the learner without using the hidden latent field, constructs the infinite- or finite-path estimating root, and compares it with the fixed completed-panel root. Cluster folds, seeds, M, diagnostic thresholds, and the rule for retaining failures are specified before examining the repeated-mask outcomes and then held fixed. Preprocessing is learned within training folds. The sampling unit for validation is a mask on a fixed panel, not another draw from the survey population; repeated person-waves or workers within a household remain together in folds, masks, paths, and covariance sums.

The principal orthogonal estimator combines an integrated conditional score with an inverse-probability residual correction from visible records. Its comparators are chosen to identify failure modes, not to create a tournament of unrelated prediction algorithms. Complete case uses only visible records. Known-design IPW uses the researcher visibility probability in the controlled experiments. Estimated-propensity IPW in the HRS observed-pattern application uses $.5q_J(H)$. Conditional-mean substitution inserts the predicted latent mean into the nonlinear score. Nonorthogonal generated integration averages the generated score without the residual correction. IPW is the reference method for the utility comparison: validity of generated integration and improvement in root mean-squared error are reported separately.

SIPP solves a mean-functional problem, whereas HRS and CPS solve multivariable logistic estimating equations. In all cases, generated paths are averaged inside a record before records are summed. An analysis with M paths therefore still has the original number of survey units; it does not have MN observations. Paired cluster scores compare each feasible estimate with its completed-panel oracle under the same mask. HRS and CPS covariance matrices aggregate these differences by longitudinal or monthly household. No bootstrap is used as the main uncertainty calculation.

4.3 Evaluation criteria and finite-path diagnostics

For target θ_F, outer replication r, and finite-path estimate $\widehat{\theta}_{rM}$, we report bias, empirical standard deviation, root mean squared error, mean reported standard error, the ratio of mean standard error to empirical standard deviation, and 95% and 80% coverage. No failed root is silently removed. If B is the number of retained outer masks,

$$\widehat{\text{bias}}_M = B^{-1}\sum_{r=1}^{B}(\widehat{\theta}_{rM} - \theta_F), \qquad \widehat{\text{SD}}_M^2 = (B-1)^{-1}\sum_{r=1}^{B}\{\widehat{\theta}_{rM} - \bar{\theta}_M\}^2.$$

Coverage uses the paired cluster standard error attached to the feasible-minus-completed estimating root. It is evaluated directly against the conditional remasking truth rather than inferred from a single standard-error ratio. The ratio of mean reported SE to empirical mask SD diagnoses average scale, while inclusion evaluates the full interval procedure. These can disagree, as the CPS results demonstrate.

The displayed summaries are common, but the diagnostic tolerances are system specific. The HRS completed-panel experiment evaluates absolute bias no larger than .10 empirical SD, mean-SE-to-SD in [.90,1.10], 95% coverage in [.92,.98], 80% coverage in [.76,.84], root convergence of at least .995, and a 99th-percentile Jacobian condition number below 10,000. SIPP uses its own centering, 95% inclusion, and finite-path covariance checks; 80% inclusion was not one of its evaluation criteria. Moreover, the high-leverage shared-residual $M = 1$ absolute bias is .1078 empirical SD, so the HRS .10 cutoff is not imported after the fact to characterize every SIPP cell. CPS evaluates absolute bias no larger than .10 empirical SD, 95% coverage in [.925,.975], 80% coverage in [.76,.84], and root convergence of at least .995, together with its task, quadrature, and finite-path checks. CPS SE/SD and Jacobian condition are reported as scale and stability diagnostics rather than binary criteria. A separate utility comparison asks whether RMSE is at least 5% lower than for IPW. These tolerances apply to the stated conditional experiments; they are not proposed as universal survey standards.

For the HRS completed-panel experiment, the same-mask difference

$$\Delta_{rM} = \widehat{\beta}_{rM} - \widehat{\beta}_{r\infty}$$

isolates the finite-path contribution from most outer mask variation. Under the independent-replica identity, $\mathrm{Var}(\Delta_{rM})$ should be proportional to $1/M$. We therefore report the log variance slope, its R^2, and the stability of $M\,\mathrm{Var}(\Delta_{rM})$. The prespecified criteria are a slope in [-1.15,-.85], $R^2 \geq .95$, and a maximum-to-minimum scaled-variance ratio no larger than 1.50.

SIPP and HRS use both joint and independent paths with identical task means. Their standardised comparison is

$$\rho_K = K_{\text{independent}}/K_{\text{joint}}.$$

A value above one means that preserving the within-cluster shared path reduces task-specific Monte Carlo covariance; a value below one means the shared path increases it.

The HRS finite-path evaluation adds a calculation that is unavailable from a one-mask covariance projection. For every one of 1,000 masks per task, the task learner and $M = \infty$ root are refitted and the same nested path batch supplies prefixes at $M = 20, 100, 500$ under both path laws. This produces 12,000 actual finite-path roots. Computation used an NVIDIA RTX 3090 with PyTorch 2.13.0 and CUDA 12.6. Elementwise path probabilities were evaluated in single precision and score reductions and Newton systems in double precision, with TF32 disabled. All 2,000 analytic $M = \infty$ roots were reproduced to a maximum absolute difference of 9.93×10^{-17}. Eight GPU roots were compared with a CPU reference; the maximum coefficient difference was 3.23×10^{-6}. The diagnostic conclusions were unchanged when evaluated after 800 rather than 1,000 masks.

5 Results

5.1 SIPP: full finite-path recovery

The completed-panel references are .101169 for high leverage and .475462 for capped debt-to-value. Table 4 reports selected rows from the full grid. Bias is small compared with conditional mask dispersion for both tasks. At $M = 1$, finite-path noise is visible: for capped debt-to-value the empirical SD is .00752 under joint paths and .00908 under independent paths. By $M = 500$, both are close to the $M = \infty$ SD of .00344.

Table 4: SIPP full-refit conditional performance over selected finite-M cells

Target	Path, M	Bias	Empirical SD	Mean SE	SE/SD	Coverage .95
High leverage	Joint, 1	-.000777	.007209	.006881	.955	.943
	Joint, 20	-.000297	.004368	.004257	.974	.952
	Joint, 500	-.000236	.004149	.004080	.983	.958
	$M = \infty$	-.000232	.004137	.004072	.984	.958
	Independent, 1	-.000568	.007465	.007294	.977	.939
	Independent, 20	-.000237	.004431	.004291	.968	.953
	Independent, 500	-.000227	.004145	.004081	.985	.959
Capped DTV	Joint, 1	-.000867	.007519	.007477	.994	.948
	Joint, 20	-.000122	.003761	.003709	.986	.951
	Joint, 500	-.000073	.003464	.003410	.984	.945
	$M = \infty$	-.000066	.003437	.003397	.988	.947
	Independent, 1	-.000532	.009078	.008924	.983	.946
	Independent, 20	-.000055	.003991	.003865	.968	.934
	Independent, 500	-.000062	.003451	.003417	.990	.943

Notes: $B = 1{,}000$ household masks in every row. The $M = \infty$ row is common to the two path laws. Coverage is conditional on the fixed completed direct-response panel and researcher mask.

The finite-path covariance is task-specific. The estimated independent-to-joint K ratio is 1.191 for high leverage and 1.536 for capped debt-to-value. Thus the joint path reduces Monte Carlo covariance in these two SIPP tasks, particularly for capped debt-to-value. This is not a generic endorsement of the joint Gaussian learner: the native uncalibrated path score is worse than a constant benchmark for high leverage, and the favourable result appears only after task calibration.

Across the full evaluation grid, conditional 95% coverage ranges from .934 to .959 and SE/SD from .955 to 1.025. The controlled experiment therefore validates recovery under the researcher mask with full refitting. It does not establish that the Census allocation process is ignorable or that the intervals have national design-based coverage.

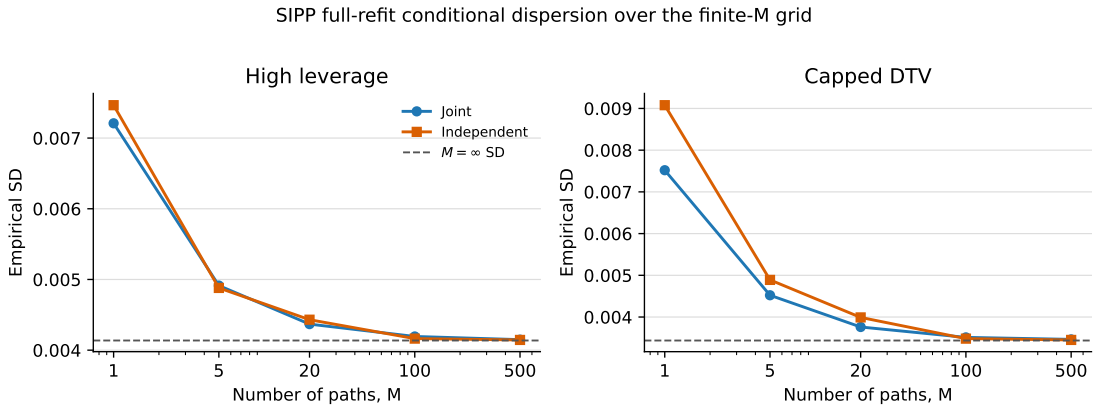

Figure 2: SIPP empirical standard deviation over the finite-M grid. Dashed lines show the $M = \infty$ full-refit dispersion. The larger independent-path covariance for capped debt-to-value is most visible at small M.

Alt text: Line plots of empirical standard deviation against the number of generated paths for two SIPP housing targets. Curves approach the dashed infinite-path limits as M increases, with the largest path-law difference at small M.

The grid also gives an operational reading of M. Under joint paths, the estimated finite-path share of total variance is 1.8% for high leverage and 3.6% for capped DTV at $M = 100$; under independent paths the corresponding shares are 2.1% and 5.5%. At $M = 500$, all four shares are about 1.1% or less. Thus 500 is conservative for these two tasks, but the smaller acceptable budget depends on both the task and path covariance. The conclusion cannot be obtained from the marginal fit of V and D alone: the independent generator preserves those marginals yet requires appreciably more paths for capped DTV.

5.2 HRS: full-refit finite-replica validation

Table 5 reports the HRS completed-panel experiment. All 12,000 finite-M roots converge. Bias is at most .016 empirical standard deviations, SE/SD lies between 1.003 and 1.015, 95% coverage is .938–.953, and 80% coverage is .813–.825. These results are not a K/M projection: every row uses an actual calibrated finite-M root after the task learner has been refitted at the corresponding household mask.

The finite-replica experiment extends, rather than replaces, the analytic $M = \infty$ validation. At $M = \infty$, work bias was -.00018, empirical SD .03414, mean paired SE .03467, and 95% inclusion .951. The health counterparts were -.00058, .03933, .03962, and .941. Held-out task-score MSE was .798 and .781 of the constant-distribution benchmark,

and focal-derivative MSE was .792 and .795. A 250-mask learner-remainder check found remainder SDs equal to 4.3% and 5.2% of the direct paired-score SD. These checks matter because the reported intervals refit the learner; a fixed-nuisance calculation would not address dependence induced by refitting.

Table 5: HRS finite-M full-refit validation under household remasking

Target	Path, M	Bias/SD	Empirical SD	SE/SD	Cov. .95	Cov. .80	$M\,\mathrm{Var}(\Delta_M)$
Work	Shared-household, 20	.0024	.034719	1.014	.950	.816	.000745
	Shared-household, 100	.0061	.034304	1.014	.950	.825	.000755
	Shared-household, 500	.0043	.034168	1.015	.950	.824	.000755
	Row-independent, 20	.0053	.034606	1.015	.953	.819	.000657
	Row-independent, 100	.0098	.034274	1.014	.949	.822	.000652
	Row-independent, 500	.0075	.034242	1.013	.949	.825	.000632
Health	Shared-household, 20	.0122	.040100	1.003	.942	.813	.000973
	Shared-household, 100	.0150	.039458	1.007	.942	.821	.000999
	Shared-household, 500	.0145	.039418	1.006	.939	.823	.000899
	Row-independent, 20	.0065	.040046	1.003	.943	.814	.000856
	Row-independent, 100	.0112	.039532	1.005	.938	.813	.000855
	Row-independent, 500	.0154	.039381	1.007	.938	.820	.000843

Notes: $B = 1{,}000$ full-refit household masks per target. $\Delta_M = \widehat{\beta}_M - \widehat{\beta}_\infty$ within the same mask. The fixed-panel target coefficients are .008352 for work and .064089 for health.

The log variance slopes are -.996 and -1.012 for shared-household and row-independent work paths, and -1.024 and -1.005 for health. Their R^2 values are at least .99946. The maximum-to-minimum ratios of $M\,\mathrm{Var}(\Delta_M)$ are 1.014–1.111. All finite-path law criteria are therefore satisfied.

The covariance direction reverses relative to SIPP. Averaging the scaled variance across M, $K_{\text{independent}}/K_{\text{joint}}$ is .860 for work and .889 for health. The high-precision one-mask projections were .853 and .885, so the absolute differences are only .0075 and .0048. Shared-household paths increase, rather than decrease, task covariance for these two HRS projections. At $M = 500$ the absolute effect on total dispersion is small because mask variation dominates, which separates mechanism evidence from a claim of material efficiency.

Validity also remains separate from improvement over a simpler estimator. In the completed-panel experiment, known-design IPW RMSE was .03369 for work and .03917 for health, compared with .03413 and .03932 for orthogonal integration. The corresponding relative changes, -1.30% and -.39%, do not attain the 5% improvement benchmark.

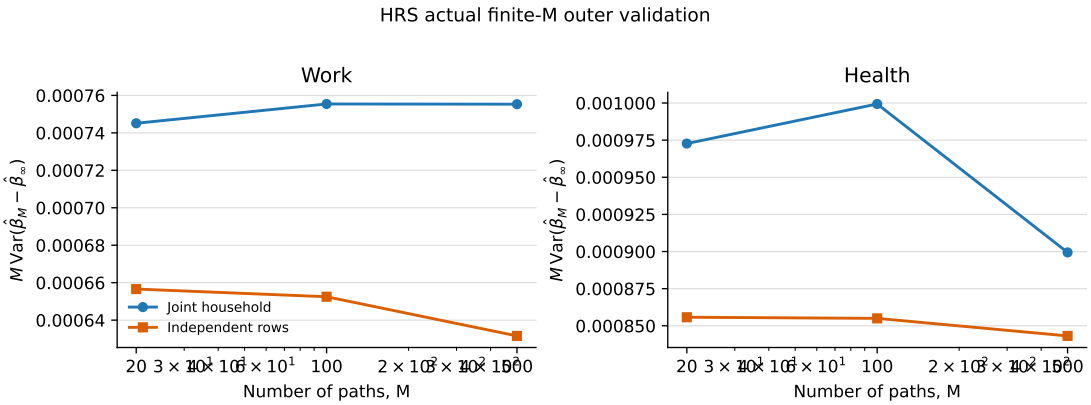

Figure 3: HRS scaled finite-path variance from actual roots. Stability of $M\,\mathrm{Var}(\widehat{\beta}_M - \widehat{\beta}_\infty)$ across $M = 20, 100, 500$ is the empirical counterpart of the $1/M$ identity.
Alt text: Scaled finite-path variances for work and health targets under shared-household and row-independent paths. Values remain nearly horizontal across $M = 20, 100, 500$, supporting inverse-M variance decay.

Generated integration permits evaluation of the conditional nonlinear score and transparent control of finite-path precision; it does not add survey information or guarantee lower RMSE.

5.3 HRS actual planned-missing application

Table 6 translates the conditional validation exercise to the observed HRS $R = SJ$ pattern. Complete-case, inverse-probability-weighted, and orthogonal estimates lead to the same substantive conclusion. For current paid work, the orthogonal $M = \infty$ coefficient is -.00169 (SE .03500); for poor or fair health it is .07573 (SE .04013). The $M = 500$ one-step approximations differ from $M = \infty$ by -.00052 and .00032, respectively—less than .015 baseline standard errors. These one-step rows are not exact replayed finite-M roots. The separate completed-panel full-refit experiment supports the numerical adequacy of $M = 500$, but neither calculation turns the imprecise application estimates into evidence of an effect.

The method contrasts explain why a generated-value point estimate is not sufficient. Conditional-mean substitution gives .02090 for work and .10716 for health. Nonorthogonal generated integration gives .00912 and .04669 with much smaller fixed-score standard errors of .01476 and .01663. Those standard errors do not include the same first-order

Table 6: HRS application under designed half-sample assignment and observed-score constructibility

Target	Estimator	Estimate	Household SE	CI low	CI high
Work	Complete case	.008352	.034219	-.058717	.075422
	Survey-weighted complete case	.019865	.044247	-.066860	.106590
	Inverse-probability weighting	-.003228	.034843	-.071521	.065065
	Conditional-mean substitution	.020901	.033821	-.045387	.087190
	Nonorthogonal generated integration	.009124	.014762	-.019811	.038058
	Orthogonal integration, $M = \infty$	-.001689	.034998	-.070284	.066907
	Orthogonal integration, $M = 500$ one-step	-.002204	.034982	-.070768	.066361
Poor health	Complete case	.064089	.039364	-.013064	.141242
	Survey-weighted complete case	.082351	.052319	-.020195	.184896
	Inverse-probability weighting	.071260	.039770	-.006689	.149209
	Conditional-mean substitution	.107159	.038003	.032672	.181645
	Nonorthogonal generated integration	.046688	.016626	.014100	.079275
	Orthogonal integration, $M = \infty$	.075727	.040133	-.002933	.154387
	Orthogonal integration, $M = 500$ one-step	.076048	.040188	-.002720	.154816

Notes: Coefficients are on raw loneliness. Standard errors aggregate estimating contributions by canonical longitudinal household. The nonorthogonal rows use fixed-score standard errors that omit the first-order learner and residual-correction uncertainty included in the orthogonal analysis. The two $M = 500$ rows are one-step approximations around the $M = \infty$ solution, not exact replayed finite-M roots. Official psychosocial weights are a complete-case sensitivity, not the main target or a national design-based analysis.

learner and residual-correction architecture and are not interpreted as valid substitutes for the orthogonal household uncertainty. A stacked one-standard-deviation marginal-risk functional, which estimates the latent loneliness mean and variance jointly with the regression, is -.00016 (SE .00341) for work and .00522 (SE .00277) for health; both 95% intervals include zero.

The estimated observation propensities show usable overlap rather than identification by decree. Their first percentiles are .174 and .176, and IPW effective-sample-size ratios relative to observed records are .947 for both tasks; score constructibility among assigned records is approximately .723. The pattern-mixture sensitivity shifts the unobserved loneliness distribution by $\delta = -.50, -.25, 0, .25, .50$ conditional SD units. The maximum coefficient shift is .255 baseline SE for work and .081 for health. Replacing the main score-observation propensity model by an earlier alternative specification shifts work by .434 SE and health by .092 SE. The application is therefore more sensitive to the work propensity specification than to the examined local response tilt. Neither exercise tests

score-CAR, and the tilt grid is a sensitivity set rather than a confidence region.

5.4 CPS regular and weak-information boundary

Table 7 reports the CPS orthogonal $M = \infty$ roots against the fixed completed-array coefficient -.101927. All masks, including failed roots, remain in the all-mask summaries. This adverse-retention rule matters at $p = .25$: summarizing only the roots that happen to converge would change the question from performance of the specified procedure to performance conditional on numerical survival.

Table 7: CPS full-refit boundary experiment

Visibility	Bias	Empirical SD	Mean SE	SE/SD	Cov. .95	Cov. .80	Convergence	RMSE gain
$p = .50$	.005604	.074538	.058865	.790	.942	.806	.996	-.188
$p = .25$	.025665	.305714	.134084	.439	.897	.730	.891	-1.556

Notes: Bias is relative to the completed-array raw log-earnings coefficient -.101927. Visibility is a researcher-assigned household probability, not the CPS outgoing-rotation share. RMSE gain is $1 - \text{RMSE}_{orth}/\text{RMSE}_{IPW}$; negative values favour IPW.

The CPS experiment supplies the negative control that the favourable SIPP and HRS results require. At $p = .50$, 996 of 1,000 roots converge. Over all masks, conditional 95% and 80% inclusion are .942 and .806, and the stated bias, coverage, and convergence criteria are satisfied. The all-mask SE/SD ratio of .790 remains an important scale warning; among converged roots it is .973 and 95% inclusion is .945. The orthogonal RMSE is .07471 versus .06288 for known-design IPW, so orthogonal integration does not attain the separate 5% RMSE-improvement benchmark.

At $p = .25$, only 891 roots converge, SE/SD falls to .439, and all-mask coverage falls to .897 at 95% and .730 at 80%. Jacobian instability is severe in the adverse roots. The failure is not caused by too few generated paths: it occurs at the analytic integrated score after full nuisance refitting. Nor is positivity by itself enough; a known .25 observation probability does not guarantee a stable finite-sample 26-parameter root built on a 24-column adjustment span. Shared-household and row-independent finite-path projections have similar K values, and ten exact $M = 200$ check roots agree with their local increments to within 7.28×10^{-6}, but that numerical identity cannot repair the outer root failure.

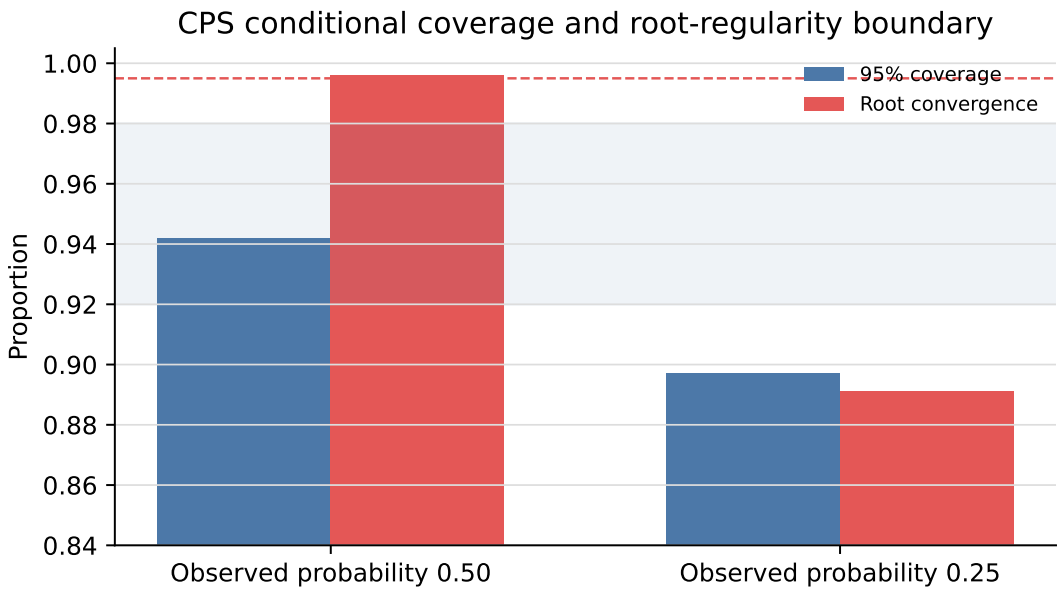

Figure 4: CPS boundary experiment. The vertical reference lines mark nominal 80% and 95% coverage. Lower information at the .25 household-visibility probability produces nonconvergence and substantial standard-error underestimation.

Alt text: Comparison of CPS validation outcomes at household visibility probabilities .50 and .25. The .25 condition has lower coverage, more failed roots, and stronger standard-error underestimation than the .50 condition.

5.5 Cross-system synthesis

Three findings survive the move from algebra to full refitting. First, the finite-path component follows the predicted $1/M$ rate when the fitted root is locally regular. Second, the sign and size of the joint-versus-independent covariance contrast depend on the target: joint paths reduce K in the SIPP tasks but increase it in the HRS tasks. Third, successful finite-path recovery is weaker than satisfactory survey inference. SIPP validates a controlled conditional experiment; HRS validates both actual finite-M roots and numerical implementation but yields imprecise substantive estimates; CPS shows that inadequate information can defeat the root and its variance estimate.

The sequence of samples is as informative as the numerical comparison. SIPP begins with a complete direct-report block and asks what would happen if that entire block were hidden. The HRS completed-responder experiment begins with constructible psychosocial scores and hides all waves within a household, whereas the observed-pattern application restores the actual separation between randomised assignment and observed-score constructibility; questionnaire return is retained as a descriptive sample-flow measure.

CPS uses actual outgoing-rotation records only to form a complete benchmark and then applies an independent researcher mask. This progression prevents a common transfer error: evidence under a controlled mask can validate an algorithm, but it cannot by itself identify an actual response mechanism.

These comparisons also clarify what should be reported. A single pooled predictive score cannot establish task validity. The relevant evidence is target-specific calibration, positivity, full-refit convergence, empirical-to-estimated uncertainty, and the incremental finite-path covariance. Increasing M addresses only the last item.

Cross-system conclusions are kept separate rather than averaging unlike metrics or imposing one system's cutoffs on another. At $p = .50$, CPS satisfies its stated bias, coverage, and convergence tolerances despite the all-mask SE/SD warning of .790, while its RMSE is higher than that of IPW. At $p = .25$, CPS root convergence and coverage are inadequate. Its finite-M evidence remains a projection check rather than an actual outer validation experiment.

6 Conclusion

Generated quantities must be validated through the complete map from released records to the downstream estimator. The proposed ladder separates exact averaging identities, calibration of the induced task mean, task-specific finite-replica covariance, and mask-level full-refit behaviour. These checks answer different questions: an analytic identity does not establish task calibration, a small K/M term does not ensure stable roots, and either result alone does not justify a probability-law interpretation. Consequently, validation is conditional on the stated estimand, analysis unit, generator and path law, replica budget, and refitting procedure; it is not a universal certification of generated records.

The applications show why these distinctions matter. In SIPP, the benefit of joint generation emerged only after calibration to the high-leverage capped debt-to-value task. In HRS, common household innovations increased the task-specific covariance term, while the finite-M root experiment supported the proposed integration-variance calculation. In

CPS, weak observed information led to root failures and deficient coverage despite small numerical integration error. These findings concern the studied conditional mask laws rather than national survey-design coverage, and they do not imply universal dominance of a path construction. Finite-path uncertainty can be quantified or made negligible only after task calibration and root stability have been established.

Data and code availability

The empirical analyses use public-use SIPP, HRS, and CPS files. The 2023 SIPP public-use files and January 2025 Basic CPS file and documentation are available from the U.S. Census Bureau [U.S. Census Bureau, 2024b, 2025, 2019]. The 2006–2018 HRS core and psychosocial public-use products are available through the University of Michigan HRS data-products portal [Health and Retirement Study, 2026]. Raw public-use microdata are not redistributed; researchers should obtain them directly from the official distributors and comply with the applicable registration and use conditions. A public replication package will include executable analysis and manuscript scripts, machine-readable result tables, and documentation mapping each reported table and figure to the code that produces it; users will supply the official source microdata. A persistent repository identifier will be added after deposit, and none is claimed here.

Funding and disclosure

This work was supported by the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2025S1A5C3A0201187111). The author reports no conflict of interest.

Acknowledgements and AI use disclosure

OpenAI Codex assisted with drafting, editing, coding, computational checks, literature searches, and submission preparation. The author verified all outputs and takes full

responsibility.

References

Kobi Abayomi, Andrew Gelman, and Marc Levy. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):273–291, 2008.

Julian B. Axenfeld, Christian Bruch, Christof Wolf, and Annelies G. Blom. The performance of multiple imputation in social surveys with missing data from planned missingness and item nonresponse. *Survey Research Methods*, 18(2):137–151, 2024.

John Barnard and Donald B. Rubin. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999.

Jonathan W. Bartlett and Rachael A. Hughes. Bootstrap inference for multiple imputation under uncongeniality and misspecification. *Statistical Methods in Medical Research*, 29 (12):3533–3546, 2020.

Todd E. Bodner. What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4):651–675, 2008.

Irina Bondarenko and Trivellore E. Raghunathan. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, 35(17):3007–3020, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

John W. Graham, Allison E. Olchowski, and Tamika D. Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213, 2007.

Ofer Harel. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, 4(1):75–89, 2007.

David Haziza and Jean-François Beaumont. Construction of weights in surveys: A review. *Statistical Science*, 32(2):206–226, 2017.

Yulei He and Alan M. Zaslavsky. Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Statistics in Medicine*, 2012.

Health and Retirement Study. Public survey data: Biennial core and psychosocial products. Survey Research Center, University of Michigan, 2026. URL https://hrsdata.isr.umich.edu/data-products/public-survey-data. 2006–2018 public data products.

Jae Kwang Kim and David Haziza. Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24(1):375–394, 2014.

Oliver Lipps and Ursina Kuhn. Income imputation in longitudinal surveys: A within-individual panel-regression approach. *Survey Research Methods*, 17(2):159–175, 2023.

Roderick J. A. Little and Sonya Vartivarian. Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161–168, 2005. URL https://www150.statcan.gc.ca/n1/pub/12-001-x/2005002/article/9046-eng.pdf.

Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

Kevin M. Murphy and Robert H. Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 3(4):370–379, 1985.

Adrian Pagan. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1):221–247, 1984.

Trivellore E. Raghunathan, Jerome P. Reiter, and Donald B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16, 2003. URL https://www2.stat.duke.edu/courses/Spring06/sta395/raghunathan2003.pdf.

Jerome P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003. URL https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6785-eng.pdf.

Jerome P. Reiter and Trivellore E. Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471, 2007.

James M. Robins and Naisyin Wang. Inference for imputation estimators. *Biometrika*, 87 (1):113–124, 2000.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

Donald B. Rubin. Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1):3–18, 2003.

Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):663–688, 2018.

U.S. Census Bureau. *Current Population Survey: Design and Methodology, Technical Paper 77*, 2019. URL https://www.census.gov/programs-surveys/cps/technical-documentation/complete.html.

U.S. Census Bureau. *2023 Survey of Income and Program Participation Users' Guide*, 2024a. URL https://www2.census.gov/programs-surveys/sipp/tech-documentation/methodology/2023_SIPP_Users_Guide_OCT24.pdf.

U.S. Census Bureau. 2023 survey of income and program participation public-use files, 2024b. URL https://www2.census.gov/programs-surveys/sipp/data/datasets/2023/.

U.S. Census Bureau. 2025 basic monthly current population survey public-use files, 2025. URL https://www.census.gov/data/datasets/2025/demo/cps/cps-basic-2025.html. January 2025 basic monthly file and data dictionary.

Paul T. von Hippel. How many imputations do you need? a two-stage calculation using a quadratic rule. *Sociological Methods & Research*, 49(3):699–718, 2020.

Paul T. von Hippel and Jonathan W. Bartlett. Maximum likelihood multiple imputation: Faster imputations and consistent standard errors without posterior draws. *Statistical Science*, 36(3):400–420, 2021.

Maaike Walraad, Jonas Klingwort, and Joep Burger. Incorporating machine learning in capture-recapture estimation of survey measurement error. *Survey Research Methods*, 18(2):99–112, 2024.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

Xianchao Xie and Xiao-Li Meng. Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, 27(4):1485–1545, 2017.