Predicting Real-World Quantitative Data with Synthetic Populations

Artur Ventura
Synthetic Users
artur@syntheticusers.com

Henrique Rodrigues
Synthetic Users
henrique@syntheticusers.com

Abstract. Synthetic populations—microsimulation models, statistically matched synthetic respondents, and, increasingly, populations of large-language-model personas—are now used to predict real-world quantitative data: survey toplines, poll margins, choice shares, and experimental treatment effects. Yet the practice lacks a shared statistical account of what such predictions estimate, how they should be scored, and when they can succeed. We give one. A population is formalized as a probability measure over latent individuals; each individual is a family of parametric per-question response kernels tied together by a low-dimensional latent trait; and a survey observes only finite, noisy *mixtures* (marginals) of those kernels. Building a synthetic population is then the ill-posed inverse problem of recovering a measure over latents from its projections. Within this framework we derive four operational results. First, calibration is a marginal-matching minimum-distance estimator, with the choice of divergence and weights made explicit. Second, evaluation admits an exact finite-sample semantics: even a perfect synthetic engine exhibits a computable sampling *floor*, and a skill statistic anchored between that floor and a zero-information reference upgrades, via a simulated null band, to a goodness-of-fit test of indistinguishability from perfection. Third, calibrating one population to many instruments should be posed as a constrained program whose *infeasibility* is a misspecification alarm rather than a numerical nuisance. Fourth, all prediction error decomposes into exactly two sources —population composition and response mechanism—which require different corrections. Identifiability analysis shows regularization is mandatory, and two Monte-Carlo-verified worked examples instantiate the floor, the skill test, and the failure of marginals to identify the mixing measure.

1 Introduction

Quantitative social measurement has always rested on a costly primitive: ask real people real questions. A growing family of *synthetic* methods proposes to relax that constraint. Microsimulation and population synthesis reconstruct populations from census margins; statistical matching and raking reweight existing samples toward new targets; and, most recently, populations of large-language-model (LLM) personas are prompted to stand in for human respondents altogether (Argyle et al., 2023; Dillion et al., 2023; Filippas et al., 2024). The ambition of these methods is no longer descriptive convenience but genuine *prediction of real-world quantitative data*: survey toplines, poll margins, product choice shares, and experimental treatment effects. The ambition is increasingly credible. In a large-scale evaluation across 70 preregistered survey experiments, treatment effects inferred from LLM-simulated samples correlated with the actual effects about as strongly as pooled human forecasts, and the correlation persisted for studies published after the model's training cutoff (Ashokkumar et al., 2026). At the same time, the same evaluation found that predictions *systematically overestimated effect sizes*, and other audits report brittle failures of synthetic respondents on exactly the subpopulation structure that matters for inference (Bisbee et al., 2024).

This mixed picture—strong headline correlation, systematic distributional bias—is precisely what one should expect from a method whose statistical foundations have not been stated. Correlation with held-out outcomes is evidence that a synthetic population carries signal; it is silent on the questions a survey statistician would ask first. What population parameter is the synthetic engine estimating? Against what benchmark should its error be judged, given that a finite run of a *perfect* engine still misses any finite target? When one engine must reproduce many surveys at once, what does it mean that no configuration satisfies all of them? And when a prediction is wrong, is the fault in *who* the synthetic individuals are, or in *how* they answer? The literature on LLM simulation currently answers these questions ad hoc, study by study; the older literatures that could answer them in general—item response theory, mixture models, survey statistics, and the theory of ill-posed inverse problems—have not been assembled around the synthetic-prediction task.

This paper assembles them. We develop a statistical formalism in which the synthetic-population enterprise can be stated, evaluated, and criticized with standard tools. The picture in one paragraph is as follows. There is a real target population we can never see in full; a survey is a finite sample from it. Each person is not a fixed answer but a *bundle of response distributions*—one per question—tied together by a latent internal state. What a survey reports is therefore not "the answer of the population" but a *mixture*: the average of individual response distributions over whoever happens to be in the population. Building a synthetic population means constructing a distribution over latent individuals, plus their response mechanisms, whose induced mixtures reproduce what surveys saw and forecast what they will see. The catch that defines the whole problem is that we observe only *marginals*, and only noisy finite-sample estimates of them; and marginals do not pin down a population. Synthetic prediction is an ill-posed inverse problem: recovering a measure over latents from its

projections.

Contributions. Within this framework we obtain four operational results.

1. **A formalism for the task** (Sections 2 to 4). Populations are probability measures over latent states; individuals are families of parametric response kernels sharing a low-dimensional trait; observables are mixtures. The construction is the natural meeting point of item response theory (Lord and Novick, 1968; Reckase, 2009) and de Finetti's representation of exchangeable populations (de Finetti, 1937).
2. **Calibration as marginal matching** (Section 5). Fitting a synthetic population is a minimum-distance / generalized method of moments estimator (Wolfowitz, 1957; Hansen, 1982), which makes the choice of divergence and of weights a substantive modeling decision rather than an implementation detail.
3. **Finite-sample evaluation** (Section 6). A finite synthetic run of a perfect engine still shows computable, nonzero divergence from any finite target: the sampling *floor*. We derive the floor in closed form, correct it for noise in the target itself, define a skill statistic anchored between the floor and a zero-information reference, and upgrade it to a hypothesis test of "indistinguishable from perfect" via a simulated null band.
4. **Multi-instrument calibration, identifiability, and error anatomy** (Sections 7 to 9). Calibrating one population to many surveys should be posed as a constrained program whose infeasibility is a misspecification alarm; marginals identify neither the joint nor the mixing measure, so regularization is mandatory; and all prediction error decomposes into population composition versus response mechanism—two sources requiring different corrections.

Two worked examples (Section 10), verified against 2×10^5-replicate Monte Carlo, instantiate the floor, the skill test, and the failure of low-order observables to identify the mixing measure.

Throughout we distinguish *calibration*—agreement with the marginals the engine was fitted to—from *prediction*—agreement with instruments held out of the fit. The distinction matters because the inverse problem is ill-posed: many synthetic populations match any finite set of marginals, and they disagree on everything else. Credit accrues only on held-out instruments, and Section 7 makes that requirement formal. Our diagnostic machinery also speaks directly to current empirical findings: the effect-size overestimation of Ashokkumar et al. (2026) is, in the vocabulary of Section 9, a between-condition contrast inflated by either over-dispersed composition or miscalibrated response channels, and the decomposition tells the two apart.

Related work. The ingredients are classical even where the application is new. The latent-trait response model is multidimensional item response theory (Rasch, 1960; Birnbaum, 1968; Lord and Novick, 1968; Reckase, 2009), with conditional independence as the local-independence axiom of latent-structure analysis (Lazarsfeld and Henry, 1968). Population-level observables as mixtures go back to de Finetti (de Finetti, 1937); identifiability of the mixing measure is the theory of Teicher (1963), Lindsay (1995), and, for conditionally independent items, Kruskal (1977) and Allman et al. (2009). Marginal matching is minimum-distance estimation (Wolfowitz, 1957; Newey and McFadden, 1994) and GMM (Hansen, 1982). In applied population synthesis the same reconstruction is performed by iterative proportional fitting (Deming and Stephan, 1940) and multilevel regression with poststratification (Gelman and Little, 1997; Park et al., 2004). Our evaluation machinery draws on goodness-of-fit theory (Read and Cressie, 1988), and our treatment of inter-survey disagreement on differential item functioning (Holland and Wainer, 1993) and house-effect pooling (Jackman, 2005). What is new here is not any single ingredient but the assembly: a single formal frame in which LLM-persona engines, microsimulators, and classical reweighting are instances of one estimation problem with one evaluation theory.

Roadmap. Section 2 states the problem. Sections 3 and 4 build the generative model and its observable. Section 5 defines the estimator, Section 6 its finite-sample evaluation, Section 7 the multi-instrument program. Section 8 develops identifiability and regularization, Section 9 the error decomposition, Section 10 the worked examples. Section 11 concludes with practice recommendations and limitations.

2 Problem Setup: Populations, Observations, and the Prediction Task

Let Ω be the space of **individual latent states**—everything about a person that governs how they answer anything: dispositions, demographics, context. The real target population is a probability measure

$$P^* \in \mathscr{P}(\Omega), \qquad \omega \sim P^*.$$

We call P^* the *super-population* to emphasize that it is the generator of respondents, not any particular roster of them. A **survey** is a finite i.i.d. draw $\omega_1, \ldots, \omega_n \sim P^*$, giving the empirical measure $\hat{P}_n = \frac{1}{n}\sum_i \delta_{\omega_i}$. Crucially, we never observe ω_i directly—only answers to questions, through the response mechanism of Section 3.

The i.i.d. idealization is rarely exact for real surveys, which use stratification, clustering, and unequal selection probabilities, and are subject to nonresponse (Cochran, 1977). Under a complex design the effective sample size $n_{\text{eff}} = n/\text{deff}$ (with deff the design effect) replaces n throughout, so the $n^{-1/2}$ sampling floor of Sections 4 and 6 should be read with n_{eff}, and design or calibration weights enter the empirical marginals as $\hat{p}_j(c) = \sum_i w_i \mathbb{1}[r_{ij} = c] / \sum_i w_i$. Nothing else in the development changes, and we suppress the distinction below.

The prediction task. A *synthetic population* is a constructed distribution over latent individuals together with modeled re-

sponse mechanisms—whether realized by a microsimulator, a reweighted donor sample, or a pool of LLM personas. The task this paper formalizes is:

> construct a synthetic population whose induced observables (i) match the finite, noisy data that real surveys have already reported (*calibration*) and (ii) forecast the quantitative results of instruments not used in the fit (*prediction*).

The remainder of the paper gives the three components this statement needs: a generative model of individuals making "induced observables" precise (Sections 3 and 4); an estimator making "match" precise (Section 5); and a finite-sample evaluation theory making "forecast" scorable against an achievable benchmark (Sections 6 and 7).

3 A Generative Model of Individuals

The setup of Section 2 left one object undefined: the mechanism that turns an unobservable latent state ω into the answers a survey actually records. This section supplies it in two steps. First we model a single individual's response behavior, question by question, as a family of stochastic channels; then we impose the structure that ties those channels together across questions—a shared low-dimensional latent trait. The two steps are deliberately separated because they carry different commitments: the first is close to unavoidable, while the second is the substantive modeling assumption on which identifiability (Section 8) and multi-instrument diagnostics (Section 7) will later hinge.

3.1 An individual is a family of parametric response laws

A question j (with answer set $\mathscr{C}_j$) is a **stochastic measurement channel**: a Markov kernel

$$R_j \mid \omega \;\sim\; \kappa_j(\cdot \mid \omega) \in \mathscr{P}(\mathscr{C}_j). \tag{1}$$

The same person is *not* deterministic—asked repeatedly they scatter (mood, framing, genuine ambivalence). Concretely each κ_j is parametric, its parameters a function of the latent state:

$$\kappa_j(\cdot \mid \omega) = \mathscr{F}_j\big(\cdot\,; \eta_j(\omega)\big),$$

$$\mathscr{F}_j = \begin{cases} \text{Categorical}\big(\pi_j(\omega)\big) & \text{nominal} \\ \text{OrderedProbit}\big(\eta_j(\omega),\, \tau_j\big) & \text{ordinal/Likert} \\ \mathscr{N}\big(\mu_j(\omega),\, \sigma_j^2\big) & \text{scale.} \end{cases}$$

So an individual *is* the whole collection $\{\kappa_j(\cdot \mid \omega)\}_{j=1}^{J}$—a per-question parametric distribution, not a vector of answers. This distinction is not pedantic: it is where within-individual response variance lives, and Section 9 shows that collapsing it (a deterministic engine) distorts exactly the population quantities a synthetic method is built to recover.

3.2 What ties the questions together: a latent trait

The natural parameters share a low-dimensional latent cause $\theta(\omega) \in \mathbb{R}^d$ (the "axes" or traits), e.g.

$$\eta_j(\omega) = a_j^\top \theta(\omega) - b_j. \tag{2}$$

Assumption 3.1 (Local independence)**.** Conditional on the latent trait, answers to distinct questions are independent:

$$R_1, \ldots, R_J \;\perp\!\!\!\perp\; \text{given } \theta(\omega).$$

This is item response theory embedded in a mixture (de Finetti) model: **all cross-question correlation flows through the latent state**. Two questions look correlated in the population only because the same θ drives both. The linear form (2) is the multidimensional two-parameter normal-ogive / logistic model, with a_j the discriminations (loadings) and b_j the difficulty; the unidimensional binary case is Rasch (Rasch, 1960) and Birnbaum (Birnbaum, 1968), with foundations in Lord and Novick (1968) and the multidimensional theory in Reckase (2009). Assumption 3.1 is precisely the "local independence" axiom of latent-structure analysis (Lazarsfeld and Henry, 1968). And the model class is not merely convenient: de Finetti's theorem (de Finetti, 1937) says a population of exchangeable respondents *must* be representable as a mixture of conditionally independent response laws—the latent trait is the de Finetti mixing variable. What is a modeling choice is the *dimension* d and the *parametric form* of the channels; what is not optional is the mixture structure itself.

Assumption 3.1 does real work throughout: it makes the population joint factorize inside the mixture integral (Section 4), it is the structure under which enough items identify the mixing measure (Section 8), and its failure is one of the named culprits when multi-instrument calibration turns infeasible (Section 7).

4 The Population Observable as a Mixture

A survey does not report κ_j; it reports the **marginal**, the response law integrated over the population:

$$p_j(c) \;=\; \mathbb{E}_{\omega\sim P^*}\big[\kappa_j(c \mid \omega)\big] \;=\; \int_\Omega \kappa_j(c \mid \omega)\, dP^*(\omega), \tag{3}$$

and, under Assumption 3.1, the full population joint is likewise a mixture,

$$P_R^*(r_{1:J}) = \int_\Omega \prod_j \kappa_j(r_j \mid \omega)\, dP^*(\omega). \tag{4}$$

Equation (3) is the object a synthetic population must reproduce, and (4) is the object it would need to reproduce to be trusted on *new* cross-question claims. The gap between the two is the identifiability problem of Section 8.

From n respondents we get only the **noisy empirical marginal**

$$\hat{p}_j(c) = \tfrac{1}{n}\sum_i \mathbb{1}[r_{ij} = c], \qquad \hat{p}_j = p_j + O_p\big(n^{-1/2}\big). \tag{5}$$

That $n^{-1/2}$ is an **irreducible floor**: even a perfect model cannot match $\hat{p}_j$ better than a fresh real sample of the same size would. Section 6 turns this remark into an exact, testable statement.

5 Synthetic Prediction as Marginal Matching

Posit a model class $\{Q_\theta\} \subset \mathscr{P}(\Omega)$ for the population and (optionally) modeled channels $\tilde{\kappa}_j$. A **synthetic population** is a draw $\tilde{\omega}_1, \ldots, \tilde{\omega}_N \sim Q_\theta$ with answers $\tilde{R}_{ij} \sim \tilde{\kappa}_j(\cdot \mid \tilde{\omega}_i)$, inducing the synthetic marginal

$$q_j(\theta) = \int_\Omega \tilde{\kappa}_j(\cdot \mid \omega)\, dQ_\theta(\omega). \tag{6}$$

For an LLM-persona engine, Q_θ is the (implicit) distribution over personas induced by the prompting scheme and $\tilde{\kappa}_j$ the model's response distribution for a persona; for a microsimulator both objects are explicit. The formalism does not care which.

The estimand is the θ that makes the induced observables agree with the real ones:

$$\boxed{\theta^* = \arg\min_\theta \sum_j w_j D_j\big(q_j(\theta), \hat{p}_j\big)} \tag{7}$$

a **moment/marginal-matching** estimator, with D_j a distributional divergence. Formally (7) is a minimum-distance / M-estimator (Wolfowitz, 1957; Newey and McFadden, 1994); when D_j is a quadratic form in the marginal residuals it is exactly the generalized method of moments (Hansen, 1982), the moment conditions being $\mathbb{E}[\mathbb{1}(R_{ij} = c)] - q_j(c;\theta) = 0$.

Two design choices in (7) are substantive, not clerical.

The weights. The w_j are not free parameters: the efficient (minimum-variance) choice is the inverse of the covariance of the moment residuals, which for a multinomial cell is $w_c \propto 1/(q_c(1-q_c))$. Uniform weights silently over-weight high-entropy questions.

The divergence. Total variation and the power-divergence family (Pearson χ^2, likelihood-ratio, Freeman–Tukey) (Read and Cressie, 1988) treat $\mathscr{C}_j$ as unordered and so ignore the metric structure of ordinal and scale questions. For the ordered channels of Section 3, a transport metric such as the 1-Wasserstein distance (Villani, 2009)—or a kernel MMD (Gretton et al., 2012)—respects inter-category distance and is usually the honest choice: predicting "4" when the truth is "5" should cost less than predicting "1".

The synthetic sample is thus a sample from Q_θ, itself an approximation of a sample from P^*:

$$\hat{P}_N^{\text{syn}} \longrightarrow Q_\theta \approx P^* \longleftarrow \hat{P}_n^{\text{real}}.$$

Calibration is not prediction. Minimizing (7) certifies agreement only with the marginals in the fit. Because the inverse problem is ill-posed (Section 8), many θ achieve near-identical calibration losses while disagreeing on unfitted quantities. A synthetic engine therefore earns predictive credit only on *held-out instruments*—surveys, questions, or experimental contrasts excluded from (7)—evaluated against the finite-sample benchmarks of Section 6, and jointly across instruments as in Section 7.

6 Finite-Sample Semantics and Evaluation

The estimand $q_j(\theta)$ in (6) is an *integral*—an infinite-population object we never possess. A run produces a **finite sample** of N cases, and the quantity we actually compute is random. Making this explicit is what turns the fit into a genuine sampling statement, and what furnishes an achievable benchmark for prediction.

6.1 Probability space of a run

Fix θ and a question j. A run draws, independently for $i = 1, \ldots, N$, a latent $\tilde{\omega}_i \sim Q_\theta$ and an answer $\tilde{R}_{ij} \sim \tilde{\kappa}_j(\cdot \mid \tilde{\omega}_i)$. The pairs $(\tilde{\omega}_i, \tilde{R}_{ij})$ are i.i.d., so marginally the answers are i.i.d. from the mixture: $\tilde{R}_{1j}, \ldots, \tilde{R}_{Nj} \overset{\text{iid}}{\sim} q_j(\theta)$. The estimator is the empirical measure

$$\begin{aligned} \hat{q}_j^N(c) &= \frac{1}{N}\sum_{i=1}^N \mathbb{1}[\tilde{R}_{ij} = c], \\ N\hat{q}_j^N &\sim \text{Multinomial}\big(N, q_j(\theta)\big), \end{aligned} \tag{8}$$

with $\mathbb{E}[\hat{q}_j^N] = q_j(\theta)$ and $\text{Var}[\hat{q}_j^N(c)] = q_c(1-q_c)/N$. The population marginal $q_j(\theta)$ is deterministic; $\hat{q}_j^N$ is what we see.

Two samplings. The target is itself an estimate: the reported topline $\hat{p}_j^n$ comes from n real respondents, $\hat{p}_j^n = p_j + O_p(n^{-1/2})$ by (5). What we compute is $D_j(\hat{q}_j^N, \hat{p}_j^n)$—random in *both* N and n.

6.2 The sampling floor is a null expectation

Even a *perfect* model ($H_0 : q_j(\theta) = p_j$) shows nonzero divergence, because $\hat{q}_j^N$ jitters. Its expectation under H_0 is the floor.

Proposition 6.1 (Sampling floor, total variation)**.** *Under $H_0 : q_j(\theta) = p_j$, using the multinomial CLT $\sqrt{N}(\hat{q}^N(c) - p_c) \Rightarrow \mathcal{N}(0, p_c(1-p_c))$ and the folded-normal mean $\mathbb{E}|Z| = \sigma\sqrt{2/\pi}$ (Leone et al., 1961),*

$$\begin{aligned} \mathbb{E}_{H_0}\big[D_{\text{TVD}}(\hat{q}_j^N, p_j)\big] &= \tfrac{1}{2}\sum_c \mathbb{E}\big|\hat{q}^N(c) - p_c\big| \\ &\approx \tfrac{1}{2}\sum_c \sqrt{\tfrac{2}{\pi}\tfrac{p_c(1-p_c)}{N}} =: d_{\text{floor},j}, \end{aligned} \tag{9}$$

and, comparing two independent multinomials of sizes N and n from a common p, the two-sample floor is

$$\begin{aligned} &\mathbb{E}\big[D_{\text{TVD}}(\hat{q}^N, \hat{p}^n) \,\big|\, q = p\big] \\ &\quad\approx \tfrac{1}{2}\sum_c \sqrt{\tfrac{2}{\pi}\, p_c(1-p_c)\Big(\tfrac{1}{N} + \tfrac{1}{n}\Big)}. \end{aligned} \tag{10}$$

The ordinal and scale floors are the same statement on the CDF increments ($\mathrm{Var}\hat{F}(c) = F_c(1-F_c)/N$) and on the standardized standard error of the mean ($\sigma/\sqrt{N}$ over σ). So d_{floor} is not an arbitrary anchor—it is $\mathbb{E}_{H_0}[D_j]$, the residual divergence a flawless model of run-size N would still exhibit. The folded-normal step treats the cell counts as marginally normal and drops the (negative) inter-cell covariance of the multinomial, so it is a mild upper bound; Example 1 confirms it against Monte Carlo to four decimals.

The two-sample form (10) matters in practice: using $1/N$ alone (topline treated as exact) is valid only when $n \gg N$; otherwise the floor is understated and skill reads pessimistically—a correct engine is scored as deficient for failing to out-predict the target's own noise.

6.3 Skill, and a test of indistinguishability from perfection

Anchoring between a zero-information reference $d_{\text{unif},j} = D_j(\text{Unif}, p_j)$ and the floor,

$$\widehat{\text{skill}}_j = \frac{d_{\text{unif},j} - D_j(\hat{q}_j^N, \hat{p}_j^n)}{d_{\text{unif},j} - d_{\text{floor},j}}, \qquad \mathbb{E}[\widehat{\text{skill}}_j] \approx \begin{cases} 1 & q_j(\theta) = p_j, \\ 0 & q_j(\theta) = \text{Unif}. \end{cases} \tag{11}$$

Replacing the point floor with a **simulated null band**—draw B multinomial samples of size N from $\hat{p}_j$, compute D_j for each, read the interval $[d_{2.5\%}, d_{97.5\%}]$—upgrades skill to a hypothesis test: the engine is *indistinguishable from perfect* exactly when the observed D_j lands inside the band. This is a parametric-bootstrap goodness-of-fit test (Read and Cressie, 1988): d_{floor} is the null *mean* of the discrepancy and the band is its null sampling distribution, so "inside the band" is failing to reject $H_0 : q_j = p_j$ at the 5% level. Because we resample from the estimate $\hat{p}_j$ rather than the unknown p_j, the band is a bootstrap approximation to the true null; drawing at the paired sizes (N, n) rather than N alone reproduces the two-sample floor (10). Example 1 shows the band detecting a misspecification that the scalar skill score alone would soften.

7 Predicting from Many Instruments

The population P^* is a single object, but it is probed by many surveys $\mathscr{J}^{(1)}, \ldots, \mathscr{J}^{(M)}$, each reporting its own marginals $p_j^{(m)}$. One calibrated Q_θ must answer to all of them at once, and—for prediction—to instruments held out of the fit entirely. Write study m's aggregate divergence as

$$\bar{D}^{(m)}(\theta) = \sum_{j \in \mathscr{J}^{(m)}} w_j D_j\big(q_j(\theta), \hat{p}_j^{(m)}\big).$$

There are two honest ways to combine the M objectives, and the choice is not cosmetic—it fixes the accept/revert rule of any calibration loop.

(A) Weighted expectation (scalarization).

$$\theta^* = \arg\min_\theta \sum_m \rho_m \bar{D}^{(m)}(\theta). \tag{12}$$

One pooled number: smooth, easy to optimize, trading studies off freely. But the minimum may *sacrifice* a study—push its error up to buy a larger gain elsewhere—and the weights ρ_m silently encode that tradeoff, with no guarantee any single study ends up acceptable.

(B) Hard constraint / minimax.

$$\text{find } \theta \;\text{ s.t. }\; \bar{D}^{(m)}(\theta) \le \tau_m \;\; \forall m, \qquad \text{or} \qquad \theta^* = \arg\min_\theta \max_m \big[\bar{D}^{(m)}(\theta) - \tau_m\big]_+. \tag{13}$$

Every study must clear its own bar τ_m; none is sacrificed. Crucially τ_m is *not* arbitrary—the natural tolerance is study m's **sampling floor** (the null band of Section 6): "match each survey to within its own sampling noise." The price is that the program can be **infeasible**—and that is informative.

The real nature of the choice. This is a multi-objective (Pareto) problem: (A) selects one frontier point via ρ; (B) selects the feasible corner. The decision encodes a belief:

- If the surveys are noisy views of *one coherent* population and the model is well specified, they are jointly satisfiable near the floor ⇒ use the **constraint** form (13), and read infeasibility as a **misspecification alarm**: no single θ explains all instruments (house effects, wording/mode effects, temporal drift, or a genuine violation of Assumption 3.1). Two literatures make these causes estimable rather than merely detectable: in psychometrics a channel $\tilde{\kappa}_j$ that depends on the instrument beyond θ is *differential item functioning*, a failure of measurement invariance (Holland and Wainer, 1993), while in survey and election modelling a per-house additive bias is a *house effect*, standardly recovered with a random-effects or state-space pooling model (Jackman, 2005)—either of which upgrades the τ_m-infeasibility signal into an explicit nuisance parameter.
- If the surveys are *not* mutually commensurable (different eras, instruments, houses), matching all of them exactly is the wrong goal; a **weighted expectation** (12) with ρ_m set by recency/trust is the honest bias–variance compromise.

Consequence for the calibration loop. The two forms are two accept rules. A weighted-expectation objective accepts any move that lowers the pooled sum, tolerating a regression in one study; a constrained objective rejects any move that pushes a study back above its band—exactly a per-study collateral guard. *Recommendation:* default to the constrained / minimax form (13) with τ_m the per-study sampling floor—the bar is principled (Proposition 6.1), no study is sacrificed, and infeasibility is a real signal; fall back to a soft weighted expectation only for studies deliberately judged non-comparable.

Held-out instruments. The same machinery scores genuine prediction. Fit θ on $M-1$ instruments, and ask whether $\bar{D}^{(M)}(\theta)$ for the held-out study lands inside *its* null band. This is the formal version of the credit criterion of Section 5: a synthetic engine is predictively adequate for an instrument class

exactly when held-out studies from that class are matched to within their own sampling noise. Evaluations that report only correlations between predicted and observed effects across studies (Ashokkumar et al., 2026) measure association on the study axis; the per-study band test is stricter, asking distributional agreement study by study.

8 Identifiability: Why Synthetic Prediction Is Hard

The map $Q \mapsto (q_1, \ldots, q_J)$ takes a rich distribution over Ω to a handful of low-dimensional **projections**. It is massively many-to-one:

Proposition 8.1 (Non-identification from marginals)**.** *Matching all fitted marginals does not determine the population:*

$$q_j(Q) = q_j(Q') \;\; \forall j \quad \not\Longrightarrow \quad Q = Q'.$$

In particular, marginals identify neither the joint (4) *nor the mixing measure.*

This is the same ill-posedness as tomography and deconvolution: recover a density from its projections. The mixing-measure version is the identifiability theory of mixtures (Teicher, 1963; Lindsay, 1995): a mixture over a latent is pinned down only up to features its finitely many projections can see. For exchangeable binary items the statement is sharp —J items identify the mixing measure only through its first J moments (Example 2)—and it degrades gracefully rather than failing outright. Recovering a continuous mixing density from noisy marginals is statistical deconvolution, with the notoriously slow logarithmic minimax rates of Fan (1991); yet nonparametric identifiability is recoverable from *enough* conditionally independent items via Kruskal's theorem (Kruskal, 1977; Allman et al., 2009), which quantifies how many items the structure of Assumption 3.1 needs. Three consequences follow.

Partial identification. The data confine Q to a *set*, not a point. The honest report is then a set, not a point estimate: the identified set and its sampling uncertainty are the objects of the partial-identification literature (Manski, 2003; Imbens and Manski, 2004).

Regularization is mandatory. The low-dimensional latent θ, priors on P^*, and Assumption 3.1 are what make the inverse problem tractable—they are modeling assumptions, not free lunches. These are the three standard regularizers of an ill-posed inverse problem: a rank/dimension cap (small d), a prior on the mixing measure (Tikhonov/Bayes shrinkage), and the local-independence constraint. In the applied population-synthesis literature the same role is played by iterative proportional fitting / raking to margins (Deming and Stephan, 1940) and, when covariates are available, by multilevel regression and poststratification (Gelman and Little, 1997; Park et al., 2004), which is precisely a regularized reconstruction of a population from its marginals.

Calibration can succeed while the joint fails. Matching every marginal perfectly can still leave the joint (4)—the cross-question structure—badly wrong. This is the formal reason calibration is not prediction (Section 5) and why held-out instruments (Section 7) are the only currency of predictive credit: the unfitted directions of the identified set are exactly where two calibration-equivalent engines part company.

9 Decomposing Prediction Error

Everything decomposes by the **law of total variance** applied to each answer:

$$\mathrm{Var}(R_j) = \underbrace{\mathrm{Var}_{P^*}\big(\mathbb{E}[R_j \mid \omega]\big)}_{\substack{\text{between-individual:} \\ \textit{who} \text{ is in the population}}} + \underbrace{\mathbb{E}_{P^*}\big[\mathrm{Var}(R_j \mid \omega)\big]}_{\substack{\text{within-individual:} \\ \text{response noise}}}. \qquad (14)$$

A mismatch in q_j vs p_j is therefore attributable to one of two things: the **composition** of the population (the measure P^* vs Q_θ) or the **response mechanism** (the channels κ_j vs $\tilde{\kappa}_j$). These are the only two places error can live, and they require different corrections: composition errors call for reweighting or re-synthesis of Q_θ (the IPF/MRP toolbox of Section 8); mechanism errors call for recalibrating the channels themselves.

Equation (14) is the variance-components decomposition of classical test theory and generalizability theory (Cronbach et al., 1972): the between-individual term is true-score variance and the within-individual term is measurement-error variance, whose ratio is the reliability. It cautions against the naive fix of shrinking within-individual noise to zero—a deterministic engine ($\tilde{\kappa}_j = \delta$) drives the second term to 0 and *overstates* between-individual spread, distorting exactly the composition it is meant to recover. An LLM persona queried at temperature zero is such an engine.

A diagnostic for effect-size inflation. The decomposition speaks directly to current empirical findings. LLM-simulated experiments predict treatment-effect *directions* well but systematically overestimate effect *sizes* (Ashokkumar et al., 2026). A treatment effect is a between-condition contrast of means; standardized, its denominator is the total variance (14). Inflation must therefore come from an understated denominator or an overstated numerator, and the two terms of (14) separate the candidate mechanisms: response channels that are too deterministic (within-individual variance collapsed—the temperature-zero pathology above) shrink the denominator, while a synthetic population that is too homogeneous in the traits driving the outcome, or persona responses that caricature the treatment contrast, inflate the numerator. Each cause implies a different correction—raise response dispersion to match test–retest reliability, re-synthesize composition against trait margins, or damp the channels' treatment sensitivity—and the decomposition, computed question by question, says which is operative. The scalar correlations reported in current evaluations cannot make this distinction; that is the practical case for evaluating synthetic engines distributionally (Section 6) rather than by association alone.

10 Worked Examples

The formalism is stated in symbols; the following two examples instantiate its two load-bearing claims with explicit numbers. All figures are computed exactly and cross-checked against 2×10^5-replicate Monte Carlo; the driver script (`population_formalism_examples.py`) accompanies this paper.

Example 1 (The floor, the skill statistic, and the null band)**.** Take a five-point Likert item with population marginal

$$p = (0.10, 0.20, 0.40, 0.20, 0.10),$$

a synthetic run of $N = 1000$ cases, and a real topline from $n = 800$ respondents. The zero-information anchor is $d_{\text{unif}} = D_{\text{TVD}}(\text{Unif}, p) = \frac{1}{2}\sum_c |0.2 - p_c| = 0.200$. The floors of Proposition 6.1 are

$$d_{\text{floor}}^{(1/N)} = \tfrac{1}{2}\sum_c \sqrt{\tfrac{2}{\pi}\tfrac{p_c(1-p_c)}{N}} = 0.0238,$$

$$d_{\text{floor}}^{(1/N+1/n)} = \tfrac{1}{2}\sum_c \sqrt{\tfrac{2}{\pi}p_c(1-p_c)\left(\tfrac{1}{N}+\tfrac{1}{n}\right)} = 0.0358.$$

Monte-Carlo expectations of the total-variation discrepancy reproduce both to four decimals (0.0238 and 0.0358), confirming the folded-normal approximation in (9). Note the honest two-sample floor is 50% larger than the $1/N$-only figure: with $n = 800$ comparable to $N = 1000$, treating the topline as exact would understate the floor and make a correct engine look deficient. The simulated null band (two-sample, $B = 2 \times 10^5$) is $[0.0127, 0.0665]$. We compare a perfect engine, $q = p$, against a biased one with $q = (0.14, 0.24, 0.34, 0.18, 0.10)$:

Engine	D_{TVD}	$\mathbb{E}[\widehat{\text{skill}}]$	in band
Perfect	0.0358	1.00	95%
Biased	0.0872	0.69	17%

The perfect engine sits at the floor with unit skill and lands inside the band exactly 95% of the time—by construction it is indistinguishable from perfect. The biased engine, whose true discrepancy from p is only $D_{\text{TVD}}(q, p) = 0.080$, still posts a respectable skill of 0.69, yet lands inside the band just 17% of the time: the hypothesis-test reading detects the misspecification that the scalar skill score alone would soften. This is the practical payoff of Section 6—report skill *against the two-sample floor* (10), and gate acceptance on band membership rather than on a skill threshold.

Example 2 (Marginals do not identify the mixing measure)**.** Let the latent be a success probability $\pi \in [0, 1]$ and let items be conditionally i.i.d. Bernoulli(π) given π—the exchangeable binary case of Section 3. Every observable on k items is then a moment of the mixing measure of order $\leq k$. Compare two mixing measures:

$$\text{(A) } \pi \sim \text{Beta}(1,2), \qquad \text{(B) } \pi = \tfrac{1}{2} \text{ w.p. } \tfrac{2}{3},\ \ \pi = 0 \text{ w.p. } \tfrac{1}{3}.$$

Both have first two moments $\mathbb{E}[\pi] = \frac{1}{3}$ and $\mathbb{E}[\pi^2] = \frac{1}{6}$, hence identical between-individual variance $\text{Var}(\pi) = \frac{1}{6} - \frac{1}{9} = 0.0556$. Consequently they induce *identical* distributions on any one or two items:

$$P(\text{item} = 1) = \tfrac{1}{3}, \qquad P(1,1) = \tfrac{1}{6},$$
$$P(0,0) = \tfrac{1}{2}, \qquad P(\text{exactly one}) = \tfrac{1}{3}$$

for both A and B. No amount of one- or two-item survey data—marginals and pairwise cross-tabs—can tell them apart. They part company only at the third moment: the probability that three items all read 1 is $\mathbb{E}[\pi^3] = 0.1000$ under A but 0.0833 under B, a 17% relative gap. So a rich distributional feature is left completely unconstrained by the lower-order observables, and choosing between A and B is a modeling act, not an inference—exactly the sense in which the regularization of Section 8 (here, committing to the Beta family) is mandatory rather than optional, and exactly the mechanism by which two calibration-equivalent synthetic engines diverge on held-out instruments.

11 Discussion and Conclusion

The foundational statement of this paper is compact. A population is a probability measure over latent individuals; each individual is a family of parametric per-question response kernels sharing a latent trait; a survey observes finite, noisy mixtures (marginals) of those kernels; and building a synthetic population is the ill-posed inverse problem of recovering a measure over latents from its projections. Everything operational followed from taking that statement seriously: the marginal-matching estimator (7), the sampling floor (9) and its two-sample correction, the skill statistic and null-band test, the constrained multi-instrument program (13), the non-identification result of Proposition 8.1, and the two-source error anatomy (14).

Recommendations for practice. For anyone predicting real quantitative data with a synthetic engine—LLM personas or otherwise—the framework reduces to five rules. Report skill against the two-sample floor (10), not against zero error: a finite target cannot be matched better than its own noise, and floors computed with $1/N$ alone flatter no one—they understate the bar and read correct engines as deficient. Gate acceptance on null-band membership, not on a skill threshold: Example 1 shows a meaningfully biased engine posting a respectable skill of 0.69 while the band test rejects it 83% of the time. Calibrate to many instruments with per-study constraints (13) and treat infeasibility as a finding—the data are telling you about house effects, mode effects, drift, or a broken independence assumption—rather than as an optimizer failure to be weighted away. Evaluate distributionally on held-out instruments, because Proposition 8.1 guarantees that calibration-equivalent engines exist that disagree wherever you did not look. And keep response channels stochastic: a temperature-zero persona is a degenerate channel that collapses within-individual variance and inflates every standardized contrast downstream (14).

Relation to current evidence. The empirical state of the art is encouraging and, we argue, under-analyzed. Correlations between LLM-predicted and observed treatment effects rivaling pooled human forecasters (Ashokkumar et al., 2026) establish that synthetic populations carry real predictive signal; the accompanying systematic overestimation of effect sizes is precisely the signature the decomposition of Section 9 anticipates, and it is diagnosable—channel over-determinism versus composition homogeneity versus caricatured treatment sensitivity—with the tools given here. Conversely, documented failures of synthetic respondents on subpopulation structure (Bisbee et al., 2024) are the joint-versus-marginal gap of Section 8 made empirical: engines calibrated on toplines were asked cross-question, cross-group questions the marginals never constrained. Neither finding is anomalous; both are what the formalism predicts.

Limitations. Three idealizations bound the scope. First, the i.i.d. survey model: real designs are stratified, clustered, and weighted, and while the design-effect correction of Section 2 handles the floor, a full treatment of informative nonresponse is not attempted here. Second, Assumption 3.1: local independence given a low-dimensional trait is a regularizing fiction, and although Section 7 turns its failure into a detectable signal, the framework does not yet estimate richer dependence (testlet effects, context effects, question-order effects) within the calibration itself. Third, estimator theory: we positioned (7) within minimum-distance and GMM estimation, but consistency, asymptotic distribution, and efficiency under partial identification—where the estimand is the identified *set* of Manski (2003)—are stated by reference rather than derived, and set-valued reporting with valid coverage (Imbens and Manski, 2004) remains to be worked out for this problem class.

Directions. The most immediate extensions are formal identifiability conditions for the trait model (how many items, of what discrimination geometry, pin down Q to a given resolution—Kruskal-type bounds (Kruskal, 1977; Allman et al., 2009) specialized to survey batteries); channel families richer than the three of Section 3, including open-ended and behavioral outcomes; and joint estimation of instrument nuisances (house effects, DIF) inside the constrained program, so that infeasibility is not merely an alarm but the beginning of an attribution.

The closing point is the one the field most needs stated plainly. A synthetic population is not validated by fitting the data it was built from; it is validated by matching instruments it never saw, to within their own sampling noise, distribution by distribution. Credit is earned on held-out instruments—everything else is curve fitting with extra steps.

References

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 2026. doi: 10.1038/s41586-026-10742-x. Published online 8 July 2026.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick, editors, *Statistical Theories of Mental Test Scores*, pages 397–479. Addison-Wesley, Reading, MA, 1968.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024.

William G. Cochran. *Sampling Techniques*. Wiley, New York, 3rd edition, 1977.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972.

Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937.

W. Edwards Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4): 427–444, 1940.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272, 1991.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 614–615. ACM, 2024.

Andrew Gelman and Thomas C. Little. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23(2):127–135, 1997.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

Paul W. Holland and Howard Wainer, editors. *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Simon Jackman. Pooling the polls over an election campaign. *Australian Journal of Political Science*, 40(4):499–517, 2005.

Joseph B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, 1977.

Paul F. Lazarsfeld and Neil W. Henry. *Latent Structure Analysis*. Houghton Mifflin, Boston, 1968.

F. C. Leone, L. S. Nelson, and R. B. Nottingham. The folded normal distribution. *Technometrics*, 3(4):543–550, 1961.

Bruce G. Lindsay. *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Institute of Mathematical Statistics, 1995.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editors, *Handbook of Econometrics, Volume 4*, pages 2111–2245. Elsevier, 1994.

David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, 1960.

Timothy R. C. Read and Noel A. C. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York, 1988.

Mark D. Reckase. *Multidimensional Item Response Theory*. Springer, New York, 2009.

Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963.

Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Jacob Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, 28(1):75–88, 1957.