mathematics

MDPI

Article

Disentangling Signal from Noise: A Bayesian Hybrid Framework for Variance Decomposition in Complex Surveys with Post-Hoc Domains

JoonHo Lee [1,*] and Alison Hooper [2]

[1] Department of Educational Studies in Psychology, Research Methodology, and Counseling, College of Education, The University of Alabama, Tuscaloosa, AL 35487, USA
[2] Department of Curriculum and Instruction, College of Education, The University of Alabama, Tuscaloosa, AL 35487, USA; alhooper2@ua.edu
* Correspondence: jlee296@ua.edu

Abstract

Quantifying geographic variation is crucial for policy evaluation, yet researchers often rely on complex national surveys not designed for sub-national inference. This design-analysis mismatch creates two challenges when decomposing variance across domains like states: informative sampling confounds substantive heterogeneity with design artifacts, and finite-sample variance inflation conflates sampling noise with signal. We introduce the Bayesian Hybrid Framework that reconciles design-based and model-based inference through Bayesian Pseudo-Likelihood for design consistency and a hybrid generalized linear mixed model that simultaneously estimates substantive domain effects and nuisance design effects (strata, PSUs). We propose a Dual Estimand Framework distinguishing between Descriptive (total observed variance) and Policy (substantive variance net of design) estimands, with explicit de-attenuation to correct finite-sample inflation. Simulations based on the 2019 National Survey of Early Care and Education demonstrate negligible bias and superior efficiency compared to standard alternatives. Applied to subsidy receipt among home-based child care providers, we find the observed between-state variation (16.7%) reduces to only 5.4% after accounting for design artifacts and sampling noise. This three-fold reduction reveals that local factors, not state policies, drive most heterogeneity, highlighting the necessity of our framework for rigorous geographic variance decomposition in complex surveys. An accompanying R package (version 0.3.0), `bhfvar`, implements the complete framework.

Keywords: complex survey design; variance decomposition; Bayesian Pseudo-Likelihood; domain estimation; multilevel modeling; survey weights; finite-sample bias; informative sampling

MSC: 62D05; 62F15; 62J12; 62J05; 62P25

Academic Editor: Simeon Reich

Received: 22 December 2025
Revised: 25 January 2026
Accepted: 29 January 2026
Published: 31 January 2026

1. Introduction

The landscape of public policy in the United States is inherently geographic. While federal frameworks often provide overarching structure, critical decisions in domains such as education, social services, and public health are implemented at the state or local level [1–5]. In Early Care and Education (ECE), for example, states exercise significant autonomy in administering the Child Care and Development Fund (CCDF), leading to

substantial variation in eligibility criteria, reimbursement rates, and regulatory frameworks [6–10]. Consequently, policymakers and researchers naturally ask how much of the observed heterogeneity in outcomes—such as subsidy participation among providers—is attributable to systematic differences between states versus variation within states. Statistically, this requires decomposing the variance of an outcome into between- and within-state components, a crucial step for assessing policy impacts, evaluating equity, and identifying the drivers of disparity.

To investigate these questions, researchers rely heavily on large-scale, nationally representative surveys, such as the National Survey of Early Care and Education [NSECE] [11]. These surveys employ complex, multistage sampling designs optimized for national precision. However, this optimization creates a fundamental tension when the analytical goal shifts to sub-national domains. The central methodological challenge arises from a *design–analysis mismatch*. While the analytical interest is often the state, these geographic domains are rarely incorporated as explicit design features (e.g., strata or primary sampling units (PSUs)) in national surveys. In the NSECE, the design involves stratification and clustering based on factors like community poverty density [11]. Depending on design features, strata may cross state boundaries. States are embedded within this structure rather than defining it, meaning state-level sample sizes are random, often small, and highly unequal. This mismatch leads many applied researchers to question the validity of state-level variance decompositions, concerned that results may be biased or misleading if the state was not part of the sampling scheme [12–14]. The NSECE team offered states the opportunity to supplement federal data collection to achieve state-level representativeness; however, only Minnesota committed funds to do so [11].

This skepticism is well-founded, as the design–analysis mismatch introduces two significant statistical obstacles. The first is *informative sampling*. A design is informative when inclusion probabilities remain correlated with the outcome of interest, even after conditioning on the analytical domains [15]. In the NSECE, the intentional oversampling of low-income areas likely induces informativeness for outcomes like subsidy receipt. If an analytical model ignores the actual design structure (the strata and PSUs used for sampling), the estimated "state effects" risk absorbing variation that is actually attributable to the specific realization of the design within those states. This confounding leads to biased estimates of the true substantive differences between states.

The second, distinct obstacle is *finite-sample variance inflation*. Because states are not design strata, the sample size within any given state can be very small. Design-based estimates for these small domains (e.g., the state mean) are inherently noisy [16]. Traditional variance decomposition methods, such as a design-based analysis of variance (ANOVA), conflate this sampling noise with true substantive heterogeneity across states. This "noise-as-signal" problem results in a severe upward bias in the estimate of between-state variance, often misleading researchers into believing that state-level differences are more substantial than they truly are.

Our work is closely related to, but distinct from, the small area estimation (SAE) literature [17]. SAE methods, including the Fay–Herriot area-level model [18] and unit-level nested error models [19], focus primarily on producing reliable point estimates of domain means or totals for areas with small samples, leveraging Bayesian shrinkage or empirical Bayes techniques (see Pfeffermann [20] for a comprehensive review). While our framework shares the mechanism of partial pooling for stabilizing small-domain estimates, our estimand is fundamentally different: we target the *variance decomposition itself*—specifically, the proportion of total heterogeneity attributable to systematic differences between domains. This focus on variance components, rather than domain-specific predictions, addresses distinct policy questions about the magnitude and sources of geographic inequality. Fur-

thermore, whereas traditional SAE often assumes noninformative sampling conditional on auxiliary variables, our Bayesian Hybrid Framework explicitly incorporates design features (strata and PSUs) as nuisance random effects and employs pseudo-likelihood weighting to achieve design consistency under informative sampling.

Existing approaches address parts of this problem but do not offer a fully satisfactory solution. Design-based ANOVA on domain means uses survey weights to obtain design-consistent state means, but the naive between-state variance of those estimates is highly susceptible to inflation by the average sampling variance of the state means. While method-of-moments corrections exist, they often do not fully propagate uncertainty and can be unstable. Naive multilevel modeling (MLM) provides partial pooling for small domains but ignores survey weights and the design structure, failing to achieve design consistency under informative sampling [21]. By omitting design features, these models incorrectly partition the variance, confounding design-induced clustering with substantive state-level heterogeneity. Frequentist pseudo-maximum likelihood (PML) incorporates weights into multilevel models and can achieve design consistency [22–24]. However, implementing PML for complex models with multiple, crossed random effects (e.g., states, strata, and PSUs) is computationally intensive and technically complex, and it does not inherently resolve finite-sample inflation or yield directly interpretable variance components on the probability scale [25].

These limitations highlight a critical methodological gap: no existing approach simultaneously (i) achieves design consistency under informative sampling, (ii) correctly partitions substantive domain variance from design-induced clustering, and (iii) addresses finite-sample noise inflation while providing coherent uncertainty quantification for nonlinear outcomes on the probability scale. To fill this gap, this paper develops the *Bayesian Hybrid Framework*, which reconciles design-based and model-based perspectives [16,26–28] to enable rigorous variance decomposition across domains that were not explicitly part of the survey design. The framework offers several key innovations. First, it achieves design consistency through the *Bayesian pseudo-likelihood* (BPL)—the Bayesian analogue of PML—which incorporates survey weights into a generalized linear mixed model (GLMM) [15,29,30]. This delivers design-consistent inference for finite-population quantities while retaining the advantages of Bayesian estimation, such as coherent uncertainty quantification and partial pooling (shrinkage) that stabilizes estimates in small domains. Second, we propose a hybrid model structure that includes both *substantive* random effects for states and *nuisance* design random effects for strata and PSUs. This explicit modeling of design hierarchies prevents design-induced clustering from being spuriously interpreted as state-level heterogeneity.

Third, we introduce the *Dual Estimand Framework* on the probability scale, recognizing that under informative sampling, the concept of "between-state variance" is inherently ambiguous. We define *Estimand A (Policy Estimand)* as the between-state variance that would remain if the distribution of design features were conceptually equalized across states, isolating the component plausibly attributable to state-level policies and contexts. We define *Estimand B (Descriptive Estimand)* as the variance that incorporates the realized pattern of design features within each state, capturing the total geographic heterogeneity in the actual finite population. Finally, to address finite-sample noise, we integrate an explicit *de-attenuation* step. We construct *Estimand A**, a de-attenuated version of A that subtracts the estimated average design-based sampling variance of state means. The gap between A and B diagnoses the impact of informative sampling; the gap between A and A* quantifies the impact of finite-sample noise.

For clarity, the paper's contributions can be summarized as follows:

- **Design-consistent hybrid modeling:** We integrate Bayesian pseudo-likelihood with a multilevel random-effects structure that includes both substantive domains (states) and nuisance design features (strata/PSUs).
- **Probability-scale variance decomposition:** We target finite-population between- and within-domain variance components directly on the probability scale, enabling direct policy interpretation.
- **Dual Estimand Framework:** We distinguish policy-relevant state heterogeneity (Estimand A) from realized descriptive variance (Estimand B), using their gap to diagnose informativeness.
- **Explicit de-attenuation with uncertainty propagation:** We incorporate a method-of-moments correction (Estimand A*) to remove finite-sample variance inflation while retaining coherent posterior uncertainty.
- **Open-source software implementation:** We provide the `bhfvar` R package (version 0.3.0) (https://joonho112.github.io/bhfvar/, accessed on 28 January 2026) which implements the complete framework with user-friendly functions for data preparation, model fitting, and estimand computation, enabling researchers to readily apply these methods to their own complex survey data.

Our empirical focus is on subsidy participation among listed home-based providers in the 2019 NSECE. Specifically, states vary in their policies related to child care subsidies, federal funds awarded to states through the CCDF to help lower-income families pay for child care, and these policy differences may relate to child care providers' decisions to participate in the child care subsidy program. Using an NSECE-based simulation design that preserves the actual state–stratum–PSU skeleton, we evaluate the performance of the Bayesian Hybrid Framework relative to standard alternatives. We then apply the framework to the NSECE data to answer a concrete policy question: How much substantive variation in subsidy participation across states remains once the influence of the complex sample design and finite-sample noise is removed? Formally, the paper is organized around three research questions:

1. **Validity and design consistency.** Under complex multistage designs where states are not sampling strata and sampling may be informative, can the Bayesian Hybrid Framework recover finite-population between-state variance on the probability scale with negligible bias, and how does its performance compare to standard alternatives?
2. **Disentanglement and interpretation.** How should we define and interpret multiple notions of "between-state variance" under informative sampling, and to what extent do design features inflate or distort the apparent magnitude of state-level heterogeneity?
3. **Bias correction and precision.** To what degree does finite-sample noise in small domains inflate between-state variance estimates, and can an integrated de-attenuation strategy (Estimand A*) improve the bias–variance trade-off beyond what is achieved by Bayesian shrinkage alone?

The remainder of the paper proceeds as follows. Section 2 formalizes the target finite-population variance decomposition and the statistical challenges of informative sampling and variance inflation. Section 3 introduces the Bayesian Hybrid GLMM, detailing the BPL construction, the hybrid random-effects structure, and the centering strategies that ensure identifiability. Section 4 develops the Dual Estimand Framework and defines Estimands A, A*, and B. Section 5 presents the NSECE-based simulation study. Section 6 applies the framework to the 2019 NSECE data to quantify state-level heterogeneity in subsidy receipt. Section 7 concludes with methodological implications, substantive findings for ECE policy, and practical guidance for applied researchers.

2. The Statistical Challenge: Domain Estimation and Variance Inflation

In this section we formalize the statistical problem that arises when analysts wish to decompose variation across domains (states) that were *not* explicitly built into the complex survey design. The target is a population-level variance decomposition across states; the obstacles are (i) a multistage, stratified, clustered design that induces informative sampling with respect to the outcome, and (ii) finite-sample variance inflation when some state samples are very small, as in the NSECE listed home-based provider sample.

2.1. Formalizing the Population Structure and Variance Decomposition

Let $\mathcal{U}$ denote a finite target population of size N. For unit $i \in \mathcal{U}$, let $Y_i \in \{0,1\}$ be the outcome of interest (e.g., an indicator of subsidy receipt for provider i), and let $S_i \in \{1,\ldots,S\}$ denote the state membership of unit i. In the NSECE application, $\mathcal{U}$ is the population of listed home-based providers in the U.S., and $S = 51$ indexes the 50 states plus the District of Columbia [11].

Let $N_s = \sum_{i\in\mathcal{U}} \mathbf{1}(S_i = s)$ be the number of units in state s, and $\pi_s = N_s/N$ the population share of state s. Define the finite-population state mean $p_s = N_s^{-1}\sum_{i:S_i=s} Y_i$, and the overall mean $\bar{p} = \sum_{s=1}^{S} \pi_s p_s$.

If we view S as a random label taking value s with probability π_s, the law of total variance yields

$$\mathrm{Var}(Y) = \mathrm{Var}_S\{E(Y \mid S)\} + E_S\{\mathrm{Var}(Y \mid S)\}. \tag{1}$$

For binary Y, $E(Y \mid S = s) = p_s$ and $\mathrm{Var}(Y \mid S = s) = p_s(1 - p_s)$. Hence

$$\mathrm{Var}(Y) = E_S\{p_s(1 - p_s)\} + \mathrm{Var}_S(p_s), \tag{2}$$

where

$$\sigma^2_{\text{within}} = E_S\{p_s(1 - p_s)\}, \qquad \sigma^2_{\text{between}} = \mathrm{Var}_S(p_s)$$

denote, respectively, the within-state and between-state components of the population variance on the probability scale.

A convenient summary of the relative importance of state-level heterogeneity is the finite-population analogue of an intraclass correlation coefficient (ICC),

$$\rho_{\text{state}} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}. \tag{3}$$

Remark (Probability-scale decomposition). We emphasize that our variance decomposition targets the probability scale rather than the latent (logit) scale commonly used in GLMM analyses. This choice is deliberate for two reasons. First, variance components on the latent scale lack direct interpretability; they can only be compared as ratios to the logistic distribution's inherent variance ($\pi^2/3 \approx 3.29$), yielding ICCs that do not translate to meaningful policy magnitudes. Second, and more importantly, policy questions are naturally framed in terms of probabilities: "How much do subsidy participation rates vary across states?" is answered directly by $\sigma^2_{\text{between}}$ on the probability scale, which captures differences in percentage points. A between-state standard deviation of 0.05 on the probability scale, for instance, implies that states differ by roughly 5 percentage points on average—a quantity with immediate policy relevance.

Equations (1)–(3) define the *population* estimands of interest. In particular, $\sigma^2_{\text{between}}$ and ρ_{state} quantify how much of the overall heterogeneity in subsidy receipt is attributable to systematic differences across states, as opposed to heterogeneity among providers within states. The rest of the section shows why these apparently simple quantities become non-

trivial to estimate under a complex national survey design that does not treat states as design strata.

The finite-population quantities defined in Equations (1)–(3) provide clear targets for inference. However, obtaining valid estimates of these quantities from complex survey data introduces two fundamental challenges, which we address in turn: the design-induced confounding between substantive and sampling-related variation (Sections 2.2 and 2.3) and the inflation of apparent between-domain heterogeneity by finite-sample noise (Section 2.4).

2.2. Complex Multistage Sampling and Domain Estimation

Let $\mathcal{S} \subset \mathcal{U}$ denote the sample of size n obtained from a multistage, stratified, clustered design. In a design such as NSECE, the first stage selects primary sampling units (PSUs), often counties or county clusters, within sampling strata indexed by $h = 1, \ldots, H$ (defined by geography and socio-demographic characteristics such as child poverty rates). At subsequent stages, secondary sampling units and, ultimately, individual providers are sampled within the selected PSUs. Listed providers are drawn from administrative frames (licensing and registry lists) restricted to the sampled geographic units.

Let $\pi_i = \Pr(i \in \mathcal{S})$ denote the inclusion probability for unit i, and $w_i = 1/\pi_i$ the corresponding design weight. In the NSECE, these weights incorporate the multistage selection probabilities, adjustments for nonresponse, and post-stratification or raking to population controls [22]. Crucially, states are *not* design strata: strata often cross state boundaries, and states spread across several strata. Consequently, the state-specific sample size $n_s = \sum_{i \in \mathcal{S}} \mathbf{1}(S_i = s)$ is itself random, ranging from 1 to several hundred in the listed provider sample.

For a given state s, a standard design-consistent estimator of the finite-population proportion p_s is the ratio (Horvitz–Thompson/Hájek) estimator [31]

$$\hat{p}_s^{\mathrm{HT}} = \frac{\sum_{i \in \mathcal{S}} w_i Y_i \mathbf{1}(S_i = s)}{\sum_{i \in \mathcal{S}} w_i \mathbf{1}(S_i = s)}. \tag{4}$$

Under the randomization distribution induced by the sampling design, $\hat{p}_s^{\mathrm{HT}}$ is design-unbiased (or asymptotically design-consistent) for p_s provided all units in state s have positive inclusion probability.

The overall proportion can be estimated by the weighted mean

$$\hat{\bar{p}} = \frac{\sum_{i \in \mathcal{S}} w_i Y_i}{\sum_{i \in \mathcal{S}} w_i} = \sum_{s=1}^{S} \hat{\pi}_s \hat{p}_s^{\mathrm{HT}}, \tag{5}$$

where $\hat{\pi}_s = \left(\sum_{i \in \mathcal{S}} w_i \mathbf{1}(S_i = s)\right) / \sum_{i \in \mathcal{S}} w_i$ is the estimated population share of state s.

A natural design-based analogue of the population between-state variance in (2) is then the weighted sample variance of the estimated state means,

$$\hat{\sigma}^2_{\mathrm{between}} = \sum_{s=1}^{S} \hat{\pi}_s \left(\hat{p}_s^{\mathrm{HT}} - \hat{\bar{p}}\right)^2. \tag{6}$$

If the $\hat{p}_s^{\mathrm{HT}}$ were known without error, $\hat{\sigma}^2_{\mathrm{between}}$ would be a plug-in estimator of $\sigma^2_{\mathrm{between}}$. In reality, however, each $\hat{p}_s^{\mathrm{HT}}$ is itself subject to design-based sampling variability that can be large when n_s is small, and the design structure (strata and PSUs) introduces additional dependence that must be accounted for.

The design-based estimators in Equations (4)–(6) provide the foundation for domain-level inference. However, their validity depends critically on how the sampling design relates to the outcome of interest—a relationship we formalize next.

2.3. The Challenge of Informative Sampling

The first conceptual challenge is *informative sampling*. Broadly, a design is non-informative for Y if, conditional on the analytical domains and any covariates included in the model, the inclusion probabilities do not depend on the outcome [22,32,33]. In contrast, under informative sampling, units with certain values of Y are systematically more or less likely to appear in the sample, even after conditioning on the domains.

One way to formalize this is to consider the conditional expectations. Sampling is non-informative at the domain level if

$$E(Y_i \mid i \in \mathcal{S}, S_i = s) = E(Y_i \mid S_i = s) \quad \text{for all } s, \tag{7}$$

that is, the sample mean within state s equals the population mean within s in expectation. When $E(Y_i \mid i \in \mathcal{S}, S_i = s) \neq E(Y_i \mid S_i = s)$, the design is informative, and analyses that ignore the design will generally be biased [33–35].

In the NSECE, oversampling of low-income and high-provider-density areas is likely to induce informativeness for subsidy receipt. If low-income PSUs are more likely both to be sampled and to contain providers with higher probabilities of subsidy receipt—regardless of which state they belong to—then the distribution of Y in the sample differs systematically from that in the population, even within a given state.

Let H_i and J_i denote the stratum and PSU to which unit i belongs. The variance of Y can be decomposed further by inserting these design features into the law of total variance:

$$\begin{aligned} \mathrm{Var}(Y) = {} & \mathrm{Var}_S\{E_{H,J|S}[E(Y|S,H,J)]\} \\ & + E_S\{\mathrm{Var}_{H,J|S}[E(Y|S,H,J)]\} + E_{S,H,J}\{\mathrm{Var}(Y|S,H,J)\}. \end{aligned} \tag{8}$$

The first term represents the between-state variance in the mean outcome *after averaging over the design structure within each state*; this is the object most closely aligned with policy-relevant "state effects." The second term captures additional between-PSU/stratum heterogeneity *within* states, and the third term captures residual individual-level variation. If a model uses only state indicators and ignores the design structure (H, J), the estimated "between-state" component risks absorbing some of the variability attributable to the design, leading to distorted estimates of $\sigma^2_{\text{between}}$ and ρ_{state}.

This observation motivates the modelling strategy adopted later: explicitly include random effects for strata and PSUs, so that state-level variation is defined *conditional on* the design-induced clustering.

The preceding discussion establishes that ignoring the design structure leads to confounded variance estimates under informative sampling. A distinct but equally important challenge arises from the finite sample sizes within domains, which we address in the following section.

2.4. The Challenge of Finite-Sample Bias (Variance Inflation)

A second, distinct challenge is finite-sample bias in $\hat{\sigma}^2_{\text{between}}$ when some domains are small. Even if the design were non-informative in the sense of (7), replacing the unknown p_s in (2) with noisy estimators $\hat{p}_s^{\text{HT}}$ leads to an upwardly biased estimate of between-state variance, because the sampling variability of the $\hat{p}_s^{\text{HT}}$ is itself treated as signal.

To see this heuristically, write $\hat{p}_s^{\text{HT}} = p_s + e_s$, where $E_{\text{design}}(e_s) = 0$ and $\mathrm{Var}_{\text{design}}(e_s) = V_s$, the design-based variance of the state estimator. In practice, V_s can be estimated via Taylor series linearization or replication methods (jackknife, balanced repeated replication, or bootstrap). We employ Taylor linearization, which is computationally efficient and yields similar results to replication methods for ratio estimators under typical survey conditions [31]. The choice is implemented via standard survey software (the `survey` package in R). Replication

methods may be preferred when the estimator involves highly nonlinear transformations, when replicate weights are provided with the data, or when explicit variance estimates for complex functionals are required. Importantly, our framework only requires a design-consistent estimate of V_s; thus, replication-based estimators can be substituted for Taylor linearization when appropriate without altering the methodological framework.

Substituting into (6) and taking expectations over the sampling design yields

$$E_{\text{design}}\left[\hat{\sigma}^2_{\text{between}}\right] \approx \sigma^2_{\text{between}} + \sum_{s=1}^{S} \pi_s V_s. \tag{9}$$

The first term is the true between-state variance defined in (2), while the second term is the *average sampling variance* of the state estimates, weighted by the population shares π_s. Equation (9) shows that the naive between-state variance estimator conflates substantive heterogeneity across states with sampling noise.

When state samples are small or highly variable—precisely the situation in the NSECE listed provider sample, where some states have only one or two providers—the contribution of $\sum_s \pi_s V_s$ can be large relative to $\sigma^2_{\text{between}}$. A convenient way to express the magnitude of this distortion is via an inflation factor,

$$\text{Inflation Factor} = \frac{E_{\text{design}}\left[\hat{\sigma}^2_{\text{between}}\right]}{\sigma^2_{\text{between}}} \approx 1 + \frac{\sum_{s=1}^{S} \pi_s V_s}{\sigma^2_{\text{between}}}. \tag{10}$$

In our motivating application, this factor can comfortably exceed 1.5 when the true between-state variation is modest but sampling variances in small states are large.

A simple method-of-moments correction subtracts an estimate of the average sampling variance from the naive estimator,

$$\hat{\sigma}^2_{\text{between, de-att}} = \max\left\{0, \hat{\sigma}^2_{\text{between}} - \sum_{s=1}^{S} \pi_s \widehat{V}_s\right\},$$

where $\widehat{V}_s$ is a design-based estimate of V_s obtained from Taylor linearization or replication methods. This *de-attenuated* estimator directly targets $\sigma^2_{\text{between}}$, but it treats the $\widehat{V}_s$ as known and does not propagate uncertainty. The Bayesian multilevel approach developed in Section 3 can be viewed as a principled, model-based analogue that simultaneously (i) separates state effects from design-induced clustering and (ii) accounts for finite-sample noise in the state estimates when decomposing variance across states. Design-based alternatives for variance component estimation have been proposed [36], but these moment-based approaches do not provide the coherent uncertainty quantification offered by our Bayesian framework.

3. The Bayesian Hybrid Framework

To address the statistical challenges detailed in Section 2—namely, informative sampling (Section 2.3) and finite-sample variance inflation (Section 2.4)—we propose a Bayesian Hybrid Framework. This approach integrates design-based principles within a model-based structure, aiming to achieve design consistency while leveraging the strengths of Bayesian multilevel modeling for variance decomposition and stabilization of small-domain estimates. The framework is "Hybrid" in two key respects: first, it employs a pseudo-likelihood to incorporate survey weights; second, it simultaneously models substantive domain effects (states) and nuisance design effects (strata and PSUs) to correctly partition the variance.

Before detailing each component, we summarize the Bayesian Hybrid Framework as a four-step procedure:

1. **Weight Preparation:** Normalize survey weights so that $\sum_{i \in S} w_i^* = n$, preserving relative importance while stabilizing variance estimation (Section 3.1).
2. **Model Specification:** Specify the Hybrid GLMM with random effects for substantive domains (states) and design features (strata, PSUs) simultaneously (Section 3.2).
3. **Centering for Identifiability:** Apply population-weighted centering for state effects and within-group centering for PSU effects to ensure the intercept represents the population mean and variance components are correctly partitioned (Section 3.3).
4. **Posterior Estimation and De-attenuation:** Estimate the pseudo-posterior via MCMC, compute the Dual Estimands (A, A*, B) at each iteration, and optionally apply explicit de-attenuation to correct residual finite-sample noise (Section 3.4).

This workflow is implemented in Stan; full code is provided in Appendix B and at the public repository (https://github.com/joonho112/bpl-variance-decomposition, accessed on 28 January 2026). The `bhfvar` R package (version 0.3.0) (https://joonho112.github.io/bhfvar/, accessed on 28 January 2026) provides a user-friendly interface for implementing this framework.

3.1. Bayesian Pseudo-Likelihood (BPL) for Design Consistency

When a sampling design is informative, the distribution of the outcome in the sample systematically differs from that in the population (see Equation (7)). Consequently, standard likelihood-based inference will yield biased estimates of population parameters.

To achieve design consistency, we must incorporate the sampling weights into the estimation procedure. In a frequentist context, this leads to pseudo-maximum-likelihood estimation (PML) [22,23,37–39]. We adopt the Bayesian analogue, the Bayesian pseudo-likelihood (BPL), following the framework developed by Savitsky and Toth [15], who established the theoretical foundations for pseudo-posterior inference under informative sampling.

A standard Bayesian analysis yields a posterior distribution

$$P(\theta \mid Y) \propto P(Y \mid \theta)\, P(\theta).$$

The BPL approach modifies the likelihood contribution of each sampled unit i by weighting it with the survey weight w_i [see also ref. [30] for a comprehensive treatment of uncertainty quantification in this framework]. The resulting pseudo-posterior distribution is defined as

$$P_{\mathrm{BPL}}(\theta \mid Y, w) \propto L_{\mathrm{BPL}}(Y \mid \theta)\, P(\theta), \tag{11}$$

where the Bayesian pseudo-likelihood L_{BPL} is given by

$$L_{\mathrm{BPL}}(Y \mid \theta) = \prod_{i \in \mathcal{S}} [P(Y_i \mid \theta)]^{w_i^*}. \tag{12}$$

Here, w_i^* denotes the *scaled* survey weight for unit i. The incorporation of weights adjusts the inference to reflect the target population structure.

Crucially, in multilevel models, the absolute scale of the weights w_i^* matters. As detailed by Rabe-Hesketh and Skrondal [22], the magnitude of the weights affects the "apparent" sample size within clusters. If raw weights are used and their sum greatly exceeds the actual cluster size, the estimation procedure may misallocate variance between levels, often overestimating between-cluster variance components.

To mitigate this bias and ensure computational stability, we normalize the weights w_i^* used in the likelihood such that their overall mean is one (i.e., $\sum_i w_i^* = n$) [22]. This normalization, also employed by Savitsky and Toth [15], preserves the relative importance

of the observations while stabilizing the estimation of the variance parameters and ensuring that the total posterior variance is regulated by the effective sample size.

3.2. The Hybrid GLMM Specification

We specify a generalized linear mixed model (GLMM) for the binary outcome Y_i (e.g., subsidy receipt). The core of the Hybrid approach is the simultaneous inclusion of random effects for both the substantive domains and the survey design features.

The model is defined on the logit scale—it is worth noting that for nonlinear link functions such as the logit, the conditional (cluster-specific) regression coefficients differ from the marginal (population-averaged) effects. As Rabe-Hesketh and Skrondal [22] emphasized, this distinction becomes critical when level-1 weights are used: the regression coefficients and random-intercept variance are intrinsically linked, so that scaling of weights can induce multiplicative bias in the conditional effects even when marginal effects remain relatively stable. Our framework primarily targets variance decomposition on the probability scale, which corresponds more closely to marginal interpretations. The probability of a positive outcome for unit i, conditional on the random effects $\mathbf{u}$, is modeled as [40]

$$\text{logit}\{P(Y_i = 1 \mid \mathbf{u})\} = \eta_i = \alpha + u_{\text{state}[i]} + u_{\text{stratum}[i]} + u_{\text{psu}[i]}. \tag{13}$$

The linear predictor η_i is composed of:

- α: the global intercept (representing a population-weighted average);
- $u_{\text{state}[i]}$: the substantive random effect for the state of unit i;
- $u_{\text{stratum}[i]}$ and $u_{\text{psu}[i]}$: the design random effects for the stratum and the primary sampling unit (PSU) of unit i.

We assume these effects are realizations from normal distributions with respective variances σ^2_{state}, $\sigma^2_{\text{stratum}}$, and σ^2_{psu}.

By explicitly including u_{stratum} and u_{psu}, the model partitions the total variance into components attributable to the sampling design and the component attributable to systematic differences between states. This directly addresses the challenge raised in Section 2.3: the estimated state effects $u_{\text{state}[i]}$ now capture the variation across states *conditional on* the design structure, preventing the confounding of design-induced clustering with substantive state heterogeneity (Equation (8)) [21].

The Bayesian framework requires specification of prior distributions for all parameters. Following recommendations for variance parameters in hierarchical models [41], we place weakly informative half-Student-t priors on the standard deviation parameters:

$$\sigma_{\text{state}}, \sigma_{\text{stratum}}, \sigma_{\text{psu}} \sim \text{half-}t(\nu = 3, s = 2.5).$$

These priors place substantial mass near zero while allowing for larger values if supported by the data, providing regularization without imposing strong prior beliefs about the magnitude of variation. For the global intercept α, we use an informative normal prior centered at the design-weighted overall logit of the outcome, with a moderately wide standard deviation (e.g., $\sigma = 0.5$ on the logit scale). This prior encourages consistency with the design-based national estimate while remaining sufficiently diffuse to allow updating from the data. The standardized random effects receive standard normal priors as part of the non-centered parameterization (see Section 3.3). Full prior specifications and Stan implementation details are provided in Appendix B. A sensitivity analysis examining alternative prior specifications is presented in Appendix C, demonstrating that conclusions are robust to this choice.

3.3. Handling Complex Hierarchies and Identifiability

The structure of the NSECE data involves complex relationships. States and strata are *crossed*—strata may cross state boundaries. In contrast, PSUs are strictly *nested* within strata (PSU $\subset$ stratum) [11]. The model must account for this structure while ensuring parameter identifiability through centering techniques.

We employ two key constraints.

3.3.1. Within-Group Centering for Nested Effects (PSU $\subset$ Stratum)

To correctly partition the variance, we employ within-group centering, constraining the PSU effects to average zero *within* each stratum h:

$$E\left[u_{\text{psu}[j]} \mid \text{stratum}[j] = h\right] = 0 \quad \text{for all } h. \tag{14}$$

This ensures that $u_{\text{psu}[j]}$ represents the deviation of PSU j from its stratum average.

3.3.2. Population-Weighted Centering for Substantive Effects (States)

To ensure that the global intercept α corresponds to the overall population mean (on the logit scale), the state effects must be centered relative to the population distribution. We impose a soft sum-to-zero constraint weighted by the known population shares π_s (defined in Section 2.1):

$$\sum_{s=1}^{S} \pi_s \, u_{\text{state}[s]} = 0. \tag{15}$$

This population-weighted centering aligns the model's intercept with the target finite-population quantity.

3.3.3. Implementation via Non-Centered Parameterization

We implement these constraints efficiently through a non-centered parameterization [42]. We first define standardized random effects $z \sim N(0,1)$ for all levels. The centered effects are then constructed deterministically.

For state effects, the population-weighted centering is achieved by

$$u_{\text{state}[s]} = \sigma_{\text{state}} \left(z_{\text{state}[s]} - \sum_{s'=1}^{S} \pi_{s'} \, z_{\text{state}[s']} \right). \tag{16}$$

For PSU effects, we use a flattened index j across all strata for computational efficiency. The within-group centering is achieved by

$$u_{\text{psu}[j]} = \sigma_{\text{psu}} \left(z_{\text{psu}[j]} - \bar{z}_{\text{stratum}[h(j)]} \right), \tag{17}$$

where $h(j)$ denotes the stratum of PSU j, and $\bar{z}_{\text{stratum}[h]}$ is the mean of the standardized PSU effects within stratum h. This approach improves MCMC sampling efficiency while ensuring the identifiability and interpretability of the variance components.

3.4. Integrated De-Attenuation Strategies

The framework addresses the finite-sample variance inflation bias described in Section 2.4—where sampling noise inflates estimates of between-state variance—through both implicit and explicit mechanisms.

3.4.1. Implicit De-Attenuation via Bayesian Shrinkage

The Bayesian multilevel structure naturally induces partial pooling or shrinkage. For states with small samples n_s or high design-based variance V_s, the estimate is pulled

strongly toward the global mean. This stabilizes the estimates for small domains. Consequently, the variance of the posterior estimates of the state means is inherently smaller than the variance of the raw design-based estimators $\hat{p}_s^{\mathrm{HT}}$. This model-based stabilization implicitly reduces the contribution of sampling noise to the estimated between-state variance (Equation (9)), thereby mitigating the inflation bias.

3.4.2. Explicit De-Attenuation (Method-of-Moments Correction)

While Bayesian shrinkage provides a robust approach, the framework allows for an optional, explicit correction analogous to the design-based method-of-moments estimator (Section 2.4).

First, we calculate the model-based between-state variance on the probability scale. This involves transforming the state effects from the logit scale (setting design effects to zero) and calculating their population-weighted variance:

$$\hat{\sigma}^2_{\text{between, model}} = \sum_{s=1}^{S} \pi_s \left(\hat{p}_s^{\text{model}} - \hat{\bar{p}}^{\text{model}}\right)^2, \tag{18}$$

where $\hat{p}_s^{\text{model}} = \text{logit}^{-1}(\alpha + u_{\text{state}[s]})$ and $\hat{\bar{p}}^{\text{model}} = \sum_{s=1}^{S} \pi_s \hat{p}_s^{\text{model}}$.

Next, we subtract the estimated average design-based sampling variance, using precomputed estimates $\widehat{V}_s$ obtained via standard survey methods (e.g., Taylor linearization):

$$\hat{\sigma}^2_{\text{between, de-att}} = \max\left\{0,\ \hat{\sigma}^2_{\text{between, model}} - \sum_{s=1}^{S} \pi_s \widehat{V}_s\right\}. \tag{19}$$

By integrating this correction within the Bayesian estimation procedure (e.g., at each MCMC iteration), we properly propagate the uncertainty of the model parameters into the final de-attenuated estimate, providing a principled measure of the true substantive heterogeneity across states.

4. Defining and Interpreting the Target Estimands: The Dual Estimand Framework

The Bayesian Hybrid Framework detailed in Section 3 provides a unified model that simultaneously estimates substantive domain effects (states) and nuisance design effects (strata and PSUs). However, once the model is estimated, translating the latent (logit) scale parameters into a decomposition of variance on the probability scale requires careful consideration of the target estimand. When a complex sampling design is informative (Section 2.3), the concept of "between-state variance" is inherently ambiguous. The observed variation across states is a mixture of true substantive differences and artifacts induced by the specific distribution of design features within those states.

To address this ambiguity and provide comprehensive insights, we introduce the *Dual Estimand Framework*. This framework distinguishes between a substantive quantity that isolates policy-relevant state heterogeneity and a descriptive quantity that captures the total observed variance landscape.

4.1. The Ambiguity of "Between-State Variance"

Recall the fitted Hybrid GLMM (Equation (13)):

$$\eta_i = \alpha + u_{\text{state}[i]} + u_{\text{stratum}[i]} + u_{\text{psu}[i]}.$$

The model successfully partitions the variance on the latent scale. However, to perform the population-level variance decomposition on the probability scale (Section 2.1), we must

define the state-level mean probability p_s. This requires a decision on how to handle the design effects u_{stratum} and u_{psu}.

If the design is informative—meaning design features are correlated with the outcome—the distribution of these effects systematically differs across states. For example, if one state happens to contain a disproportionate number of PSUs from a high-outcome stratum, the observed mean in that state will be elevated due to the design, not solely due to state-level factors.

This leads to two distinct interpretations of "between-state variance":

1. *Substantive interpretation (the policy question).* How much would the outcomes vary across states if we could conceptually neutralize the effects of the sampling design? This isolates the variation attributable purely to state identity (e.g., policies, economies).
2. *Descriptive interpretation (the data landscape).* How much do the average outcomes vary across states in the actual population, given their specific composition of strata and PSUs? This captures the total observed heterogeneity.

The Dual Estimand Framework explicitly targets both interpretations by defining two distinct estimands, A and B, calculated during the posterior summarizing phase (e.g., at each MCMC iteration).

4.2. Estimand A (State-Only): The Policy Estimand

Estimand A aims to isolate the substantive heterogeneity across states. We refer to this as the *Policy Estimand* because it focuses on the variance attributable purely to state identity, which is often the primary interest for policy evaluation.

Definition. To calculate Estimand A, we marginalize the design effects. This is achieved by setting the realized stratum and PSU effects to their expected value, which is zero (due to the centering constraints in Section 3.3). The state-specific mean probability under Estimand A, $p_s^{(A)}$, is defined using only the global intercept and the state random effect:

$$p_s^{(A)} = \text{logit}^{-1}\left(\alpha + u_{\text{state}[s]}\right). \tag{20}$$

It is crucial to emphasize that $u_{\text{state}[s]}$ is estimated from the *full* Hybrid model, meaning the variance has already been partitioned against the design effects during estimation. Equation (20) then calculates the resulting probability while neutralizing the contribution of those design effects in the posterior calculation.

Variance decomposition. We apply the law of total variance (Equation (2)) using the probabilities $p_s^{(A)}$. The variance components are calculated by weighting the states by their population shares π_s:

$$\sigma^2_{\text{between},A} = \text{Var}_S\left(p_s^{(A)}\right) = \sum_{s=1}^{S} \pi_s \left(p_s^{(A)} - \bar{p}^{(A)}\right)^2, \tag{21}$$

$$\sigma^2_{\text{within},A} = E_S\left[p_s^{(A)}(1 - p_s^{(A)})\right] = \sum_{s=1}^{S} \pi_s \, p_s^{(A)}\left(1 - p_s^{(A)}\right), \tag{22}$$

where $\bar{p}^{(A)} = \sum_{s=1}^{S} \pi_s p_s^{(A)}$ is the overall population mean under this estimand.

Interpretation. Estimand A provides a "purified" measure of state heterogeneity. It answers the question: *How much do states vary if we conceptually equalize the distribution of design strata and PSUs across states?* By setting the design effects to zero, we remove the variation induced by the specific realization of the sampling design. This yields a conservative, policy-relevant measure of the structural differences between states. Furthermore, as

introduced in Section 3.4, Estimand A can be explicitly de-attenuated (Estimand A*) by subtracting the average design-based sampling variance, providing the most rigorous measure of true structural heterogeneity, net of both design artifacts and finite-sample noise.

4.3. Estimand B (Full-Design): The Descriptive Estimand

Estimand B aims to describe the total observed variance landscape, incorporating the heterogeneity induced by the complex survey design. We refer to this as the *Descriptive Estimand* as it characterizes the data structure and is closer to traditional design-based variance concepts.

Definition. Estimand B incorporates the realized design effects. We first calculate the individual-level probabilities p_i using the full linear predictor:

$$p_i = \text{logit}^{-1}\big(\alpha + u_{\text{state}[i]} + u_{\text{stratum}[i]} + u_{\text{psu}[i]}\big). \tag{23}$$

To obtain the state-level mean $p_s^{(B)}$, we must average these individual probabilities within each state, reflecting the population structure. We achieve this via a weighted mixture, using the scaled survey weights w_i^* (used in the pseudo-likelihood, Section 3.1) as internal weights:

$$p_s^{(B)} = E(p_i \mid S_i = s) \approx \frac{\sum_{i:S_i=s} w_i^* \, p_i}{\sum_{i:S_i=s} w_i^*}. \tag{24}$$

This represents the model-based estimate of the finite-population state mean, accounting for the realized design structure within the state.

Variance decomposition. The between-state variance is the population-weighted variance of the state means $p_s^{(B)}$:

$$\sigma^2_{\text{between},B} = \text{Var}_S\big(p_s^{(B)}\big) = \sum_{s=1}^{S} \pi_s \big(p_s^{(B)} - \bar{p}^{(B)}\big)^2, \tag{25}$$

where $\bar{p}^{(B)} = \sum_{s=1}^{S} \pi_s p_s^{(B)}$.

The within-state variance $\sigma^2_{\text{within},B}$ is the expectation $E_S\{\text{Var}(Y \mid S)\}$. For a mixture distribution within state s, this consists of two parts:

$$\text{Var}(Y \mid S = s) = E_{i|s}\big[p_i(1 - p_i)\big] + \text{Var}_{i|s}(p_i). \tag{26}$$

The first term is the average binomial variance. The second term, $\text{Var}_{i|s}(p_i)$, captures the *mixture variance*—the heterogeneity among individuals within the state induced by the variation in design effects. The overall within-state variance is the population-weighted average of these components:

$$\sigma^2_{\text{within},B} = \sum_{s=1}^{S} \pi_s \Big\{ E_{i|s}\big[p_i(1 - p_i)\big] + \text{Var}_{i|s}(p_i) \Big\}. \tag{27}$$

The expectations and variances in (26) and (27) are calculated using the weighted mixture approach analogous to Equation (24).

Interpretation. Estimand B represents the total geographic heterogeneity present in the population as structured by the design. It answers the question: *How different are the states in practice, given their actual composition of strata and PSUs?* This estimand is useful for descriptive analyses that aim to quantify the total imprint of state differences on the observed data.

4.4. Diagnosing Informative Sampling

The Dual Estimand Framework not only provides nuanced measures of variance but also serves as a powerful diagnostic tool for detecting informative sampling (Section 2.3).

If the sampling design were non-informative, the design effects would be orthogonal to the outcome, conditional on the state. In such a scenario, Estimand A and Estimand B would converge. When they diverge, it indicates that the design features are systematically correlated with the outcome.

We can diagnose the extent of informative sampling by comparing the results:

1. **Comparing overall means ($\bar{p}^{(A)}$ vs. $\bar{p}^{(B)}$).** $\bar{p}^{(B)}$ represents the design-consistent estimate of the population mean, incorporating realized design effects. $\bar{p}^{(A)}$ represents the mean when those effects are marginalized. A significant difference indicates that the sampling design systematically captured areas where the outcome is higher or lower than what state identity alone would predict. For example, if $\bar{p}^{(B)} > \bar{p}^{(A)}$, it suggests the design oversampled areas (PSUs/strata) with higher outcome rates.
2. **Comparing variance components.** A large discrepancy between $\sigma^2_{\text{between},A}$ and $\sigma^2_{\text{between},B}$ (typically the latter being larger) indicates that the design structure significantly contributes to the observed geographic heterogeneity. If the descriptive variance (B) is substantially larger than the policy variance (A), it implies that much of the apparent difference between states is attributable to the confounding effects of the sampling design rather than pure state-level effects.

The divergence between Estimand A and Estimand B quantifies the impact of the informative design and validates the necessity of the Hybrid Framework for correctly separating substantive signal from design-induced noise.

4.5. Practical Guidance for Estimand Selection

In applied research, the choice among estimands depends on the research question. Table 1 summarizes practical guidance for selecting the appropriate estimand.

Table 1. Practical guide for selecting among Estimands A, A*, and B.

Estimand	Research Question	Recommended Use
B: Descriptive	"How different are states in practice, given their actual composition?"	Descriptive monitoring; equity assessments; characterizing total geographic heterogeneity as experienced
A: Policy	"How much variation is attributable to state identity, net of design artifacts?"	Policy evaluation; isolating structural state differences; comparing states on equal footing
A*: De-attenuated	"What is the true substantive between-state variance, net of both design and sampling noise?"	Rigorous policy analysis; contexts with small/unequal domain samples; primary recommended estimand

We recommend reporting both A* and B in applied analyses. The gap between them ($B - A^*$) serves as a diagnostic for the combined impact of informative sampling and finite-sample variance inflation. When this gap is large, as in our NSECE application, it signals that naive analyses would substantially overstate the role of domain-level factors. Section 7.3 provides a complete workflow for applied researchers.

5. Simulation Study

We conducted an extensive simulation study to evaluate the empirical performance of the proposed Bayesian Hybrid Framework. The primary objectives were to assess its

ability to recover the true substantive between-state variance ($\sigma^2_{\text{between}}$) under conditions mimicking the NSECE, specifically addressing the challenges of informative sampling (Section 2.3) and finite-sample variance inflation (Section 2.4).

5.1. Simulation Design and Setup

5.1.1. Data-Generating Process (DGP)

The simulation design is constructed by utilizing the empirical design skeleton of the NSECE analytic sample. We preserved the actual hierarchy of states ($S = 51$), strata (H), and primary sampling units (PSUs), along with the original survey weights and cluster memberships. This approach replicates the realistic challenges of the NSECE, including the crossed structure of states and strata and the highly unequal domain (state) sample sizes.

Within this fixed design structure, we generated new binary outcomes Y_i according to a generalized linear mixed model (GLMM) on the logit scale, consistent with the Hybrid model specification (Equation (13)):

$$\text{logit}\{P(Y_i = 1)\} = \alpha + u_{\text{state}[i]} + u_{\text{stratum}[i]} + u_{\text{psu}[i]}.$$

Random effects were drawn from normal distributions:

$$u_{\text{state}} \sim N(0, \sigma^2_{\text{state}}), \quad u_{\text{stratum}} \sim N(0, \sigma^2_{\text{stratum}}), \quad u_{\text{psu}} \sim N(0, \sigma^2_{\text{psu}}),$$

and the global intercept was fixed at $\alpha = -1.5$.

To ensure identifiability and interpretability of the population parameters, we applied centering constraints during the DGP, analogous to those in Section 3. In particular, state effects were centered using the empirical state population shares π_s such that $\sum_s \pi_s u_{\text{state}[s]} = 0$. Stratum and PSU effects were mean-centered within their respective levels.

The target estimand, or "ground truth", was the substantive population between-state variance on the probability scale. This was calculated by transforming the realized state effects to the probability scale (marginalizing design effects, analogous to the definition of Estimand A in Section 4) and computing their population-weighted variance as in Equation (21).

5.1.2. Manipulated Factors and Scenarios

We implemented a 2^4 factorial design, resulting in 16 unique scenarios. This design crossed four factors to test the methods across a range of realistic conditions, summarized in Table 2.

1. Variance magnitudes. We varied the standard deviations of the random effects on the logit scale. We crossed two levels of state variance ($\sigma_{\text{state}} \in \{0.3, 0.6\}$) with two levels each for stratum variance ($\sigma_{\text{stratum}} \in \{0.4, 0.7\}$) and PSU variance ($\sigma_{\text{psu}} \in \{0.5, 0.9\}$).

2. Informative sampling. We manipulated the relationship between the sampling design and the outcome ($\rho_{\text{inform}} \in \{0, 0.5\}$).

- *Non-informative* ($\rho_{\text{inform}} = 0$): The original weights were used, independent of the outcome realization.
- *Informative* ($\rho_{\text{inform}} = 0.5$): We induced a correlation between the stratum random effects and the inclusion probabilities by adjusting the weights using an exponential tilting mechanism (cf. the stratification-based informativeness in) [22,43],

$$w_i^{\text{adj}} \propto w_i \times \exp\left(-\rho_{\text{inform}} \cdot u_{\text{stratum}[i]}\right).$$

When $\rho_{\text{inform}} > 0$, this mechanism increases the weights for units in strata with negative realized effects (lower outcome prevalence) and decreases the weights for units in strata with positive effects (higher outcome prevalence). This creates the confounding between design features and outcomes described in Section 2.3. Importantly, the outcome model itself remained unchanged; informativeness entered solely through the adjusted survey weights.

We generated $R = 100$ independent replications for each of the 16 scenarios, resulting in 1600 simulated datasets.

Table 2. Simulation design parameters and values.

Parameter	Description	Values
σ_{state}	State effect standard deviation (substantive)	$\{0.3, 0.6\}$
σ_{stratum}	Stratum effect standard deviation (design)	$\{0.4, 0.7\}$
σ_{psu}	PSU effect standard deviation (design)	$\{0.5, 0.9\}$
ρ_{inform}	Informativeness correlation parameter	$\{0, 0.5\}$
Total scenarios	Number of factorial combinations	16 (2^4)
Replications	Number of replications per scenario	100

The parameter values in Table 2 were selected based on three considerations. First, the variance magnitudes (σ_{state}, σ_{stratum}, σ_{psu}) span realistic ranges informed by preliminary estimates from the NSECE data and comparable survey analyses in the education policy literature. The low/high levels capture conditions ranging from modest heterogeneity (where detection is challenging) to substantial heterogeneity (where variance decomposition is clearly warranted). Specifically, $\sigma_{\text{state}} \in \{0.3, 0.6\}$ corresponds approximately to between-state ICCs of 2–10% on the latent scale, consistent with typical findings in social policy research [44]. Second, the informativeness parameter $\rho_{\text{inform}} \in \{0, 0.5\}$ contrasts the ideal scenario (noninformative sampling) with a realistically strong correlation between design features and outcomes, reflecting the NSECE's oversampling of low-income areas. Third, we adopt a fixed-parameter factorial design rather than a fully stochastic simulation (drawing parameters from priors at each replication) for three reasons: (i) fixed parameters enable controlled comparisons of estimator performance under known conditions, facilitating interpretable bias and RMSE calculations; (ii) the 2^4 factorial structure allows systematic assessment of main effects and interactions among the manipulated factors; and (iii) computational feasibility, as each of the 1600 replications already requires fitting a complex Bayesian model via MCMC. A fully stochastic design would average over these conditions, obscuring specific scenarios where methods succeed or fail.

In the simulation study, we employ variational Bayes (VB) approximation for computational tractability given the large number of replications. The real-data application in Section 6 uses full MCMC sampling to ensure valid posterior uncertainty quantification.

5.2. Methods Compared and Evaluation Metrics

5.2.1. Estimation Methods

We compared five distinct approaches for estimating the between-state variance:

1. **Bayesian Hybrid (A).** The Policy Estimand (Section 4). This is calculated from the Hybrid GLMM (Section 3) using Bayesian pseudo-likelihood (BPL), relying on implicit Bayesian shrinkage for stabilization.
2. **Bayesian Hybrid (A*).** The explicitly de-attenuated Policy Estimand (Sections 3 and 4). This applies a method-of-moments correction by subtracting the average design-based sampling variance $\left(\sum_s \pi_s \widehat{V}_s\right)$. The sampling variances $\widehat{V}_s$ were calculated via

Taylor linearization within each simulation replication using the realized sample and adjusted weights.

3. **Bayesian Hybrid (B).** The Descriptive Estimand (Section 4), capturing the total variance landscape including realized design effects.
4. **Naive MLM.** A standard unweighted GLMM with only state-level random effects, ignoring survey weights and the design hierarchy (strata/PSU).
5. **Design-based.** A standard ANOVA-type estimator, calculated as the weighted sample variance of the Horvitz–Thompson estimators for state means (Equation (6)).

The three Hybrid estimands were computed using the same Stan model [45], fitted via mean-field variational Bayes for computational efficiency across the 1600 simulations. The Naive MLM was fitted using `lme4::glmer`, and the design-based estimator was computed using the `survey` package in R [46].

5.2.2. Performance Metrics

Let θ_r denote the true substantive between-state variance in replication r of a given scenario, and $\hat{\theta}_{mr}$ the corresponding estimate from method m. For each method and scenario, we computed the Monte Carlo bias (accuracy) and root mean square error (RMSE, overall efficiency):

$$\text{Bias}_m = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{mr} - \theta_r),$$

$$\text{RMSE}_m = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{mr} - \theta_r)^2}.$$

We also computed the Monte Carlo standard deviation (SD) of the estimates across replications as a measure of sampling variability (precision). For readability, all metrics are presented scaled by 10^3.

5.3. Simulation Results

The results strongly validate the theoretical expectations established in Sections 2–4, demonstrating the superior performance of the Bayesian Hybrid framework, particularly Estimands A and A*, in recovering the substantive between-state variance under complex survey conditions.

5.3.1. Overview of Performance and the Failure of Alternatives

Table 3 provides a quantitative summary of performance metrics aggregated across all 16 simulation scenarios. Figure 1 visualizes the bias and RMSE in key scenarios, contrasting low versus high variance regimes and sampling designs.

The results demonstrate a clear hierarchy in performance. The Bayesian Hybrid Policy Estimands (A and A*) consistently achieve the lowest RMSE (panel B of Figure 1) and exhibit near-zero bias (panel A of Figure 1). Hybrid A* exhibits the lowest overall RMSE across both non-informative (5.39) and informative (5.50) sampling designs (Table 3).

In contrast, the alternative methods exhibit substantial failures, confirming the theoretical challenges discussed in Section 2:

- **Design-based estimator.** This method shows severe positive bias (e.g., 19.65 under non-informative sampling) and the highest RMSE (22.83). This empirically confirms the variance inflation problem (Section 2.4), where sampling noise is conflated with substantive signal (Equation (9)).
- **Naive MLM.** By ignoring the weights and the design structure, the Naive MLM systematically overestimates the between-state variance (mean bias $\approx$ 5.4). It fails to

partition the variance correctly, confounding design-induced clustering with state-level heterogeneity (Section 2.3).

- **Hybrid B (Descriptive Estimand).** Hybrid B exhibits substantial positive bias (mean bias ≈ 12.0) relative to the *substantive* target. This is expected, as Estimand B targets the descriptive variance, which includes realized design effects (Section 4).

Table 3. Summary of performance metrics by method and sampling design.

Method	Non-Informative Sampling				Informative Sampling			
	Mean Bias	SD	RMSE	MAE	Mean Bias	SD	RMSE	MAE
Bayesian Hybrid (A)	−0.23	1.41	6.45	4.97	−0.66	1.39	6.55	5.12
Bayesian Hybrid (A*)	−2.40	1.92	5.39	4.46	−2.62	1.94	5.50	4.60
Bayesian Hybrid (B)	12.35	3.10	13.55	12.36	11.73	2.84	12.92	11.79
Naive MLM	5.46	2.63	6.58	5.64	5.42	2.64	6.56	5.66
Design-based	19.65	3.25	22.83	19.65	14.48	2.12	16.91	14.55

Note. All values are multiplied by 10^3 for readability. Mean and standard deviation (SD) are calculated across the 16 simulation scenarios. RMSE = root mean square error; MAE = mean absolute error.

5.3.2. Addressing Challenge 1: Mitigating Finite-Sample Variance Inflation

A key contribution of the Hybrid Framework is its approach to mitigating the upward bias caused by noisy small-domain estimates (Section 2.4). The framework employs two mechanisms: implicit shrinkage (Estimand A) and explicit de-attenuation (Estimand A*).

Estimand A, through Bayesian partial pooling, already mitigates much of the variance inflation seen in the design-based approach (panel A of Figure 1). However, Estimand A often retains a residual bias when sampling noise is substantial. Figure 2 illustrates the effectiveness of the explicit de-attenuation in A*. Points below the diagonal indicate scenarios where A* successfully reduced the magnitude of the bias present in A.

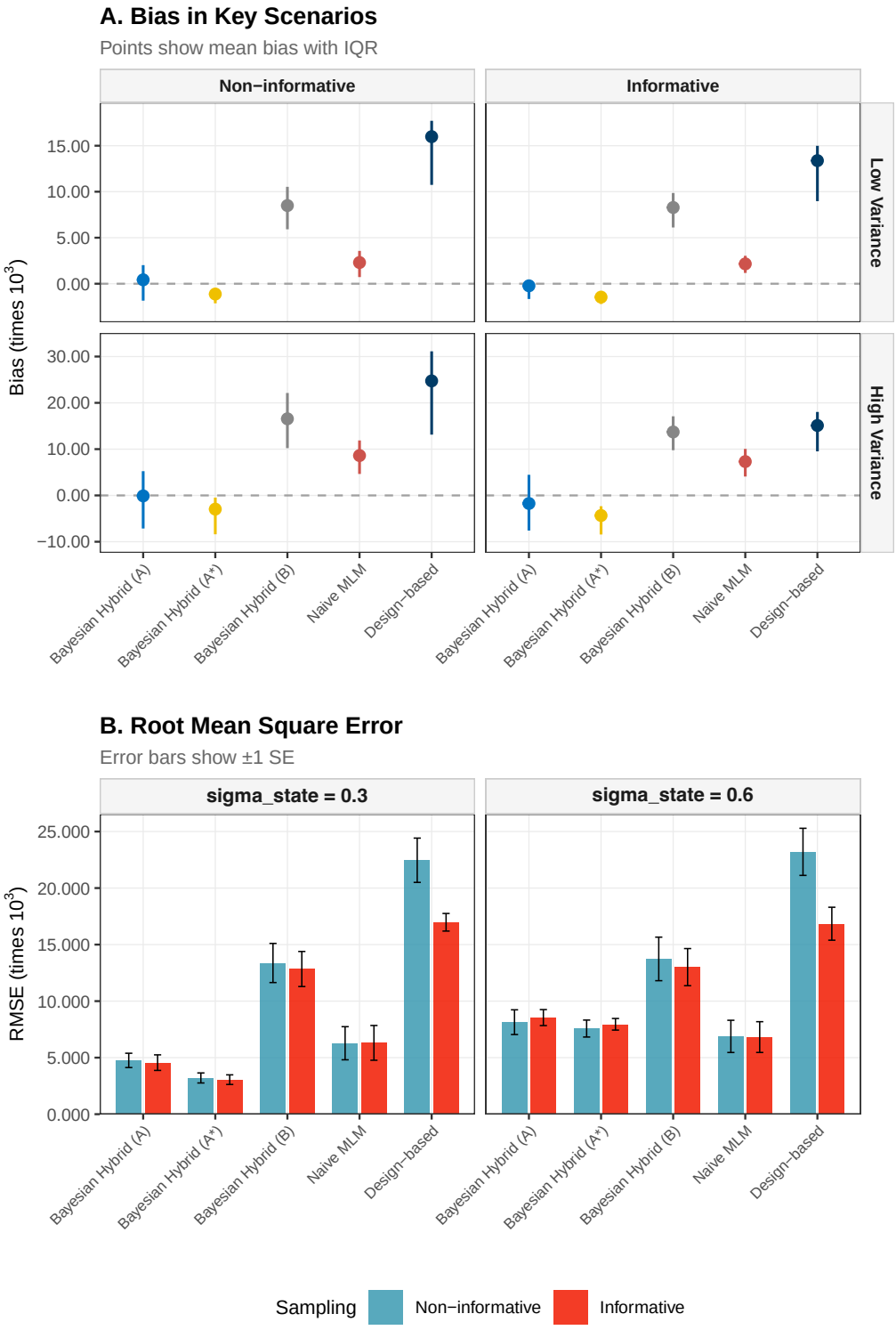
The red regression line (slope < 1) confirms that A* systematically reduces the magnitude of the bias. This correction is effective across both low-variance ($\sigma_{\text{state}} = 0.3$) and high-variance ($\sigma_{\text{state}} = 0.6$) regimes.

There is a trade-off: the explicit correction in A* introduces a slight negative bias on average (Table 3), suggesting a tendency toward over-correction. However, this negative bias is small relative to the positive bias of the alternatives, and A* ultimately achieves the lowest overall RMSE (panel B of Figure 1), indicating that the bias reduction outweighs any associated increase in variance.

5.3.3. Addressing Challenge 2: Sampling Distributions and Small Domains

Figure 3 visualizes the sampling distributions of the estimators in a challenging high-variance scenario ($\sigma_{\text{state}} = 0.6$, $\sigma_{\text{stratum}} = 0.7$, $\sigma_{\text{psu}} = 0.9$). The distributions for Hybrid A and A* are concentrated near the true value (red dashed line). Notably, both Hybrid estimators exhibit pronounced positive skewness, characterized by long right tails; the mean (blue line) is larger than the median (black line). This asymmetry is common for variance estimators bounded at zero, particularly when stabilization mechanisms (such as shrinkage in A or de-attenuation in A*) pull estimates toward the boundary.

In contrast, the design-based and Naive MLM estimators are substantially biased upward and exhibit much higher variability. While their distributions appear relatively symmetric—with the mean and median close together—they are centered far from the true value. This demonstrates the inability of these alternative methods to effectively separate substantive signal from sampling noise, resulting in systematic overestimation.



Note: Low Variance: .=(0.3, 0.4, 0.5); High Variance: .=(0.6, 0.7, 0.9). Results from 100 replications per scenario.

Figure 1. Comparison of bias and root mean square error (RMSE) across estimation methods. Panel (**A**) displays the mean bias with interquartile range across replications in key scenarios. Panel (**B**) displays the RMSE, faceted by the magnitude of true state variance (σ_{state}). Error bars show ±1 standard error.

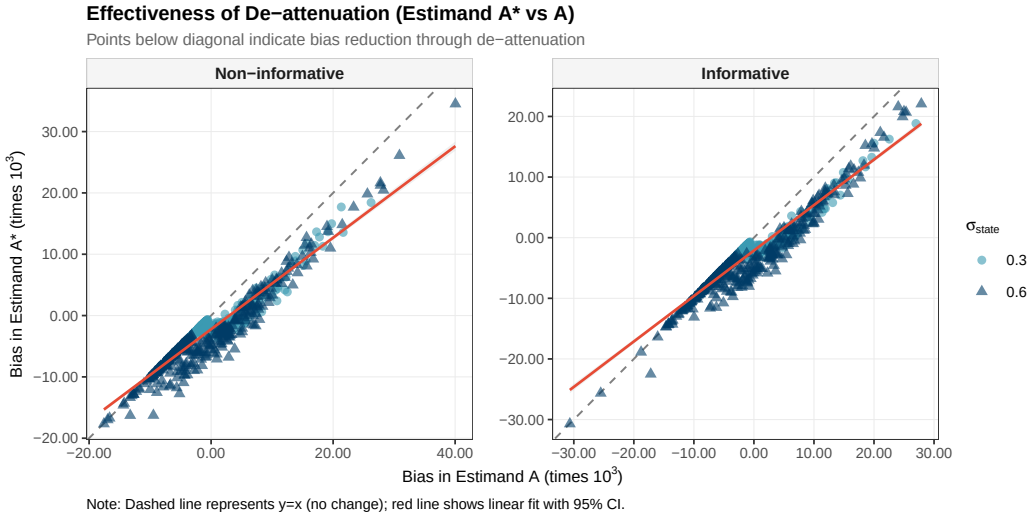


Figure 2. Effectiveness of explicit de-attenuation: comparing bias in Estimand A* versus Estimand A. Each point corresponds to a simulation scenario; points below the diagonal indicate a reduction in bias magnitude through de-attenuation. The red line shows a linear fit with 95% confidence band.

5.3.4. Robustness to Informative Sampling and Bias–Variance Trade-Off

We next evaluate the performance of the methods when the sampling design is informative (Section 2.3). A robust method must account for the correlation induced between the design features (strata) and the outcome via the weights.

Returning to Figure 1, we compare the results between the non-informative ($\rho_{\text{inform}} = 0$) and informative ($\rho_{\text{inform}} = 0.5$) sampling panels. The Bayesian Hybrid Estimands (A and A*) demonstrate remarkable stability: the shift in bias and RMSE when moving from non-informative to informative sampling is minimal (Table 3). For example, the RMSE for Hybrid A* only increases from 5.39 to 5.50.

This robustness is attributable to the two core components of the Hybrid Framework: (i) the use of BPL (Section 3) incorporates weights to ensure design consistency, and (ii) the explicit modeling of design effects (strata and PSUs) (Section 3) partitions the variance, preventing design-induced heterogeneity from contaminating the substantive state effects.

The Naive MLM, lacking these protections, remains systematically biased under informative sampling (mean bias of 5.42; Table 3), as it fails to correct for the systematic differences between the sample and the population.

Figure 4 provides a synthesis of the overall performance by visualizing the bias–variance trade-off across all 16 scenarios. The x-axis represents the absolute bias (accuracy), and the y-axis represents the standard deviation across replications (precision). Superior methods cluster near the origin.

The visualization clearly demonstrates the dominance of the Hybrid Policy Estimands. Hybrid A (blue) and A* (yellow) consistently occupy the southwest region, indicating optimal performance characterized by both low bias and low variance, regardless of whether the sampling is informative (triangles) or non-informative (circles). The curve for A* generally lies closer to the origin than A, confirming its superior RMSE performance.

The Naive MLM (red) exhibits moderate variance but significantly higher bias. The design-based estimator (dark blue) is located far in the northeast region, characterized by both high bias and high variability, illustrating the severe consequences of ignoring finite-sample variance inflation.

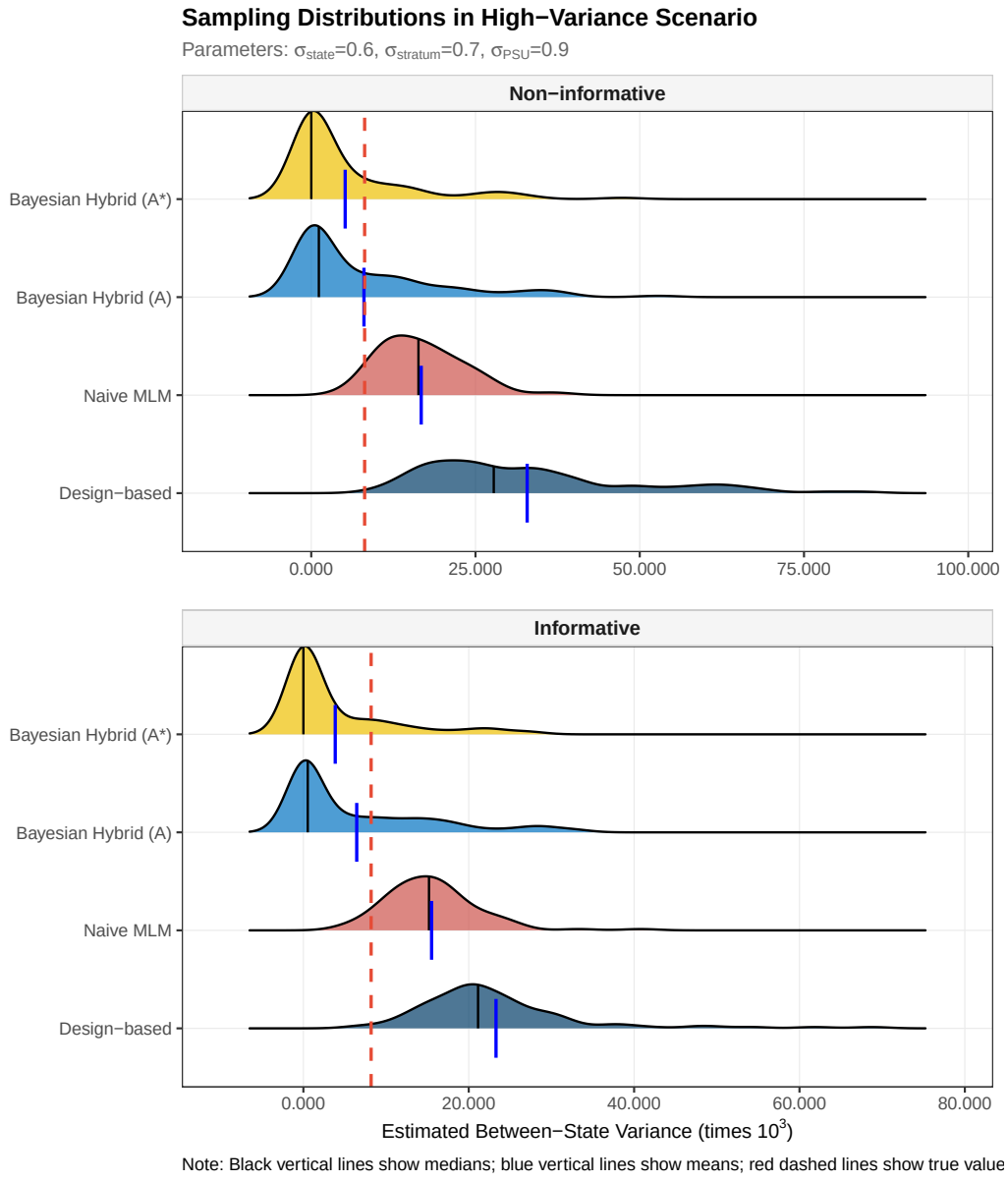


Figure 3. Sampling distributions of estimated between-state variance in a high-variance scenario ($\sigma_{\text{state}} = 0.6$, $\sigma_{\text{stratum}} = 0.7$, $\sigma_{\text{psu}} = 0.9$). Black vertical lines indicate the median; blue vertical lines indicate the mean; the red dashed line indicates the true population value.

Finally, the position of Hybrid B (grey) validates the Dual Estimand Framework (Section 4). It exhibits higher bias relative to the substantive target (by definition) but moderate variance. The consistent gap between the performance frontiers of A* and B underscores the importance of distinguishing between policy-relevant variance (A*) and the total descriptive variance (B).

The simulation study confirms the main theoretical claims of the paper. Under an NSECE-like multistage design with highly unequal domain sizes and informative sampling, the Bayesian Hybrid framework recovers the true substantive between-state variance with negligible bias and substantially lower RMSE than widely used alternatives. Three key findings emerge. First, finite-sample variance inflation is a severe issue for traditional design-based estimators, but it is effectively mitigated by the Hybrid framework's combination of Bayesian shrinkage (Estimand A) and explicit de-attenuation (Estimand A*). Second, robustness to informative sampling requires both appropriate weighting (BPL) and explicit modeling of design features; the Hybrid framework successfully disentangles

substantive variance from design-induced noise, while the Naive MLM fails to do so. Third, the bias–variance trade-off strongly favors the Hybrid approach, particularly Estimand A*, which consistently delivers the best overall performance across a wide range of scenarios. These findings provide strong empirical support for using the Hybrid estimands when analyzing domain-level variance in complex surveys.

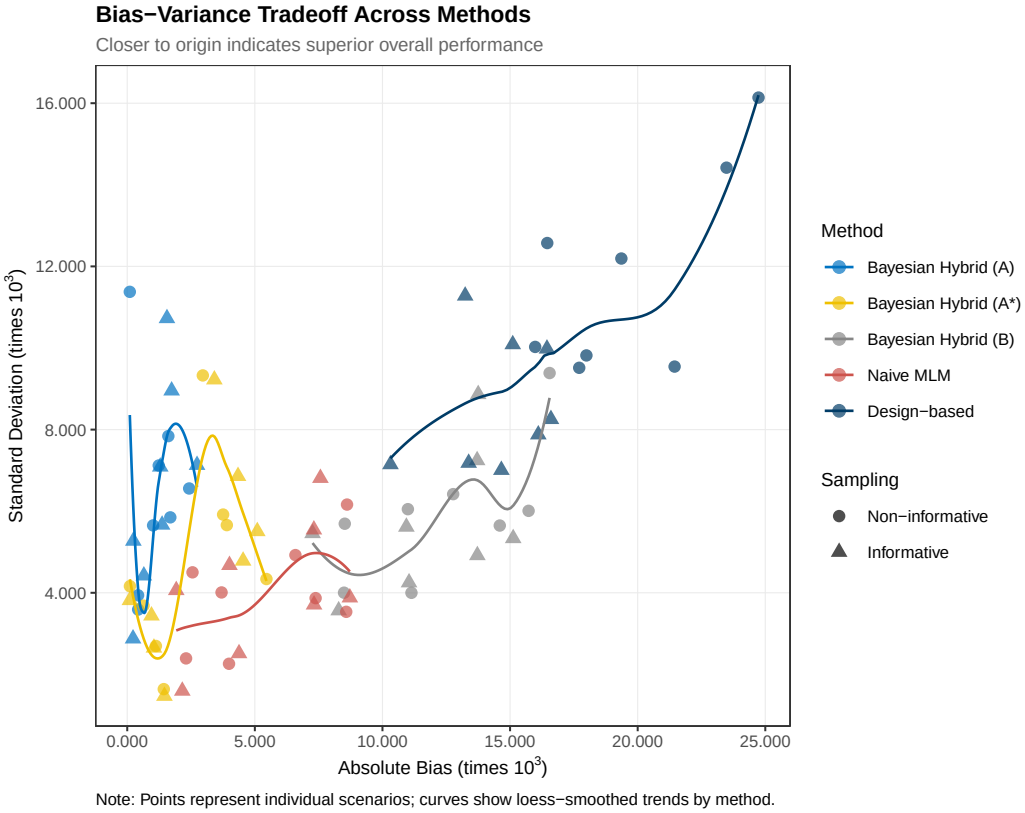


Figure 4. Bias–variance trade-off across methods for estimating between-state variance. Each point represents a simulation scenario, with the x-axis showing absolute bias and the y-axis showing standard deviation across replications. Proximity to the origin indicates superior overall performance (lower mean squared error). Curves show loess-smoothed trends by method.

6. Application: State Variation in Subsidy Receipt Among Listed Home-Based Providers

The simulation study presented in Section 5 validated the efficacy of the Bayesian Hybrid Framework, demonstrating that Estimand A* robustly recovers the true substantive between-domain variance under complex survey conditions characterized by informative sampling and finite-sample noise—conditions where traditional alternatives fail. Building on this validation, we now apply the framework to the 2019 National Survey of Early Care and Education (NSECE) data [11]. This application aims to (1) estimate the substantive between-state variance in subsidy receipt among listed home-based child care (HBCC) providers, (2) empirically diagnose the extent of design complexities in the NSECE data, and (3) interpret the substantive implications of the findings.

6.1. Data, Context, and Design Complexities

The analysis utilizes data from the 2019 NSECE, a nationally representative study designed to characterize the supply of and demand for early care and education (ECE) in the United States. We focus on the analytic sample of *listed* HBCC providers—those identified through state and national administrative lists (e.g., licensing or registry lists). This group represents the segment of HBCC most plausibly reachable by state policy levers.

We exclude unlisted providers, as unlisted unpaid providers cannot accept Child Care and Development Fund (CCDF) subsidies by definition, and very few unlisted paid providers report subsidy revenue.

The outcome variable is a binary indicator of subsidy system participation (1 = yes, 0 = no), constructed based on whether the provider reported receiving government subsidy payments for any enrolled children (derived from the NSECE variable `HB9_RVNU_GOVT_NUMCH_CCSUBS`). Responses indicating the provider "does not receive payment from government sources" were coded as 0, while responses indicating "not enough information to calculate" were treated as missing to ensure accuracy. After excluding cases with missing data on the outcome, the analytic sample consists of $N = 3603$ listed HBCC providers.

The NSECE employs a complex, multistage, stratified, and clustered sampling design [11]. Providers in the analytic sample are distributed across $S = 50$ states, $H = 30$ design strata, and $J = 371$ primary sampling units (PSUs). Table 4 summarizes the key features of the design and the analytic sample.

Table 4. Descriptive statistics of the NSECE 2019 design and analytic sample (listed home-based providers).

Statistic	Value
Total observations (analytic N)	3603
Number of states (S)	50
Number of design strata (H)	30
Number of PSUs (J)	371
State sample size: minimum	1
State sample size: median	39.5
State sample size: maximum	688
Mean survey weight	21.79
Coefficient of variation of weights	1.84
Unweighted prevalence of subsidy receipt	26.9%
Design-weighted prevalence of subsidy receipt	24.7%

Note. The analytic sample includes listed home-based providers with non-missing data on subsidy receipt. S, H, and J refer to the number of states, strata, and PSUs represented in the analytic sample.

The descriptive statistics highlight the statistical challenges formalized in Section 2. First, the potential for informative sampling (Section 2.3) is suggested by the high coefficient of variation (CV) of the survey weights (1.84). A high CV indicates unequal probabilities of selection, resulting from the NSECE's intentional oversampling of low-income areas. The fact that the unweighted prevalence (26.9%) exceeds the design-weighted prevalence (24.7%) is consistent with the sample over-representing providers in high-subsidy communities relative to their population share.

Second, the risk of finite-sample variance inflation (Section 2.4) is evident in the highly unequal state sample sizes, ranging from a minimum of 1 to a maximum of 688, with a median of only 39.5. States with very small samples will yield noisy design-based estimates, necessitating methods that can stabilize these estimates and prevent sampling noise from inflating the overall variance decomposition.

6.2. Implementation of the Bayesian Hybrid Framework

We implemented the Bayesian Hybrid Framework (Section 3) to address these complexities. The model specification follows the Hybrid GLMM (Equation (13)), simultaneously

incorporating random effects for the substantive domains (states) and the design features (strata and PSUs):

$$\text{logit}\{P(Y_i = 1)\} = \alpha + u_{\text{state}[i]} + u_{\text{stratum}[i]} + u_{\text{psu}[i]}.$$

To ensure design consistency under informative sampling, we employed the Bayesian pseudo-likelihood (BPL; Section 3), incorporating scaled survey weights into the likelihood estimation.

To ensure identifiability and correct partitioning of the variance, we applied the centering constraints detailed in Section 3. State effects were centered using population-weighted centering (Equation (15)) based on known state population shares, aligning the intercept α with the population average. PSU effects were centered within their respective strata (Equation (14)). These constraints were implemented via a non-centered parameterization in Stan.

We used a combination of informative and weakly informative priors. For the variance components $(\sigma_{\text{state}}, \sigma_{\text{stratum}}, \sigma_{\text{psu}})$, we specified weakly informative half-Student-t priors (3 degrees of freedom, scale 2.5) [41,47]. For the global intercept α, we used an informative normal prior centered at the logit of the design-weighted national prevalence (24.7%). This prior encourages consistency with the standard design-based estimate of the national mean while remaining broad enough to allow updating from the data.

For the explicit de-attenuation (Estimand A*), we pre-calculated the design-based sampling variances of the state means ($\widehat{V}_s$) using Taylor series linearization. These values were then used in the method-of-moments correction (Equation (19)) within the Bayesian estimation procedure.

The model was estimated using Markov chain Monte Carlo (MCMC) sampling in Stan (four chains, 2000 iterations each). Convergence diagnostics (Rhat and effective sample size, ESS) indicated excellent mixing and convergence for all key parameters (Rhat ≤ 1.01, ESS > 1000; see Table 5), supporting the reliability of the posterior estimates.

Table 5. Posterior summary and convergence diagnostics for key latent-scale parameters (mean [90% credible interval]).

Label	Estimate	SD	Rhat	ESS
Global intercept (log-odds)	−1.589 [−1.759, −1.427]	0.103	1.001	4654
σ_{state} (superpopulation SD, logit)	0.803 [0.357, 1.219]	0.260	1.005	1139
σ_{stratum} (design-stratum SD, logit)	0.760 [0.419, 1.129]	0.219	1.002	1853
σ_{psu} (PSU SD, logit)	1.312 [1.134, 1.505]	0.114	1.001	3722
$\text{ICC}_{\text{state}}$ (latent)	0.110 [0.021, 0.216]	0.060	1.005	1106
$\text{ICC}_{\text{stratum}}$ (latent)	0.097 [0.028, 0.189]	0.050	1.003	1736
ICC_{psu} (latent)	0.272 [0.211, 0.339]	0.039	1.001	2940
s_{state} (finite-pop SD, logit, weighted)	0.756 [0.332, 1.126]	0.241	1.006	1099

Note. Estimates are posterior means with 90% credible intervals in brackets. SD = posterior standard deviation. Rhat and ESS (effective sample size) indicate excellent convergence. Latent ICCs are calculated relative to the total latent variance including the logistic level-1 variance ($\pi^2/3$).

6.3. Results: The Latent Variance Structure and Model Justification

Before interpreting the results on the probability scale, we examine the decomposition of variance on the latent (logit) scale. This allows us to understand the relative contributions of substantive state differences versus design-induced clustering, empirically justifying the need for the Hybrid model structure.

Table 5 presents the posterior summary of the model parameters, standard deviations (SDs), and intraclass correlation coefficients (ICCs) on the logit scale. Figure 5 provides a visual comparison (ANOVA plot) of the posterior distributions of the SDs across the three levels.

The results reveal a clear hierarchy in the sources of variation. The largest component is the PSU-level variation ($\sigma_{\text{psu}} = 1.312$), corresponding to an ICC of 27.2%. This indicates substantial local clustering in subsidy receipt among listed HBCC providers, suggesting that neighborhood characteristics or local implementation factors captured by the PSUs are the dominant drivers of heterogeneity.

The state-level ($\sigma_{\text{state}} = 0.803$; ICC 11.0%) and stratum-level ($\sigma_{\text{stratum}} = 0.760$; ICC 9.7%) variations are moderate and similar in magnitude.

Crucially, the significant magnitude of the design effects (stratum and PSU) confirms the theoretical concerns raised in Section 2. If the model had ignored the design structure (as in a Naive MLM), the substantial variation attributable to strata and PSUs would likely have been confounded with the state-level estimates. The Hybrid model successfully partitions this variance, ensuring that the state effects represent substantive differences conditional on the design structure.

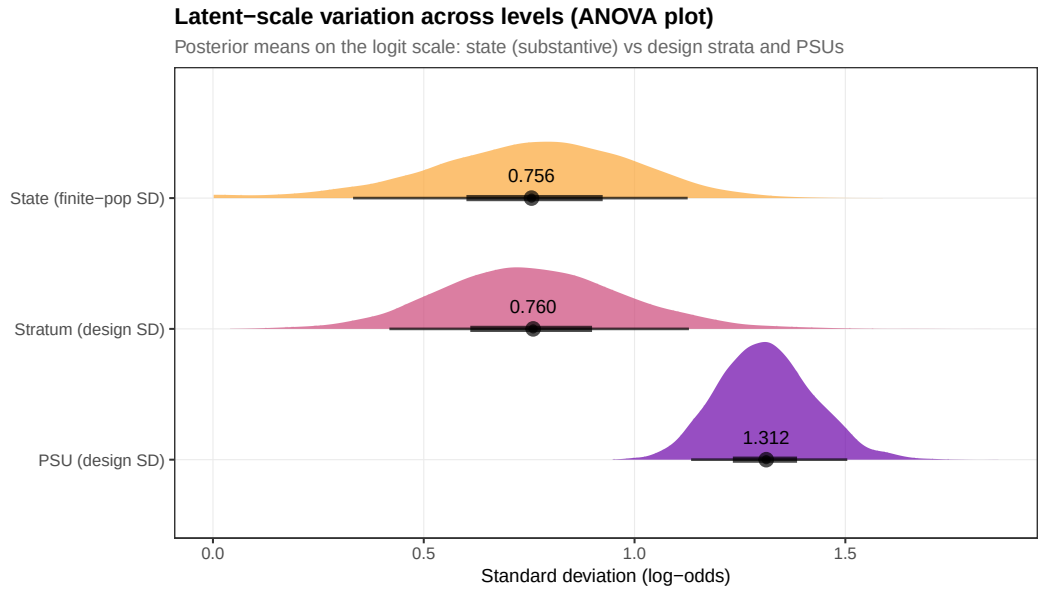


Figure 5. Latent-scale variation across levels (ANOVA plot). Posterior distributions of the standard deviations on the logit scale for state (substantive), stratum (design), and PSU (design) levels. Points indicate posterior means; thick bars show 66% credible intervals; thin bars show 90% credible intervals. The finite-population SD for states (s_{state}) is shown, reflecting the realized variation among the 50 states.

6.4. Results: Probability-Scale Decomposition and the Dual Estimands

We now transition to the probability scale, applying the Dual Estimand Framework (Section 4) to decompose the variance, diagnose the impact of the informative design, and isolate the substantive state-level heterogeneity. Table 6 presents the variance decomposition under Estimands A (Policy), B (Descriptive), and A* (De-attenuated Policy).

Diagnosing informative sampling. The comparison between Estimand A and Estimand B provides strong empirical evidence of informative sampling (Section 4). The overall population mean prevalence under the full-design (Descriptive) estimand, $\bar{p}^{(B)}$, is 24.7%. In contrast, the mean prevalence under the state-only (Policy) estimand, $\bar{p}^{(A)}$, which marginalizes the design effects, is significantly lower at 19.5%.

Table 6. Variance decomposition of subsidy receipt on the probability scale under Estimands A (state-only), A* (de-attenuated), and B (full-design).

Estimand	Component	Estimate
A: State-only (policy)	Population mean prevalence (%)	19.5 [17.1, 22.0]
	Between-state variance	0.013 [0.002, 0.026]
	Within-state variance	0.143 [0.131, 0.156]
	Total variance	0.157 [0.141, 0.172]
	Proportion between-state (%)	8.4 [1.6, 15.5]
B: Full-design (descriptive)	Population mean prevalence (%)	24.7 [23.7, 25.7]
	Between-state variance	0.031 [0.028, 0.035]
	Within-state variance	0.155 [0.150, 0.160]
	Total variance	0.186 [0.181, 0.191]
	Proportion between-state (%)	16.7 [15.0, 18.6]
A*: State-only de-attenuated	Between-state variance	0.008 [0.000, 0.020]
	Within-state variance	0.143 [0.131, 0.156]
	Total variance	0.152 [0.137, 0.166]
	Proportion between-state (%)	5.4 [0.0, 12.6]

Note. Entries are posterior means with 90% credible intervals in brackets.

This substantial difference (over five percentage points) confirms that the NSECE sampling design systematically oversampled areas (strata and PSUs) where subsidy receipt rates among listed HBCC providers are higher than what state identity alone would predict. This aligns with the NSECE's design goal of oversampling low-income areas, which likely have higher subsidy utilization.

Isolating substantive variance from design artifacts. The impact of the design on the variance decomposition is equally striking. The proportion of variance between states under Estimand B is 16.7%. This represents the total observed geographic heterogeneity, including the realized distribution of design features within states.

When we isolate the substantive state heterogeneity using Estimand A, the proportion between states drops by half, to 8.4%. The gap between B and A quantifies the extent to which the observed geographic variation is inflated by the confounding effects of the sampling design. Nearly half of the apparent state-level variation in the descriptive landscape is attributable to the specific realization of strata and PSUs within those states.

Quantifying and correcting finite-sample inflation. Finally, we address the finite-sample variance inflation (Section 2) by comparing Estimand A with the explicitly de-attenuated Estimand A*. Estimand A (8.4%) relies on implicit Bayesian shrinkage to stabilize small-domain estimates. Estimand A* applies an explicit method-of-moments correction, subtracting the average design-based sampling variance.

This correction further reduces the estimated proportion of between-state variance to 5.4%. The difference between A and A* indicates that even after partial pooling, residual sampling noise inflated the variance estimate. This validates the need for explicit de-attenuation, consistent with the findings of the simulation study (Section 5).

In synthesis, the analysis reveals that the true substantive variance attributable to the state level is modest (5.4%). This contrasts sharply with the apparent variation suggested by descriptive approaches (16.7%), demonstrating the critical importance of the Hybrid Framework for separating substantive signal from the artifacts of informative sampling and finite-sample noise.

Interpreting the Variance Reduction. The three-fold reduction from 16.7% (Estimand B) to 5.4% (Estimand A*) carries important substantive implications. The large descriptive variance (B) might initially suggest that state policies—such as CCDF reimbursement rates, eligibility thresholds, or administrative complexity—are major drivers of subsidy participation among listed HBCC providers. However, our analysis reveals that most of

this apparent state-level heterogeneity is attributable to two artifacts: the confounding of state identity with the realized distribution of design features (strata and PSUs), and sampling noise in states with small samples.

The modest substantive variance (5.4%) indicates that true systematic differences across states play a limited role in explaining variation in subsidy receipt. Instead, the dominant source of variation is local—specifically, PSU-level effects account for an ICC of 27.2% on the latent scale (Table 6). This finding suggests that whether a listed HBCC provider participates in the subsidy system depends more on neighborhood characteristics, local administrative infrastructure, community-level outreach, and provider-level factors than on state-wide policy parameters.

For policymakers, this implies that interventions targeting local support systems—such as staffed family child care networks, community-based technical assistance, and streamlined enrollment processes at the county or regional level—may be more effective than incremental adjustments to state-level eligibility criteria or payment schedules. The framework thus redirects attention from *which* states have low participation to *where within states* providers face barriers to subsidy access.

6.5. Results: State-Level Estimates and the Role of Shrinkage

The Bayesian multilevel approach not only provides a robust global variance decomposition but also yields stabilized estimates for individual states. This is particularly important given the highly unequal sample sizes observed in the NSECE data (Table 4).

Figure 6 illustrates the impact of Bayesian shrinkage by comparing the raw design-based state prevalence estimates (Horvitz–Thompson estimators; x-axis) with the model-based estimates under Estimand A (y-axis).

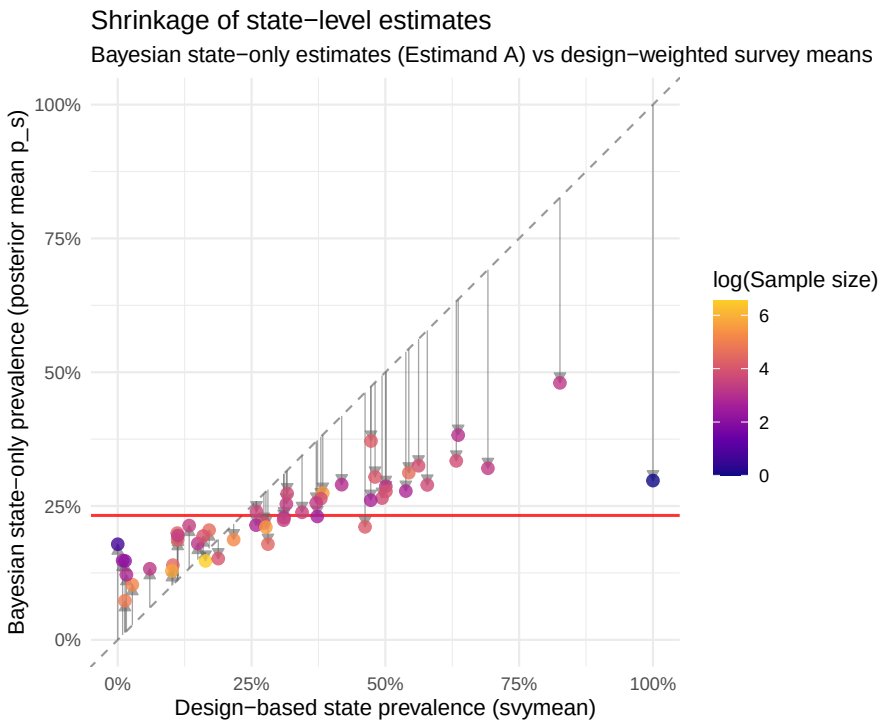


Figure 6. Shrinkage of state-level estimates: Bayesian state-only estimates (Estimand A) versus design-weighted survey means. Each point represents a state. The color indicates the logarithm of the state sample size. Arrows indicate the movement from the design-based estimate to the Bayesian posterior mean. The red horizontal line indicates the overall population mean under Estimand A (19.5%). The dashed diagonal line represents equality.

The plot demonstrates the stabilization provided by partial pooling. States with extreme design-based estimates—those near 0% or approaching 100%—are pulled substantially toward the global mean (the red line at 19.5%). This shrinkage effect is particularly pronounced for states with small sample sizes (indicated by darker colors). For example, several states with design-based estimates exceeding 75% are shrunk to model-based estimates below 50%.

This stabilization directly addresses the challenge of finite-sample noise (Section 2). By pooling information across states, the Bayesian model mitigates the influence of sampling variability in small domains. Importantly, the shrinkage does not erase genuine heterogeneity; the model effectively re-weights the evidence such that persistent, design-adjusted patterns are retained, while idiosyncratic fluctuations in small samples are downweighted, yielding more reliable state-level estimates [48].

6.6. Synthesis of Statistical Findings

The application of the Bayesian Hybrid Framework to the 2019 NSECE data successfully disentangled the complex variance structure of subsidy receipt among listed HBCC providers. The analysis provides strong empirical evidence confirming the methodological challenges outlined in Section 2.

First, the significant divergence between the Descriptive Estimand (B) and the Policy Estimand (A) confirmed the presence of informative sampling. The overall prevalence under Estimand B (24.7%) was notably higher than under Estimand A (19.5%), indicating that the NSECE design systematically oversampled areas with higher subsidy utilization.

Second, the analysis quantified the substantial impact of the design structure on observed geographic heterogeneity. The descriptive between-state variance proportion (Estimand B, 16.7%) was double the policy estimand (Estimand A, 8.4%). This demonstrates that nearly half of the observed variation across states is attributable to the realization of the sampling design (the specific distribution of strata and PSUs) rather than substantive state differences.

Third, the explicit de-attenuation (Estimand A*) further corrected for finite-sample variance inflation. Even after Bayesian shrinkage (Estimand A), residual sampling noise inflated the variance estimate. The correction reduced the final estimate of the substantive between-state variance proportion to a modest 5.4%.

In summary, the Bayesian Hybrid Framework demonstrated its necessity for analyzing geographic domains in the NSECE data. By rigorously accounting for informative sampling and finite-sample noise, the analysis reveals that the true substantive variation attributable to the state level is significantly smaller than descriptive or naive approaches suggest.

To illustrate the practical consequences of our framework, Table 7 contrasts the between-state variance estimates obtained from different analytical approaches.

Table 7. Comparison of between-state variance estimates across methods.

Method	Proportion Between-State (%)	Interpretation
Design-based ANOVA	∼25–30	Severely inflated by sampling noise
Naive MLM (unweighted)	∼15–18	Confounded by design effects
Bayesian Hybrid (B)	16.7	Total geographic heterogeneity
Bayesian Hybrid (A)	8.4	State effects, shrinkage only
Bayesian Hybrid (A*)	5.4	True substantive variance

A naive design-based approach would suggest that roughly one-quarter of the variance in subsidy receipt lies between states—a figure that could mislead policymakers into overestimating the importance of state-level factors. The unweighted multilevel model

partially corrects this but still conflates design-induced clustering with substantive state effects. Only the Bayesian Hybrid Framework, with its explicit modeling of design features and de-attenuation correction, reveals that the true state contribution is less than 6%. This five-fold difference between naive and corrected estimates underscores the necessity of our framework for accurate policy inference from complex survey data.

7. Discussion and Conclusions

The challenge of decomposing variance across geographic domains using complex survey data is pervasive in policy research. When substantive domains (like states) are not explicitly incorporated into the sampling design, analysts face the dual threats of informative sampling and finite-sample variance inflation. This study introduced and validated the Bayesian Hybrid Framework, integrating design-based principles within a multilevel modeling structure to address these challenges. Through simulation and an application to the 2019 NSECE data, we demonstrated the framework's ability to separate substantive signal from design-induced noise, providing a rigorous foundation for geographic variance decomposition.

7.1. Summary of Methodological Contributions

The Bayesian Hybrid Framework offers several key methodological advantages over traditional approaches.

First, it directly addresses informative sampling through a two-pronged strategy. The use of the Bayesian pseudo-likelihood (BPL) incorporates survey weights to achieve design consistency, aligning the model estimates with the target population structure [22]. Simultaneously, the Hybrid GLMM explicitly models design features (strata and PSUs) as random effects. This partitioning ensures that the estimated state effects are conditional on the design structure, preventing the confounding of design-induced clustering with substantive heterogeneity—a critical failure of naive multilevel models.

Second, the framework mitigates finite-sample variance inflation. Implicit Bayesian shrinkage stabilizes estimates in small domains, reducing the impact of sampling noise. The explicit de-attenuation strategy (Estimand A*) provides a further correction by subtracting the average design-based sampling variance. The simulation study confirmed that A* consistently achieved the lowest RMSE and effectively reduced the bias present in estimators relying solely on shrinkage (Estimand A).

Third, the Dual Estimand Framework resolves the ambiguity inherent in defining "between-state variance" under informative designs. By distinguishing between the Policy Estimands (A/A*) and the Descriptive Estimand (B), the framework allows researchers to target the quantity most relevant to their research question. Furthermore, the divergence between A and B serves as a diagnostic tool, quantifying the impact of the informative design on the variance landscape.

In the NSECE application, this framework revealed that the observed geographic heterogeneity (16.7%) was substantially inflated. The most rigorous estimate of substantive between-state variance was significantly lower (5.4%), highlighting the necessity of these methods for accurate policy inference.

7.2. Substantive Implications for ECE Policy

Although not the primary focus of this paper, the empirical findings from the NSECE application offer critical insights into the landscape of subsidy receipt among listed home-based child care (HBCC) providers.

First, the substantive role of state-level differences is limited. The central substantive finding is that systematic differences across states account for a modest proportion (5.4%) of the

variance in subsidy participation (Estimand A*). This suggests that variations in high-level state policies—such as CCDF reimbursement rates, eligibility criteria, or overall state ECE economies—are not the primary drivers of heterogeneity in this sector. This estimate is notably lower than previous analyses that could not fully account for complex survey designs, implying that past research may have overestimated the influence of state-level factors.

Second, local factors are the dominant drivers of variation. The latent-scale analysis (Section 6) revealed that PSU-level effects were the dominant source of variation (ICC 27.2%). This indicates that local clustering is far more consequential than state identity. Factors operating at the local level—such as neighborhood demographics, local families' child care decision-making patterns, or the presence of localized support networks—appear to be the critical mediators between state policy and provider participation.

Policy recommendations. These findings underscore that efforts to increase HBCC participation cannot rely solely on state-level policy adjustments. The dominance of local variation suggests that the realization of state policies is heavily dependent on local context. This highlights the importance of local support structures, such as staffed family child care networks or resource and referral agencies, which often help HBCC providers navigate administrative burdens. Strengthening these local supports may be a more effective strategy than incremental changes to state-level parameters. Furthermore, reducing the complexity of enrollment, reporting, and payment processes, and aligning requirements across different ECE systems (licensing, QRIS, and subsidy) is essential for reducing the administrative burden that disproportionately affects HBCC providers [49].

7.3. Practical Guidance for Applied Researchers

The Bayesian Hybrid Framework provides a flexible toolkit for analyzing domain-level variation in complex surveys. Applied researchers should carefully consider their research objectives when selecting the appropriate estimand.

Choosing among Estimands A, A, and B.* We can conceptualize the Dual Estimand Framework as a menu aligned with distinct research questions:

- *Estimand B (descriptive).* Use B when the goal is to describe the actual pattern of geographic heterogeneity in the population as structured by the survey design. It answers questions like, "How different are states in terms of what households experience in practice?" This is the natural choice for descriptive statistics or monitoring exercises and aligns closely with traditional design-based concepts.
- *Estimand A (policy).* Use A when the interest is in policy-relevant differences that are conceptually independent of the realized configuration of strata and PSUs. It addresses questions such as, "How much variation would remain if states were placed on an equal footing with respect to local composition?" or "How much of the heterogeneity could state-level policy levers plausibly influence?" It provides stabilized, model-based estimates for individual domains via implicit Bayesian shrinkage.
- *Estimand A* (de-attenuated policy).* Use A* when small-domain sample sizes or noisy state estimates make variance inflation a major concern, and when accurate quantification of substantive between-state variance is paramount. By subtracting the average sampling variance, A* provides a conservative, noise-corrected estimate of state-level heterogeneity. This is the preferred estimand for rigorous policy analysis in highly variable designs like the NSECE.

In practice, we recommend reporting both A and B, using their gap as a diagnostic for informative sampling, and presenting A* as the primary estimate when the focus is on policy-relevant variance.

Recommended workflow. Based on the simulations and the NSECE application, we recommend the following workflow (implemented in the `bhfvar` R package (version 0.3.0), https://joonho112.github.io/bhfvar/, accessed on 28 January 2026):

1. *Diagnose design challenges.* Inspect the distribution of survey weights (e.g., coefficient of variation) and domain sample sizes. Large weight variability and highly unequal or small domain sizes signal that naive estimators are likely biased.
2. *Fit the Hybrid GLMM.* Include random effects for both substantive domains and design features (strata, PSUs). Incorporate scaled weights via BPL. Ensure appropriate centering constraints (e.g., population-weighted centering for domains) are applied.
3. *Examine latent-scale results.* Use posterior distributions of SDs and ICCs (ANOVA plots) to understand the relative importance of substantive versus design effects and to justify the model structure.
4. *Compute and interpret the Dual Estimands.* Calculate Estimands A, B, and A* on the probability scale. Compare the results to quantify the impact of the informative design and finite-sample noise, selecting the estimand aligned with the research question.

7.4. Limitations and Future Directions

While the Bayesian Hybrid Framework offers significant advantages, it is subject to several limitations that suggest avenues for future research.

Interpretation of BPL uncertainty and weight scaling. The use of the Bayesian pseudo-likelihood (BPL) aims to achieve design consistency. Savitsky and Toth [15] established that the pseudo-posterior contracts on the true generating distribution in L_1 under certain regularity conditions on the sampling design, providing theoretical justification for this approach. However, the interpretation of the resulting uncertainty estimates (credible intervals) is nuanced. The BPL is not a true likelihood representing the generative model of the population; rather, it aligns the inference with the finite-population estimates. Consequently, the resulting pseudo-posterior does not have the same interpretation as a standard Bayesian posterior under a correctly specified generative model [50]. Williams et al. [30] demonstrate that the survey-weighted pseudo-posterior produces asymptotically incorrect uncertainty estimates, with credible intervals that fail to achieve nominal frequentist coverage [21]. They propose a post-hoc sandwich adjustment that inflates the posterior variance to account for the complex sampling design [51], which represents a promising direction for future integration with our framework.

Furthermore, as noted by Rabe-Hesketh and Skrondal [22], the scaling of level-1 weights in multilevel models can affect parameter estimates, particularly when cluster sizes are small. Although we employed standard scaling practices, sensitivity to alternative scaling methods remains an area for continued investigation.

Reliance on weights and model specification. The framework's design consistency relies on the assumption that the survey weights correctly capture the inclusion probabilities and non-response adjustments. Misspecification of the weights can impact the results. Additionally, the model relies on the appropriateness of the GLMM specification, including the assumption of normality of the random effects on the logit scale.

Extensions to regression models. This study focused on variance decomposition in an intercept-only (ANOVA) model. A critical extension is the application of the Hybrid Framework to models including covariates (e.g., multilevel regression and poststratification [52,53]). In generalized linear mixed models, the estimation of regression coefficients is intrinsically linked to the estimation of variance components. Informative sampling may introduce biases not only in the variance estimates but also in the regression coefficients themselves. Future research should investigate how the Hybrid Framework performs in estimating both conditional and marginal regression effects under complex designs.

We also examined prior sensitivity for the state-level variance component. As shown in Appendix C, posterior summaries for σ_{state} and Estimand A* are stable across a reasonable range of weakly informative priors, leaving our substantive conclusions unchanged.

7.5. Conclusions

Decomposing geographic variation using complex survey data requires statistical methods capable of navigating the inherent complexities of the sampling design. This study demonstrates that failing to account for informative sampling and finite-sample noise leads to severely distorted conclusions. In the application to HBCC subsidy receipt, traditional descriptive estimates suggested that 16.7% of the variance was attributable to the state level. In contrast, the Bayesian Hybrid Framework revealed that the true substantive proportion was only 5.4%, with the majority of variation driven by local (PSU-level) factors.

This difference underscores the importance of integrating design-based principles within model-based inference. The Bayesian Hybrid Framework, with its explicit modeling of design effects and the clarity provided by the Dual Estimand Framework, offers a principled approach for researchers seeking to understand the true sources of heterogeneity in complex populations.

Author Contributions: Conceptualization, J.L. and A.H.; methodology, J.L. and A.H.; software, J.L.; validation, J.L. and A.H.; formal analysis, J.L.; investigation, J.L. and A.H.; resources, J.L. and A.H.; data curation, J.L. and A.H.; writing—original draft preparation, J.L.; writing—review and editing, J.L. and A.H.; visualization, J.L.; supervision, J.L. and A.H.; project administration, J.L. and A.H.; funding acquisition, J.L. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Administration for Children and Families (ACF) of the United States (U.S.) Department of Health and Human Services (HHS) as part of a financial assistance award (Grant #: 90YE0272) totaling $100,000 with 100 percent funded by ACF/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by ACF/HHS, or the U.S. Government.

Institutional Review Board Statement: This study was approved by the Institutional Review Board of The University of Alabama (protocol #18-05-1224).

Informed Consent Statement: Not applicable.

Data Availability Statement: Most variables used in this study are available from the National Survey of Early Care and Education (NSECE) Public-Use Files, United States, 2019 (ICPSR 37941), available at https://www.icpsr.umich.edu/web/ICPSR/studies/37941 (accessed on 28 January 2026). However, the state identifier indicating which state each listed home-based provider belongs to requires access to the Level 1 Restricted-Use Files. The application procedure is available at https://www.icpsr.umich.edu/web/ICPSR/studies/38445 (accessed on 28 January 2026) through the "Access Restricted Data" menu. All R and Stan code used to generate the results and figures in this study are available in a public repository at https://github.com/joonho112/bpl-variance-decomposition (accessed on 28 January 2026). The `bhfvar` R package (version 0.3.0), which implements the complete Bayesian Hybrid Framework for variance decomposition in complex surveys, is available at https://joonho112.github.io/bhfvar/ (accessed on 28 January 2026).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Mathematical Proofs and Derivations

This appendix provides mathematical derivations and proofs supporting key methodological components of the Bayesian Hybrid Framework.

Appendix A.1. Formal Derivation of Finite-Sample Variance Inflation

This section provides a formal derivation of the bias in the naive design-based estimator of between-state variance, $\hat{\sigma}^2_{\text{between}}$ (Equation (6)), due to finite-sample noise, justifying the approximation presented in Equation (9).

Let p_s be the true finite-population mean for state s, and $\hat{p}_s$ be a design-consistent estimator (e.g., the Hájek estimator, Equation (4)). We assume $E_{\text{design}}[\hat{p}_s] \approx p_s$ and let $V_s = \text{Var}_{\text{design}}(\hat{p}_s)$. The true population variance is

$$\sigma^2_{\text{between}} = \sum_{s=1}^{S} \pi_s (p_s - \bar{p})^2. \tag{A1}$$

We analyze the naive estimator, assuming the population shares π_s are known (a common simplification when analyzing the properties of the variance estimator):

$$\hat{\sigma}^2_{\text{between}} = \sum_{s=1}^{S} \pi_s (\hat{p}_s - \hat{\bar{p}})^2, \tag{A2}$$

where $\hat{\bar{p}} = \sum_s \pi_s \hat{p}_s$.

Derivation. First, expand the estimator:

$$\hat{\sigma}^2_{\text{between}} = \sum_s \pi_s (\hat{p}_s)^2 - 2 \sum_s \pi_s \hat{p}_s \hat{\bar{p}} + \sum_s \pi_s (\hat{\bar{p}})^2. \tag{A3}$$

Since $\sum_s \pi_s \hat{p}_s = \hat{\bar{p}}$ and $\sum_s \pi_s = 1$, this simplifies to

$$\hat{\sigma}^2_{\text{between}} = \sum_s \pi_s (\hat{p}_s)^2 - (\hat{\bar{p}})^2. \tag{A4}$$

Now take the expectation over the sampling design, using $E[X^2] = \text{Var}(X) + (E[X])^2$. Assuming approximate unbiasedness,

1. $E_{\text{design}}[(\hat{p}_s)^2] \approx V_s + p_s^2$,
2. $E_{\text{design}}[(\hat{\bar{p}})^2] \approx \text{Var}_{\text{design}}(\hat{\bar{p}}) + \bar{p}^2$.

Substituting into the expectation of the estimator,

$$\begin{aligned} E_{\text{design}}[\hat{\sigma}^2_{\text{between}}] &= \sum_s \pi_s E_{\text{design}}[(\hat{p}_s)^2] - E_{\text{design}}[(\hat{\bar{p}})^2] \\ &\approx \sum_s \pi_s (V_s + p_s^2) - \left(\text{Var}_{\text{design}}(\hat{\bar{p}}) + \bar{p}^2\right) \\ &= \left(\sum_s \pi_s p_s^2 - \bar{p}^2\right) + \sum_s \pi_s V_s - \text{Var}_{\text{design}}(\hat{\bar{p}}). \end{aligned} \tag{A5}$$

The first term is the definition of the true population variance, $\sigma^2_{\text{between}}$. Thus,

$$E_{\text{design}}[\hat{\sigma}^2_{\text{between}}] \approx \sigma^2_{\text{between}} + \sum_{s=1}^{S} \pi_s V_s - \text{Var}_{\text{design}}(\hat{\bar{p}}). \tag{A6}$$

The bias is the average sampling variance, $\sum_s \pi_s V_s$, minus the variance of the overall mean estimator. In large national surveys, $\text{Var}_{\text{design}}(\hat{\bar{p}})$ is typically negligible compared to the average sampling variance of small domains. Neglecting this term yields the approximation in Equation (9):

$$E_{\text{design}}[\hat{\sigma}^2_{\text{between}}] \approx \sigma^2_{\text{between}} + \sum_{s=1}^{S} \pi_s V_s. \tag{A7}$$

Appendix A.2. Justification and Verification of Centering Constraints

Section 3 imposes centering constraints on the random effects for identifiability and interpretability. This appendix justifies the need for these constraints and verifies that the non-centered parameterizations correctly implement them.

Justification of the constraints. We aim for the global intercept α to represent the population average of the linear predictor (log-odds). The linear predictor for a unit in state s, stratum h, and PSU j is

$$\eta_{shj} = \alpha + u_{\text{state}[s]} + u_{\text{stratum}[h]} + u_{\text{psu}[j]}. \tag{A8}$$

The population average $\bar{\eta}$ is the expectation over the population distribution of the indices (s, h, j):

$$\bar{\eta} = E_{\text{pop}}[\eta_{shj}] = \alpha + E_{\text{pop}}[u_{\text{state}[s]}] + E_{\text{pop}}[u_{\text{stratum}[h]}] + E_{\text{pop}}[u_{\text{psu}[j]}]. \tag{A9}$$

To achieve the interpretation $\alpha = \bar{\eta}$, we must ensure that the expectations of all random effects are zero.

For nested effects (PSU within stratum), we use within-group centering (Equation (14)):

$$E\left[u_{\text{psu}[j]} \mid \text{stratum}[h]\right] = 0. \tag{A10}$$

By the law of iterated expectations, this implies $E_{\text{pop}}[u_{\text{psu}[j]}] = 0$. We similarly assume $E_{\text{pop}}[u_{\text{stratum}[h]}] = 0$.

For the substantive effects (states), we must enforce $E_{\text{pop}}[u_{\text{state}[s]}] = 0$. In the finite-population context, this expectation is defined by the population shares π_s, leading to the population-weighted centering constraint (Equation (15)):

$$E_{\text{pop}}[u_{\text{state}[s]}] = \sum_{s=1}^{S} \pi_s u_{\text{state}[s]} = 0. \tag{A11}$$

Verification of implementation: population-weighted centering. The implementation (Equation (16)) is

$$u_{\text{state}[s]} = \sigma_{\text{state}} \left(z_{\text{state}[s]} - \sum_{s'=1}^{S} \pi_{s'} z_{\text{state}[s']} \right). \tag{A12}$$

Let $\bar{z}_\pi = \sum_{s'=1}^{S} \pi_{s'} z_{\text{state}[s']}$. Then the population-weighted sum is

$$\begin{aligned} \sum_{s=1}^{S} \pi_s u_{\text{state}[s]} &= \sigma_{\text{state}} \sum_{s=1}^{S} \pi_s (z_{\text{state}[s]} - \bar{z}_\pi) \\ &= \sigma_{\text{state}} \left(\sum_{s=1}^{S} \pi_s z_{\text{state}[s]} - \sum_{s=1}^{S} \pi_s \bar{z}_\pi \right) \\ &= \sigma_{\text{state}} \left(\bar{z}_\pi - \bar{z}_\pi \sum_{s=1}^{S} \pi_s \right). \end{aligned} \tag{A13}$$

Since $\sum_{s=1}^{S} \pi_s = 1$, this equals $\sigma_{\text{state}}(\bar{z}_\pi - \bar{z}_\pi) = 0$. The constraint is satisfied.

Verification of implementation: within-group centering. The implementation (Equation (17)) for PSU effects within stratum h is

$$u_{\text{psu}[j]} = \sigma_{\text{psu}} \left(z_{\text{psu}[j]} - \bar{z}_{\text{stratum}[h]} \right), \tag{A14}$$

where

$$\bar{z}_{\text{stratum}[h]} = \frac{1}{|J_h|} \sum_{j \in J_h} z_{\text{psu}[j]}, \tag{A15}$$

and J_h is the set of PSU indices in stratum h.

We calculate the sum of PSU effects within stratum h:

$$\begin{aligned} \sum_{j \in J_h} u_{\text{psu}[j]} &= \sigma_{\text{psu}} \sum_{j \in J_h} \left(z_{\text{psu}[j]} - \bar{z}_{\text{stratum}[h]} \right) \\ &= \sigma_{\text{psu}} \left(\sum_{j \in J_h} z_{\text{psu}[j]} - \sum_{j \in J_h} \bar{z}_{\text{stratum}[h]} \right) \\ &= \sigma_{\text{psu}} \left(\sum_{j \in J_h} z_{\text{psu}[j]} - |J_h| \cdot \bar{z}_{\text{stratum}[h]} \right). \end{aligned} \tag{A16}$$

By the definition of $\bar{z}_{\text{stratum}[h]}$, we have $|J_h| \cdot \bar{z}_{\text{stratum}[h]} = \sum_{j \in J_h} z_{\text{psu}[j]}$. Thus, the expression equals 0. The constraint is satisfied.

Appendix A.3. Derivation of Within-State Mixture Variance for Estimand B

Estimand B (Descriptive Estimand, Section 4) incorporates realized design effects. Consequently, individual probabilities p_i vary within a state s. The distribution of the binary outcome Y within state s is a mixture of Bernoulli distributions. We derive its variance (Equation (26)) using the law of total variance.

Derivation. Apply the law of total variance, conditioning on the individual unit i within state s:

$$\text{Var}(Y \mid S = s) = E_{i|s}[\text{Var}(Y \mid i, s)] + \text{Var}_{i|s}[E(Y \mid i, s)]. \tag{A17}$$

For a binary outcome,

1. $E(Y \mid i, s) = p_i$,
2. $\text{Var}(Y \mid i, s) = p_i(1 - p_i)$.

Substituting these components back into the law of total variance gives

$$\text{Var}(Y \mid S = s) = E_{i|s}[p_i(1 - p_i)] + \text{Var}_{i|s}(p_i). \tag{A18}$$

This decomposition shows that the within-state variance in Estimand B comprises the average binomial variance and the mixture variance (the variance of the probabilities themselves, induced by heterogeneity of design effects within the state).

Appendix A.4. Asymptotic Properties and Design Consistency of the Bayesian Pseudo-Likelihood

The Bayesian Hybrid Framework relies on the BPL (Section 3) to achieve design consistency. This appendix summarizes the theoretical justification for the consistency and asymptotic normality of BPL estimators, based on foundational results in the literature (e.g., Savitsky and Toth [15], Williams et al. [30]).

Design consistency. The consistency of the BPL stems from the relationship between the BPL score function and the census estimating equation (CEE). The CEE defines the finite-population parameter $\boldsymbol{\theta}_N$:

$$\mathbf{U}_N(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \mathbf{u}(Y_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{A19}$$

where $\mathbf{u}(\cdot)$ is the score function. The pseudo-score function (gradient of the log BPL) is

$$\hat{\mathbf{U}}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i^* \mathbf{u}(Y_i; \boldsymbol{\theta}). \tag{A20}$$

Proposition A1 (design unbiasedness)**.** *The pseudo-score function is a design-unbiased estimator of the census score function.*

Proof. Assuming $w_i^* \approx 1/\pi_i$, and letting I_i be the sample inclusion indicator,

$$\begin{aligned} E_{\text{design}}[\hat{\mathbf{U}}(\boldsymbol{\theta})] &= E_{\text{design}}\left[\sum_{i \in \mathcal{U}} I_i (1/\pi_i)\, \mathbf{u}(Y_i; \boldsymbol{\theta})\right] \\ &= \sum_{i \in \mathcal{U}} E_{\text{design}}[I_i](1/\pi_i)\, \mathbf{u}(Y_i; \boldsymbol{\theta}) \\ &= \sum_{i \in \mathcal{U}} \pi_i (1/\pi_i)\, \mathbf{u}(Y_i; \boldsymbol{\theta}) \\ &= \mathbf{U}_N(\boldsymbol{\theta}). \end{aligned} \tag{A21}$$

□

Under regularity conditions detailed in Savitsky and Toth [15], this ensures that the pseudo-posterior distribution contracts around $\boldsymbol{\theta}_N$ as $N, n \to \infty$, guaranteeing design-consistent inference.

Asymptotic normality (Bernstein–von Mises for BPL). While the BPL ensures consistency, the shape of the pseudo-posterior distribution differs from the standard Bayesian posterior. The Bernstein–von Mises theorem for misspecified models (e.g., Kleijn and van der Vaart [50]) applies here.

The pseudo-posterior distribution is asymptotically normal, centered at the BPL estimator $\hat{\boldsymbol{\theta}}_{\text{BPL}}$, but with a "sandwich" covariance structure (Williams et al. [30]):

$$\mathbf{V}_{\text{BPL}} = \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}, \tag{A22}$$

where $\mathbf{H}$ is the expected pseudo-information matrix (the curvature of the BPL), and $\mathbf{J}$ is the design-based variance of the pseudo-score function. Under complex sampling, $\mathbf{H} \neq \mathbf{J}$.

Implications. This result implies that raw credible intervals derived directly from the BPL may not achieve nominal frequentist coverage without adjustment. However, in the Hybrid Framework, the explicit modeling of design features (strata and PSUs) directly accounts for much of the design-induced variance, potentially mitigating the discrepancy between $\mathbf{H}$ and $\mathbf{J}$.

Appendix B. Stan Program Documentation

This appendix provides a detailed, block-by-block explanation of the Stan program implementing the Bayesian Hybrid Framework. The program is organized according to Stan's required block structure, and each section explains how the code realizes the methodological components described in the main text. The complete program is available at https://github.com/joonho112/bpl-variance-decomposition (accessed on 20 December 2025).

We use the following conventions: indices $i \in \{1, \ldots, N\}$ denote observations, $s \in \{1, \ldots, S\}$ denote states, $h \in \{1, \ldots, H\}$ denote strata, and $j \in \{1, \ldots, J\}$ denote PSUs (flattened across strata).

Appendix B.1. Data and Transformed Data Blocks

The `data` block declares all inputs required by the model, including the sample structure, outcome, indices, survey weights, and prior hyperparameters.

```
data {
  int<lower=1> N;    // observations
  int<lower=1> S;    // states
  int<lower=1> H;    // strata
  int<lower=1> J;    // total PSUs (flattened)

  array[N] int<lower=0, upper=1> y;           // binary outcome
  array[N] int<lower=1, upper=S> state_id;
  array[N] int<lower=1, upper=H> stratum_id;

  array[H] int<lower=1> J_h;        // # PSUs in stratum h
  array[H] int<lower=1> psu_start;  // start index for stratum h
  array[N] int<lower=1> psu_in_stratum_id;

  vector<lower=0>[N] w_lik;              // pseudo-likelihood weights
  vector<lower=0>[S] w_state_pop_share;  // population shares

  real prior_alpha_mean;
  real<lower=0> prior_alpha_sd;

  int<lower=0, upper=1> use_deattenuation;
  vector<lower=0>[S] vhat_state;  // design-based sampling variance
}
```

The sample structure is defined by the number of observations (N), states (S), strata (H), and PSUs (J). The binary outcome `y` corresponds to Y_i in the main text. Three index arrays map each observation to its state, stratum, and PSU.

PSU nesting structure. PSUs are strictly nested within strata (PSU $\subset$ stratum; Section 3). Rather than using a two-dimensional array, we employ a *flattened representation* for computational efficiency. For stratum h, the PSUs occupy contiguous positions $\texttt{psu\_start}[h], \ldots, \texttt{psu\_start}[h] + J_h[h] - 1$ in the global PSU index space $\{1, \ldots, J\}$. The array `psu_in_stratum_id` provides the within-stratum PSU index (ranging from 1 to J_h) for each observation.

Pseudo-likelihood weights. The vector `w_lik` contains the scaled survey weights w_i^* used in the Bayesian pseudo-likelihood (Equation (12)). Following Section 3, these weights should be normalized so that $\sum_i w_i^* \approx n$ to stabilize variance component estimation.

Population shares. The vector `w_state_pop_share` provides the population shares π_s for each state (Section 2). These are used for both the population-weighted centering of state effects (Equation (15)) and the variance decomposition calculations in the generated quantities block.

De-attenuation inputs. The binary flag `use_deattenuation` enables the explicit de-attenuation correction (Section 3). When enabled, `vhat_state` provides the design-based sampling variances $\widehat{V}_s$ for each state, computed externally via Taylor linearization or replication methods.

The `transformed data` block pre-computes the flattened PSU index for each observation:

```
transformed data {
```

```
  array[N] int psu_flat_id;
  for (i in 1:N) {
    psu_flat_id[i] = psu_start[stratum_id[i]] + psu_in_stratum_id[i] - 1;
  }
}
```

This transformation converts the within-stratum PSU index to the global flattened index. For observation i in stratum h with within-stratum PSU index k, the flattened index is $\texttt{psu\_start}[h] + k - 1$. Pre-computing this mapping avoids redundant calculations during MCMC sampling.

Appendix B.2. Parameters and Transformed Parameters Blocks

The `parameters` block declares the model's free parameters using a non-centered parameterization (Section 3):

```
parameters {
  real alpha;                  // global intercept
  vector[S] z_state;           // standardized state effects
  vector[H] z_stratum;         // standardized stratum effects
  vector[J] z_psu;             // standardized PSU effects

  real<lower=0> sigma_state;   // state SD
  real<lower=0> sigma_stratum; // stratum SD
  real<lower=0> sigma_psu;     // PSU SD
}
```

The global intercept α corresponds to the population-weighted average log-odds. The standardized random effects $z_{\text{state}}, z_{\text{stratum}}, z_{\text{psu}}$ are assigned standard normal priors in the model block, and the corresponding centered effects are constructed deterministically in the `transformed parameters` block. This non-centered parameterization improves MCMC sampling efficiency when data are sparse or variance components are small.

The `transformed parameters` block constructs the centered random effects with the appropriate constraints:

```
transformed parameters {
  vector[S] u_state;
  vector[H] u_stratum;
  vector[J] u_psu;

  // Population-weighted centering for states (Equation (16))
  vector[S] w_state_norm = w_state_pop_share / sum(w_state_pop_share);
  real mean_z_state = dot_product(w_state_norm, z_state);
  u_state = sigma_state * (z_state - mean_z_state);

  // Mean-centering for strata
  u_stratum = sigma_stratum * (z_stratum - mean(z_stratum));

  // Within-stratum centering for PSUs (Equation (17))
  for (h in 1:H) {
    int st  = psu_start[h];
    int Jh  = J_h[h];
    real m;
```

```
    real acc = 0;
    for (j in 0:(Jh - 1)) acc += z_psu[st + j];
    m = acc / Jh;
    for (j in 0:(Jh - 1)) {
      u_psu[st + j] = sigma_psu * (z_psu[st + j] - m);
    }
  }
}
```

State effects: population-weighted centering. The construction of u_{state} implements Equation (16). First, the population shares are normalized to sum to one. The weighted mean of the standardized effects, $\bar{z}_\pi = \sum_s \pi_s z_s$, is computed. Each centered effect is then

$$u_{\text{state}[s]} = \sigma_{\text{state}} \left(z_{\text{state}[s]} - \bar{z}_\pi \right). \tag{A23}$$

As verified in Appendix A.2, this ensures $\sum_s \pi_s u_{\text{state}[s]} = 0$, so the intercept α represents the population-weighted average log-odds.

Stratum effects: simple mean-centering. The stratum effects are centered by subtracting the unweighted mean across strata, ensuring $\sum_h u_{\text{stratum}[h]} = 0$. This is appropriate when strata are treated as exchangeable design features without external population weights.

PSU effects: within-stratum centering. The construction of u_{psu} implements Equation (17). For each stratum h, the code computes the mean of the standardized PSU effects within that stratum, $\bar{z}_h = J_h^{-1} \sum_{j \in J_h} z_j$, and subtracts it:

$$u_{\text{psu}[j]} = \sigma_{\text{psu}} \left(z_{\text{psu}[j]} - \bar{z}_{h(j)} \right). \tag{A24}$$

As verified in Appendix A.2, this ensures $\sum_{j \in J_h} u_{\text{psu}[j]} = 0$ for each stratum h. This within-group centering correctly partitions variance between strata and PSUs, preventing PSU effects from absorbing stratum-level variation.

Appendix B.3. Model Block

The `model` block specifies the priors and the weighted pseudo-likelihood:

```
model {
  // Priors
  alpha         ~ normal(prior_alpha_mean, prior_alpha_sd);
  sigma_state   ~ student_t(3, 0, 2.5);
  sigma_stratum ~ student_t(3, 0, 2.5);
  sigma_psu     ~ student_t(3, 0, 2.5);
  z_state   ~ std_normal();
  z_stratum ~ std_normal();
  z_psu     ~ std_normal();

  // Linear predictor (Equation (13))
  vector[N] eta = alpha
                + u_state[state_id]
                + u_stratum[stratum_id]
                + u_psu[psu_flat_id];

  // Weighted pseudo-likelihood (Equation (12))
  for (i in 1:N) {
    target += w_lik[i] * bernoulli_logit_lpmf(y[i] | eta[i]);
```

```
  }
}
```

Prior distributions. The global intercept receives an informative normal prior centered on the design-weighted overall logit, with the hyperparameters `prior_alpha_mean` and `prior_alpha_sd` supplied as data. The variance components ($\sigma_{\text{state}}, \sigma_{\text{stratum}}, \sigma_{\text{psu}}$) receive weakly informative half-$t_3(0, 2.5)$ priors, which place substantial mass near zero while allowing for larger values if supported by the data. The standardized random effects receive standard normal priors, completing the non-centered parameterization.

Linear predictor. The vectorized construction of η implements Equation (13):

$$\eta_i = \alpha + u_{\text{state}[i]} + u_{\text{stratum}[i]} + u_{\text{psu}[i]}. \tag{A25}$$

Stan's array indexing efficiently extracts the appropriate random effect for each observation.

Bayesian pseudo-likelihood. The likelihood contribution implements Equation (12). For each observation, the log-likelihood is weighted by the survey weight w_i^*:

$$\log L_{\text{BPL}} = \sum_{i \in \mathcal{S}} w_i^* \log P(Y_i \mid \eta_i). \tag{A26}$$

This weighting adjusts the inference to target finite-population quantities, achieving design consistency as established in Appendix A.4. The loop is necessary because Stan does not support vectorized weighted likelihoods; however, the inner function `bernoulli_logit_lpmf` is highly optimized.

Appendix B.4. Generated Quantities Block

The `generated quantities` block computes posterior summaries at each MCMC iteration, including latent-scale ICCs, the dual estimands on the probability scale, finite-population standard deviations, and diagnostic quantities.

Appendix B.4.1. Latent-Scale Intraclass Correlations

```
real var_level1 = square(pi()) / 3.0;
real var_total_latent = square(sigma_state) + square(sigma_stratum)
                      + square(sigma_psu) + var_level1;
real icc_state   = square(sigma_state)   / var_total_latent;
real icc_stratum = square(sigma_stratum) / var_total_latent;
real icc_psu     = square(sigma_psu)     / var_total_latent;
```

For binary outcomes with a logit link, the level-1 (residual) variance on the latent scale is $\pi^2/3 \approx 3.29$, the variance of the standard logistic distribution. The total latent variance is the sum of all variance components:

$$\sigma^2_{\text{total, latent}} = \sigma^2_{\text{state}} + \sigma^2_{\text{stratum}} + \sigma^2_{\text{psu}} + \pi^2/3. \tag{A27}$$

The intraclass correlation for each level is the proportion of total latent variance attributable to that level. These ICCs provide a model-based decomposition on the latent (logit) scale, complementing the probability-scale estimands.

Appendix B.4.2. Estimand A: State-Only (Policy) Decomposition

```
vector[S] p_state_A;
for (s in 1:S) p_state_A[s] = inv_logit(alpha + u_state[s]);

vector[S] wS = w_state_pop_share / sum(w_state_pop_share);
```

```
real p_bar_A = dot_product(wS, p_state_A);

real var_within_A = 0.0;
for (s in 1:S) var_within_A += wS[s] * p_state_A[s] * (1.0 - p_state_A[s]);

real var_between_A = 0.0;
for (s in 1:S) var_between_A += wS[s] * square(p_state_A[s] - p_bar_A);

real var_total_A = var_within_A + var_between_A;
real prop_between_A = (var_total_A > 0) ? var_between_A / var_total_A : 0;
```

Estimand A (Section 4) marginalizes the design effects by setting them to zero. For each state s, the probability is computed using only the intercept and state effect (Equation (20)):

$$p_s^{(A)} = \text{logit}^{-1}(\alpha + u_{\text{state}[s]}). \tag{A28}$$

The variance decomposition follows Equations (21) and (22). The within-state variance is the population-weighted average binomial variance:

$$\sigma^2_{\text{within},A} = \sum_{s=1}^{S} \pi_s \, p_s^{(A)}(1 - p_s^{(A)}). \tag{A29}$$

The between-state variance is the population-weighted variance of the state probabilities:

$$\sigma^2_{\text{between},A} = \sum_{s=1}^{S} \pi_s \left(p_s^{(A)} - \bar{p}^{(A)}\right)^2. \tag{A30}$$

De-attenuation (Estimand A).* When enabled, the explicit de-attenuation correction (Section 3, Equation (19)) subtracts the weighted average of the design-based sampling variances:

```
if (use_deattenuation == 1) {
  real mean_v = dot_product(wS, vhat_state);
  var_between_A_deatt = fmax(0, var_between_A - mean_v);
} else {
  var_between_A_deatt = var_between_A;
}
```

The `fmax(0, ·)` ensures the de-attenuated variance is non-negative, as required for a variance estimate.

Appendix B.4.3. Estimand B: Full-Design (Descriptive) Decomposition

```
vector[N] eta_full = alpha + u_state[state_id]
                   + u_stratum[stratum_id] + u_psu[psu_flat_id];
vector[N] p_i;
for (i in 1:N) p_i[i] = inv_logit(eta_full[i]);

// State means via weighted mixture (Equation (24))
vector[S] p_state_B;
vector[S] state_w_obs;
for (s in 1:S) state_w_obs[s] = 0;
for (i in 1:N) state_w_obs[state_id[i]] += w_lik[i];
```

```
for (s in 1:S) {
  real den = state_w_obs[s];
  if (den > 0) {
    real num = 0;
    for (i in 1:N) if (state_id[i] == s) num += w_lik[i] * p_i[i];
    p_state_B[s] = num / den;
  } else {
    p_state_B[s] = inv_logit(alpha + u_state[s]);
  }
}
```

Estimand B (Section 4) incorporates the realized design effects. First, individual probabilities are computed using the full linear predictor (Equation (23)):

$$p_i = \text{logit}^{-1}(\alpha + u_{\text{state}[i]} + u_{\text{stratum}[i]} + u_{\text{psu}[i]}). \tag{A31}$$

The state mean $p_s^{(B)}$ is computed as a weighted average of the individual probabilities within each state (Equation (24)):

$$p_s^{(B)} = \frac{\sum_{i:S_i=s} w_i^* \, p_i}{\sum_{i:S_i=s} w_i^*}. \tag{A32}$$

For states without sampled observations (an edge case), the code falls back to the Estimand A definition for numerical stability.

The within-state variance for Estimand B includes both the average binomial variance and the mixture variance induced by heterogeneous design effects within states (Equations (26) and (27)):

```
var_within_B = 0.0;
for (s in 1:S) {
  real den = state_w_obs[s];
  real term_mean_binvar = 0.0;
  real term_mixture_var = 0.0;

  if (den > 0) {
    for (i in 1:N) if (state_id[i] == s) {
      real wnorm = w_lik[i] / den;
      term_mean_binvar += wnorm * (p_i[i] * (1 - p_i[i]));
      term_mixture_var += wnorm * square(p_i[i] - p_state_B[s]);
    }
  } else {
    real ps = inv_logit(alpha + u_state[s]);
    term_mean_binvar = ps * (1 - ps);
  }
  var_within_B += wS[s] * (term_mean_binvar + term_mixture_var);
}
```

Appendix B.4.4. Finite-Population Standard Deviations

The code computes both unweighted and population-weighted finite-population standard deviations of state effects on both the logit and probability scales:

```
real s_state_logit_unw = sd(u_state);
```

```
real s_state_logit_w;
{
  vector[S] wS = w_state_pop_share / sum(w_state_pop_share);
  real mu = dot_product(wS, u_state);
  real acc = 0;
  for (s in 1:S) acc += wS[s] * square(u_state[s] - mu);
  s_state_logit_w = sqrt(fmax(0, acc));
}
```

The unweighted SD treats all states equally, while the weighted SD accounts for population shares. These quantities complement the variance decomposition by providing direct measures of state-effect heterogeneity.

Appendix B.4.5. Diagnostics: Weighted Log-Likelihood and Posterior Predictive Checks

```
vector[N] log_lik;
array[N] int y_rep;
for (i in 1:N) {
  real e = alpha + u_state[state_id[i]]
         + u_stratum[stratum_id[i]] + u_psu[psu_flat_id[i]];
  log_lik[i] = w_lik[i] * bernoulli_logit_lpmf(y[i] | e);
  y_rep[i]   = bernoulli_logit_rng(e);
}
```

The pointwise log-likelihood values `log_lik` are weighted by the survey weights, consistent with the pseudo-likelihood formulation. These can be used for model comparison via pseudo-LOO or WAIC, though their interpretation differs from standard information criteria due to the weighting.

The replicated data `y_rep` are generated from the posterior predictive distribution (unweighted) and can be used for graphical posterior predictive checks to assess model fit.

Appendix C. Prior Sensitivity Analysis

To assess the robustness of our conclusions to prior specification for variance components, we re-estimated the Bayesian Hybrid model for the NSECE application under three alternative prior distributions for the state-level standard deviation σ_{state}, keeping all other model components (likelihood, design features, and priors for remaining parameters) identical to the main analysis.

1. **Baseline (main analysis)**: half-Student-$t(3, 2.5)$
2. **More concentrated**: half-Normal$(0, 1)$
3. **Heavier tails**: half-Cauchy$(0, 2.5)$
4. **Wider scale**: half-Student-$t(3, 5)$

Each model was fit using Hamiltonian Monte Carlo (NUTS) in Stan (4 chains; 1000 warmup and 1000 sampling iterations per chain; `adapt_delta` = 0.95; `max_treedepth` = 12). Table A1 reports posterior means and 90% credible intervals for σ_{state} and the key substantive quantity of interest (Estimand A*: the de-attenuated between-state proportion on the probability scale).

Table A1. Prior sensitivity for the state-level SD and the de-attenuated between-state proportion (Estimand A*). Entries are posterior means with 90% credible intervals, estimated via MCMC (HMC/NUTS).

Prior for σ_{state}	σ_{state} (logit SD)	Estimand A* (Between-State %, de-att.)
Baseline: half-Student-$t(3, 2.5)$	0.812 [0.351, 1.248]	5.6 [0.0, 12.9]
More concentrated: half-Normal$(0, 1)$	0.763 [0.317, 1.165]	4.8 [0.0, 12.0]
Heavier tails: half-Cauchy$(0, 2.5)$	0.821 [0.416, 1.235]	5.7 [0.0, 13.3]
Wider scale: half-Student-$t(3, 5)$	0.837 [0.434, 1.247]	5.8 [0.0, 13.1]

Across these reasonable prior perturbations, the posterior mean of σ_{state} ranges from 0.763 to 0.837, with a maximum relative deviation of 6.1% from the baseline specification. The posterior mean of Estimand A* ranges from 4.8% to 5.8%, a spread of 1.0 percentage points. Credible intervals overlap substantially across specifications, indicating that the posterior is largely data-driven in this application.

All MCMC runs exhibited good computational behavior (no divergent transitions and no maximum-treedepth saturations), and convergence diagnostics were satisfactory (for σ_{state}, $\widehat{R} \leq 1.01$ with effective sample size ≥ 540). Overall, the qualitative conclusion is unchanged: after de-attenuation, true between-state heterogeneity accounts for only a modest share of total variation in subsidy receipt. Nevertheless, for applications with fewer domains or smaller within-domain samples, we recommend conducting similar sensitivity analyses because variance components can be more weakly identified and thus more prior-dependent.

References

1. Albers, B.; Mildon, R.; Lyon, A.; Shlonsky, A. Implementation frameworks in child, youth and family services—Results from a scoping review. *Child. Youth Serv. Rev.* **2017**, *81*, 101–116. [CrossRef]
2. Baumann, A.; Shelton, R.; Kumanyika, S.; Haire-Joshu, D. Advancing healthcare equity through dissemination and implementation science. *Health Serv. Res.* **2023**, *58*, 327–344. [CrossRef] [PubMed]
3. Dunbar, P.; Keyes, L.; Browne, J. Determinants of regulatory compliance in health and social care services: A systematic review using the Consolidated Framework for Implementation Research. *PLoS ONE* **2022**, *18*, e0278007. [CrossRef]
4. Nilsen, P. Making sense of implementation theories, models and frameworks. *Implement. Sci.* **2015**, *10*, 53. [CrossRef]
5. Shankardass, K.; Muntaner, C.; Kokkinen, L.; Shahidi, F.; Freiler, A.; Oneka, G.; Bayoumi, A.; O'Campo, P. The implementation of Health in All Policies initiatives: A systems framework for government action. *Health Res. Policy Syst.* **2018**, *16*, 26. [CrossRef]
6. Connors, M.; Morris, P. Comparing state policy approaches to early care and education quality: A multidimensional assessment of quality rating and improvement systems and child care licensing regulations. *Early Child. Res. Q.* **2015**, *30*, 266–279. [CrossRef]
7. Jenkins, J.; Nguyen, T. Keeping kids in care: Reducing administrative burden in state Child Care Development Fund policy. *J. Public Adm. Res. Theory* **2021**, *32*, 23–40. [CrossRef]
8. Prusinski, E.; Mahler, P.; Collins, M.; Couch, H. Strengthening early childhood education and care in a "childcare desert". *Early Child. Educ. J.* **2022**, *51*, 1317–1333. [CrossRef]
9. Raikes, H.; Torquati, J.; Wang, C.; Shjegstad, B. Parent experiences with state child care subsidy systems and their perceptions of choice and quality in care selected. *Early Educ. Dev.* **2012**, *23*, 558–582. [CrossRef]
10. Rigby, E.; Ryan, R.; Brooks-Gunn, J. Child care quality in different state policy contexts. *J. Policy Anal. Manag.* **2007**, *26*, 887–908. [CrossRef]
11. National Survey of Early Care and Education Project Team. *2019 National Survey of Early Care and Education Data Collection and Sampling Methodology Report*; OPRE Report 2022-118; Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services: Washington, DC, USA, 2022.
12. Ambade, P.; Hoffman, Z.; Mehra, K.; Gunja, M.; Yi, M.; MacKinnon, B.; MacKinnon, N. Hospital discharge communication problems in 10 high-income nations: A secondary analysis of an international health policy survey. *BMJ Open* **2025**, *15*, e094724. [CrossRef]
13. Gangopadhyaya, A.; Karpman, M. The impact of Arkansas Medicaid work requirements on coverage and employment: Estimating effects using national survey data. In *Health Services Research*; Wiley: Hoboken, NJ, USA, 2025; p. e14624. [CrossRef]

14. Yang, C.; Zhang, N.; Gao, T.; Zhu, Y.; Gong, C.; Xu, M.; Feng, C. Association between social determinants of health and premature atherosclerotic cardiovascular disease and sex differences in US adults: A cross-sectional study. *Prev. Med. Rep.* **2025**, *50*, 102967. [CrossRef]
15. Savitsky, T.D.; Toth, D. Bayesian estimation under informative sampling. *Electron. J. Stat.* **2016**, *10*, 1677–1708. [CrossRef]
16. Makela, S.; Si, Y.; Gelman, A. Bayesian inference under cluster sampling with probability proportional to size. *Stat. Med.* **2018**, *37*, 3849–3868. [CrossRef]
17. Rao, J.N.K.; Molina, I. *Small Area Estimation*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2015. [CrossRef]
18. Fay, R.E.; Herriot, R.A. Estimates of Income for Small Places: An Application of James–Stein Procedures to Census Data. *J. Am. Stat. Assoc.* **1979**, *74*, 269–277. [CrossRef]
19. Battese, G.E.; Harter, R.M.; Fuller, W.A. An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *J. Am. Stat. Assoc.* **1988**, *83*, 28–36. [CrossRef]
20. Pfeffermann, D. New Important Developments in Small Area Estimation. *Stat. Sci.* **2013**, *28*, 40–68. [CrossRef]
21. León-Novelo, L.G.; Savitsky, T.D. Fully Bayesian estimation under dependent and informative cluster sampling. *J. Surv. Stat. Methodol.* **2023**, *11*, 484–510. [CrossRef]
22. Rabe-Hesketh, S.; Skrondal, A. Multilevel modelling of complex survey data. *J. R. Stat. Soc. Ser. A* **2006**, *169*, 805–827. [CrossRef]
23. Skinner, C.J. Domain means, regression and multi-variate analysis. In *Analysis of Complex Surveys*; Skinner, C.J., Holt, D., Smith, T.M.F., Eds.; Wiley: Chichester, UK, 1989; pp. 59–88.
24. Isaki, C.T.; Fuller, W.A. Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.* **1982**, *77*, 89–96. [CrossRef]
25. Grilli, L.; Pratesi, M. Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Surv. Methodol.* **2004**, *30*, 93–103.
26. Little, R.J. To model or not to model? Competing modes of inference for finite population sampling. *J. Am. Stat. Assoc.* **2004**, *99*, 546–556. [CrossRef]
27. Gelman, A. Struggles with survey weighting and regression modeling (with discussion). *Stat. Sci.* **2007**, *22*, 153–188.
28. Chen, Q.; Elliott, M.R.; Haziza, D.; Yang, Y.; Ghosh, M.; Little, R.J.A.; Sedransk, J.; Thompson, M. Approaches to improving survey-weighted estimates. *Stat. Sci.* **2017**, *32*, 227–248. [CrossRef]
29. Han, Q.; Wellner, J.A. Complex sampling designs: Uniform limit theorems and applications. *Ann. Stat.* **2021**, *49*, 459–485. [CrossRef]
30. Williams, M.R.; McGuire, F.H.; Savitsky, T.D. Uncertainty quantification for multi-level models using the survey-weighted pseudo-posterior. *arXiv* **2025**, arXiv:2510.09401.
31. Lohr, S.L. *Sampling: Design and Analysis*, 3rd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2022. [CrossRef]
32. Pfeffermann, D. The role of sampling weights when modeling survey data. *Int. Stat. Rev.* **1993**, *61*, 317–337. [CrossRef]
33. Pfeffermann, D. The use of sampling weights for survey data analysis. *Stat. Methods Med Res.* **1996**, *5*, 239–261. [CrossRef] [PubMed]
34. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
35. Little, R.J.A. Models for nonresponse in sample surveys. *J. Am. Stat. Assoc.* **1982**, *77*, 237–250. [CrossRef]
36. Korn, E.L.; Graubard, B.I. Estimating variance components by using survey data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2003**, *65*, 175–190. [CrossRef]
37. Longford, N.T. Model-based variance estimation in surveys with stratified clustered designs. *Aust. J. Stat.* **1996**, *38*, 333–352. [CrossRef]
38. Binder, D.A. On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **1983**, *51*, 279–292. [CrossRef]
39. Chambers, R.L., Skinner, C.J., (Eds.) *Analysis of Survey Data*; Wiley: Chichester, UK, 2003.
40. Breslow, N.E.; Clayton, D.G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25. [CrossRef]
41. Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **2006**, *1*, 515–534. [CrossRef]
42. Papaspiliopoulos, O.; Roberts, G.O.; Sköld, M. A general framework for the parametrization of hierarchical models. *Stat. Sci.* **2007**, *22*, 59–73. [CrossRef]
43. Pfeffermann, D.; Skinner, C.J.; Holmes, D.J.; Goldstein, H.; Rasbash, J. Weighting for unequal selection probabilities in multilevel models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1998**, *60*, 23–40. [CrossRef]
44. Raudenbush, S.W.; Bryk, A.S. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed.; Sage Publications: Thousand Oaks, CA, USA, 2002.
45. Carpenter, B.; Gelman, A.; Hoffman, M.D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan: A probabilistic programming language. *J. Stat. Softw.* **2017**, *76*, 1–32. [CrossRef]
46. Lumley, T. Analysis of complex survey samples. *J. Stat. Softw.* **2004**, *9*, 1–19. [CrossRef]

47. Gelman, A.; Jakulin, A.; Pittau, M.G.; Su, Y.S. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2008**, *2*, 1360–1383. [CrossRef]
48. Si, Y.; Pillai, N.S.; Gelman, A. Bayesian nonparametric weighted sampling inference. *Bayesian Anal.* **2015**, *10*, 605–625. [CrossRef]
49. Adams, G.; Dwyer, K. *Child Care Subsidies and Home-Based Child Care Providers: Expanding Participation*; Technical Report; Urban Institute: Washington, DC, USA, 2021.
50. Kleijn, B.; van der Vaart, A. The Bernstein-von Mises theorem under misspecification. *Electron. J. Stat.* **2012**, *6*, 354–381. [CrossRef]
51. Williams, M.R.; Savitsky, T.D. Uncertainty estimation for pseudo-Bayesian inference under complex sampling. *Int. Stat. Rev.* **2021**, *89*, 72–107. [CrossRef]
52. Gelman, A.; Little, T.C. Poststratification into many categories using hierarchical logistic regression. *Surv. Methodol.* **1997**, *23*, 127–135.
53. Park, D.K.; Gelman, A.; Bafumi, J. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Anal.* **2004**, *12*, 375–385. [CrossRef]