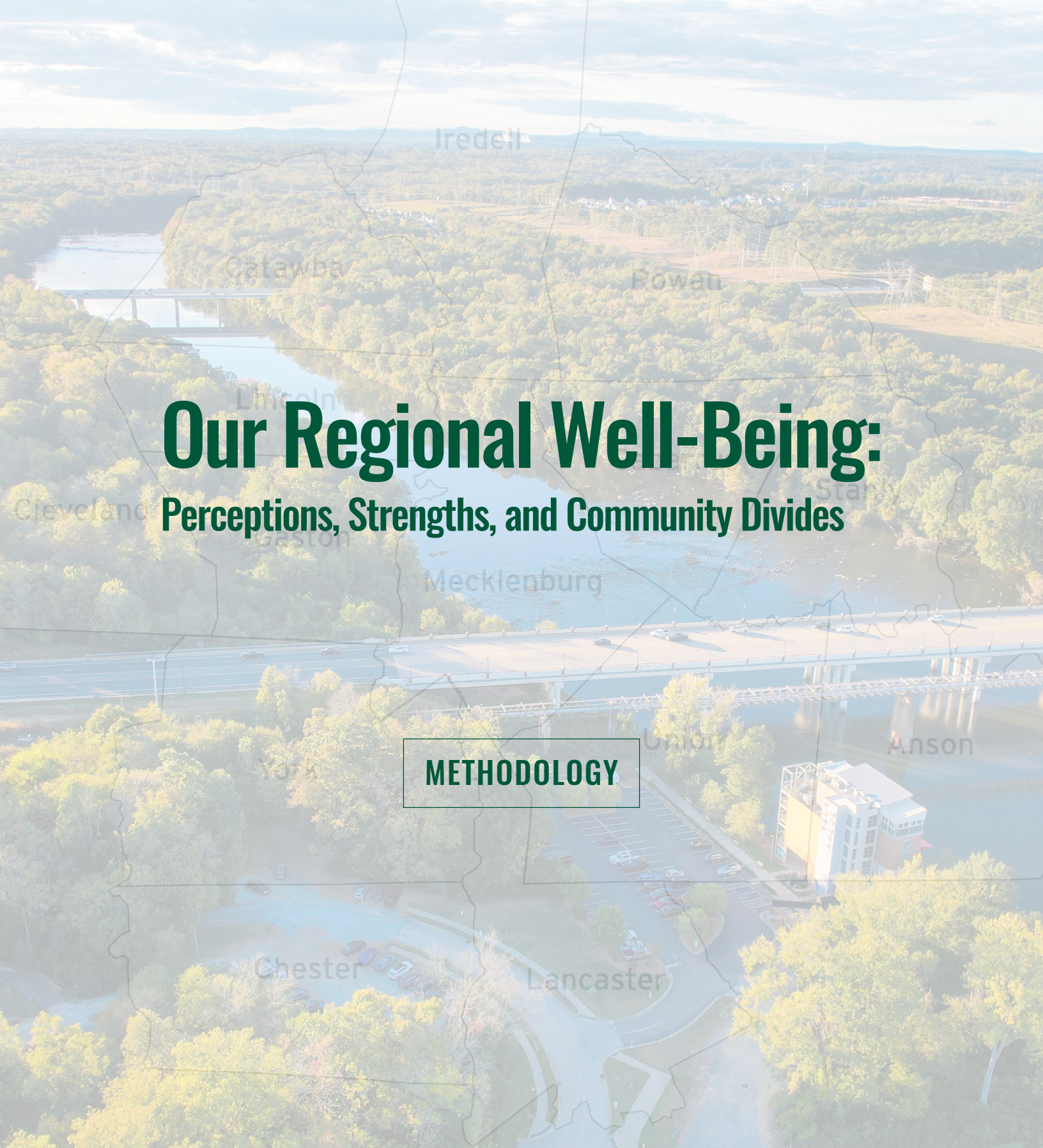




Our Regional Well-Being:

Perceptions, Strengths, and Community Divides

METHODOLOGY



UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

URBAN INSTITUTE
REGIONAL DATA TRUST

Appendices

Appendix A: Survey Weights

While we worked with ADRG to consider regional quotas when conducting the survey, several issues ensured the resulting sample would deviate from standard population metrics. First, quotas were focused on county of residence, race, Hispanic origin, gender, and age. In addition, because the quotas were somewhat relaxed, we generated them using straightforward marginal totals (i.e., how many Whites, how many Females, etc.) rather than a joint distribution.

Second, we purposely oversampled lower-density counties such as Anson and Chester. This was done to ensure that the final sample had an adequate number of respondents from every county in our region, with the knowing trade-off that it would consequently skew metrics for other variables.¹

In addition to these issues, we know that survey data often suffers from certain groups, e.g., younger individuals or renters, who are shown to be significantly less likely to participate than others (non-response bias), while the modality of the survey (phone or online) can also skew the final sample a survey provides.² This means that while the final quota counts remain within the acceptable deviation range (ADRG language), we still feel it is necessary to adjust each respondent's contribution to statistically match known Census population totals.

Variable Selection and Harmonization

We expanded the variables in our weighting strategy to include income and homeownership, in addition to those used to meet regional quotas: County, Race, Hispanic Ethnicity, Gender, and Age. We have included this expanded version of statistical weighting for the following reasons:

1. Studies have shown that economic stability is a primary driver of sentiment, particularly regarding quality of life and well-being. By weighting to Income targets, we ensure our findings on “well-being” aren’t artificially inflated by a sample that is wealthier than the actual population.³
2. Homeownership is a critical variable in local policy and quality-of-life questions, but renters are traditionally underrepresented in survey research; by including this variable, we ensure their unique perspectives on housing and community stability are given their proper proportional “voice”.⁴

¹ Richard Valliant et al., *Practical Tools for Designing and Weighting Survey Samples*, Statistics for Social and Behavioral Sciences (Springer International Publishing, 2018), <https://doi.org/10.1007/978-3-319-93632-1>.

² Richard Valliant et al., *Practical Tools for Designing and Weighting Survey Samples*, Statistics for Social and Behavioral Sciences (Springer International Publishing, 2018), <https://doi.org/10.1007/978-3-319-93632-1>.

³ Daniel Kahneman and Angus Deaton, “High Income Improves Evaluation of Life but Not Emotional Well-Being,” *Proceedings of the National Academy of Sciences* 107, no. 38 (2010): 16489–93, <https://doi.org/10.1073/pnas.1011492107>; Matthew A. Killingsworth et al., “Income and Emotional Well-Being: A Conflict Resolved,” *Proceedings of the National Academy of Sciences* 120, no. 10 (2023): e2208661120, <https://doi.org/10.1073/pnas.2208661120>.

⁴ Barry Steinberg et al., “A Nonresponse Bias Study of the Consumer Expenditure Survey for the Ten-Year Period 2010–2019,” Consumer Expenditure Surveys Program Report Series, U.S. Bureau of Labor Statistics, 2022, https://www.bls.gov/cex/research_papers/pdf/steinberg-et-al-a-nonresponse-bias-study-of-the-consumer-expenditure-survey-for-the-ten-year-period-2010-2019.pdf; Denise DiPasquale and Edward L. Glaeser, “Incentives and Social Capital: Are Homeowners Better Citizens?,” *Journal of Urban Economics* 45, no. 2 (1999): 354–84, <https://doi.org/10.1006/juec.1998.2098>.

We begin by aligning the survey data with Census targets, starting by collapsing the Race categories. As noted in Table A1, certain groups have extremely small counts and, consequently, can produce issues within the various weighting algorithms, creating unstable response propensity estimates and extreme weight variance.⁵

Responses of Online and Telephone Respondents, Unweighted Percentages (Column)

	Phone (%)	Web (%)	Total (%)
White	64.44	68.84	66.33
Black or African American	22.75	23.17	22.93
Asian	1.82	1.97	1.88
American Indian or Alaskan Native	1.56	2.25	1.86
Middle Eastern or North African	0.25	0.62	0.41
Native Hawaiian or other Pacific Islander	0.34	0.62	0.46
Other	5.37	1.91	3.89
Respondent hung up	0.17	0	0
DK/Refused	3.3	0.62	2.15

To avoid these issues and increase the likelihood that the weighting specifications converge, we merged the smaller categories of “AIAN” and “NAPI” into the “Other” category. We also manually reviewed the values and levels for each demographic variable in the survey data to ensure they matched the target list exactly, preventing errors during iterative processing.

Different Weighting Models

Traditional raking, also known as *Iterative Proportional Fitting (IPF)*, is the classic method to adjust survey samples to reflect population totals.⁶ The model works iteratively: it adjusts the weights for one demographic variable, checks how that shifts the others, and makes small course corrections. It cycles through every variable again and again until all demographic targets match their Census marginals.

Calibration differs from the traditional IPF specification in that it is a model-based technique that uses a distance function to solve for all demographic targets simultaneously and is designed to find the smallest possible adjustment to the raw weights that will satisfy all population constraints.⁷ To ensure methodological rigor, our process evaluated three distinct functional forms of calibration to determine which provided the most stable and representative results:

⁵ David Haziza and Jean-François Beaumont, “Construction of Weights in Surveys: A Review,” *Statistical Science* 32, no. 2 (2017), <https://doi.org/10.1214/16-STS608>; Michael P. Battaglia et al., “Practical Considerations in Raking Survey Data,” *Survey Practice* 2, no. 5 (2009): 1-10, <https://doi.org/10.29115/SP-2009-0019>.
⁶ W. Edwards Deming and Frederick F. Stephan, “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known,” *The Annals of Mathematical Statistics* 11, no. 4 (1940): 427-44, <https://doi.org/10.1214/aoms/1177731829>.
⁷ Jean-Claude Deville and Carl-Erik Särndal, “Calibration Estimators in Survey Sampling,” *Journal of the American Statistical Association* 87, no. 418 (1992): 376-82, <https://doi.org/10.1080/01621459.1992.10475217>.

- **Linear:** This method utilizes a squared-distance function to minimize the difference between raw and adjusted weights. It is the mathematical foundation for most large-scale government surveys. While highly efficient, it can produce negative weights if the sample requires extreme adjustments to meet population targets.
- **Raking:** This functional form replicates traditional raking within the calibration framework. It uses a multiplicative distance function to ensure that all final weights remain positive. It is often used as a more robust alternative to linear calibration when the survey sample is significantly skewed.
- **Logit (Bounded):** This is an advanced optimization technique that allows the researcher to set “hard bounds” (a floor and a ceiling) directly within the algorithm. It forces the computer to find a valid solution while strictly adhering to a predefined weight cap (e.g., 3.0), preventing any single respondent from becoming a statistical outlier during calculation.

While weighting corrects for demographic bias, it can also create “outliers”, i.e., respondents with very high weights. This raises the theoretical argument for weight truncation as extreme weights, whether caused by deliberate oversampling or nonresponse adjustments, can drastically inflate sampling variance and destabilize survey estimates.⁸ To evaluate the tradeoffs inherent within weight trimming, we run multiple iterations with different combinations of upper and lower bounds.⁹

Weighting Model Diagnostics

We compare the different weighting models using three standard metrics to gauge the most effective and appropriate methodology:

- **Design Effect:** This measure (DEFF) measures the “cost” of weighting. It quantifies how much the variance of our estimates increases as we move away from a simple unweighted sample using Kish’s formula.¹⁰
- **“Effective” Sample Size:** A calculation that estimates the accuracy and precision of our data by comparing it to a hypothetical perfect random sample. While theoretical, the metric remains critical for researchers as it indicates how much confidence we can place in our statistical tests after applying weighting.
- **Max Bias:** Another measure of accuracy that identifies the largest gaps between the weighted sample and the census targets for all weighted variables.

Table A2 evaluates the multiple weighting philosophies against three core criteria discussed above on our expanded weighting criteria (the ones the ADRG used in their quota counts in addition to our economic factors). We’ve also included the maximum weight (Max) to show the variation among the outliers for each method.

⁸ Frank Potter and Yuhong Zheng, “Methods and Issues in Trimming Extreme Weights in Sample Surveys,” paper presented at American Statistical Association, *American Statistical Association, Section on Survey Research Methods*, 2015.

⁹ Matthew DeBell and Jon A. Krosnick, *Computing Weights for American National Election Study Survey Data*, no. nes012427, ANES Technical Report Series (American National Election Studies, 2009), <http://www.electionstudies.org/>.

¹⁰ Leslie Kish, “Weighting for Unequal Pi,” *Journal of Official Statistics* 8 (1992): 183-200.

Survey Weighting Diagnostics and Efficiency Metrics (Expanded Models)

	Max Weight	Design Effect	Max Bias
Raking	5	1.414	0.027
Calibrate LM (Standard)	7.06	1.426	0
Calibrate LM (Truncated)	4.09	1.382	0
Calibrate Logit	2.99	1.4	0
Raking (T3)	3.07	1.312	0.023
Raking (T5)	5	1.413	0.027
Calibrate Ridge (0.1)	3.07	1.308	0.014
Calibrate Ridge (0.5)	5.01	1.405	0.002
Calibrate LM Bound (A)	2.99	1.347	0.005
Calibrate LM Bound (B)	4.06	1.365	0.003

After removing methods that failed to meet the accuracy threshold and those with extreme outliers in max weight and excessively high design effect, we selected the linear calibration with a weight truncation of 3 (L3) as the final weight for this project. By using the linear form and subsequently applying a conservative trim to an increasingly used methodology, we achieved a high-precision model that remains within 0.5% of all Census targets while maintaining a superior ESS. This dual-layer approach ensures that the final data is both demographically accurate and statistically powerful for regional analysis.

Appendix B: Small-Area Estimation

Understanding how residents across the fourteen-county region differ in their experiences and perceptions of quality of life and community well-being may not be the sole objective of the stratified analysis; it remains an integral component of the survey. Unfortunately, this ostensibly straightforward objective is more complicated than one might anticipate, compared with other demographic or descriptive variables. When comparing race, household income, gender, partisanship, employment status, and housing tenure, the sample's over 3,300 respondents (post-weighting) provide sufficient statistical power for reliable subgroup analyses.

This means results showing Black residents rate regional life satisfaction differently from white residents, or that renters express distinct opinions on housing affordability compared to homeowners, have sufficient numbers across groups to report descriptive cross-tabulations and weighted means directly. From a more technical perspective, larger sample sizes yield acceptable margins of error, allowing us to be confident these differences exist. This makes the results easy to interpret and provides a solid foundation for policy discussions.

County (of residence) presents a fundamentally different challenge. While the survey strove to achieve representation across all fourteen counties, the practical reality of a fixed sample size distributed across a region of this scale means that individual county samples, and therefore margins of error, vary substantially. A fact that can be seen in Table B1 below:

Total Number of Responses by Modality and Margin of Error

County	Phone	Web	Unwgt. N	Population	MOE (%)
Anson	25	22	47	22,432	14.28
Cabarrus	171	115	286	244,925	5.79
Catawba	104	110	214	167,054	6.69
Chester	23	66	89	32,323	10.37
Cleveland	85	99	184	102,194	7.22
Gaston	179	183	362	242,010	5.15
Iredell	116	93	209	206,361	6.78
Lancaster	73	52	125	111,652	8.76
Lincoln	79	63	142	97,611	8.22
Mecklenburg	957	510	1,467	1,206,285	2.56
Rowan	93	90	183	153,384	7.24
Stanly	58	46	104	67,326	9.6
Union	165	90	255	263,386	6.13
York	172	161	333	303,001	5.37
TOTAL	2,300	1,700	4,000	3,219,944	1.55

Note. MOE represents Margin of Error at the 95% confidence level.

Mecklenburg County, unsurprisingly, has the greatest number of respondents at just over 1300, a sample size that yields a margin of error of just over 2%. At the other end of the distribution, Anson County has just under 50 participants. While a higher number than would be “expected” if we strictly adhered to the region’s representative proportions, it still means the county has a margin of error of over 14%.

Why does this matter? For example, let’s say we were examining a question about overall life satisfaction, and a county estimate found that 60% of residents reported a high level of life satisfaction. A margin of error of 8 percentage points, i.e., roughly the average we theoretically expect from our sample, implies that this estimate could plausibly range from 52% to 68% in the actual population. This variability complicates our ability to distinguish between counties, and presenting raw county-level percentages as comparable to the regional subgroup figures would be methodologically misleading and could lead to actions based on false premises.

This uncertainty around certain geographies does not imply that the data is flawed or completely unreliable. Instead, it means direct estimation is problematic for small county samples, which is why the modeling approach described below aims to address these concerns.

Calculating County-level Estimates

Standard statistical models assume survey respondents are independent draws from a single population. In reality, respondents are nested within counties, sharing a local environmental context—such as housing markets, labor conditions, and public services—that individual demographics alone cannot explain. To generate defensible county-level estimates, we employ a Bayesian Hierarchical Model that explicitly captures this two-level data structure. At the individual level, the model accounts for key respondent demographics, such as race, income, employment status, and housing tenure. At the county level, it captures residual geographic variation using county-level Census covariates, including age structure, median income, racial composition, and homeownership rates.

The primary statistical advantage of this hierarchical framework lies in its use of partial pooling, which allows the model to learn the relationship between county demographics and survey outcomes across the full regional sample, using that baseline to stabilize estimates in counties with sparse data.¹¹ The degree to which a county’s estimate is pulled toward the regional trend—a phenomenon known as shrinkage—is calibrated dynamically by local sample size. Counties with large sample sizes, such as Mecklenburg, possess sufficient local data and remain largely unadjusted. In contrast, counties with small sample sizes, such as Anson, experience greater shrinkage toward the regional mean to prevent extreme sampling noise from skewing the final metrics.

While frequentist multilevel models also perform partial pooling, their reliance on large-sample asymptotic approximations creates significant limitations when working with a small number of geographic units.¹² A Bayesian specification overcomes this limitation by treating county-level random effects as parameters with their own prior distributions, estimating full posterior distributions for all parameters simultaneously so that county-level uncertainty is explicitly quantified rather than approximated. Additionally, the inclusion of weakly informative prior distributions on county-level

¹¹ Andrew Gelman and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 23rd printing, Analytical Methods for Social Research (Cambridge Univ. Press, 2021).

¹² Daniel Stegmueller, “How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches,” *American Journal of Political Science* 57, no. 3 (2013): 748–61, <https://doi.org/10.1111/ajps.12001>.

covariates and variance terms acts as a regularizing mechanism against overfitting.¹³ This ensures that the model's predictions remain appropriately conservative, responding only to genuine, data-supported geographic differences while safeguarding against sampling noise.

In summary, this two-level Bayesian framework provides the analytically rigorous foundation necessary for reliable small-area estimation.¹⁴ By capturing the nested nature of survey data, leveraging partial pooling to balance local signal with regional strength, and using Bayesian priors to fully propagate uncertainty across 14 geographic units, the model maximizes statistical efficiency without overstating precision.

The Difference in Bayesian Estimates

Unlike a traditional regression, Bayesian models treat each parameter as having a probability distribution, or a range of plausible values weighted by their consistency with the observed data and prior knowledge about reasonable parameter ranges. This means the output of a Bayesian model is not a point estimate but a full posterior distribution over all model parameters.

The computational method employed to estimate these posterior distributions is Markov Chain Monte Carlo sampling, implemented via the Stan probabilistic programming platform, which underpins both the *rstanarm* and *brms* packages used in this analysis.¹⁵ The underlying concept is intuitive: just as a survey estimates population tendencies by drawing a representative sample of individuals, MCMC estimates a model's posterior distribution by generating a large representative sample of plausible parameter values. Stan specifically employs Hamiltonian Monte Carlo, a particular form of MCMC that utilizes the derivatives of the posterior density to generate transitions that explore the parameter space more effectively than older approaches. This reduces the number of iterations required to achieve reliable estimates.

The practical difference between the county-level results reported here and those produced by a simpler approach lies in the post-estimation step. Instead of calculating a single county-level predicted probability and attaching a margin of error based on a normal approximation, the process generates 2,000 plausible county-level estimates for each county and outcome category. These estimates are then aggregated at the county level using survey-design weights, increasing the likelihood that the final predictions reflect each county's actual population composition rather than the respondent sample's composition within that county.

Another difference between the Bayesian approach we utilize and a more traditional regression lies in the uncertainty of these estimates. A confidence interval, as one might see in a fixed-effects regression, describes the long-run behavior of the estimator across hypothetical repeated samples. On the flip side, our approach provides a probability that a given estimate contains the true estimate based on the full posterior distribution of plausible values for each county.¹⁶ While it is easy to conflate these

¹³ Andrew Gelman, "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)," *Bayesian Analysis* 1, no. 3 (2006), <https://doi.org/10.1214/06-BA117A>.

¹⁴ J. N. K. Rao and Isabel Molina, *Small Area Estimation*, 1st ed. (Wiley, 2015), <https://doi.org/10.1002/9781118735855>.

¹⁵ Bob Carpenter et al., "Stan: A Probabilistic Programming Language," *Journal of Statistical Software* 76, no. 1 (2017), <https://doi.org/10.18637/jss.v076.i01>; Paul-Christian Bürkner, "Brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software* 80, no. 1 (2017), <https://doi.org/10.18637/jss.v080.i01>.

¹⁶ Andrew Gelman and Sander Greenland, "Are Confidence Intervals Better Termed 'Uncertainty Intervals'?", *BMJ*, September 10, 2019, 15381, <https://doi.org/10.1136/bmj.15381>; Richard D. Morey et al., "The Fallacy of Placing Confidence in Confidence Intervals," *Psychonomic Bulletin & Review* 23, no. 1 (2016): 103-23, <https://doi.org/10.3758/s13423-015-0947-8>; John K. Kruschke and Torrin M. Liddell, "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective," *Psychonomic Bulletin & Review* 25, no. 1 (2018): 178-206, <https://doi.org/10.3758/s13423-016-1221-4>.

two concepts, the latter allows us to analyze model results more intuitively and better aligns with practitioners' and non-specialists' ability to interpret these findings.

For binary outcome variables, we used a multilevel logistic regression implemented in the `rstanarm` package in R, which models the probability of a respondent falling into the outcome category of interest as a logistic function of individual-level demographic predictors and a county-level random intercept. For ordered categorical outcomes such as the three-category life satisfaction variable, the model used was a multilevel cumulative logistic regression implemented in the `brms` package in R, which extends the binary framework to outcomes with more than two ordered categories by estimating threshold parameters that divide the underlying continuum of the outcome into the observed categorical responses.

In both specifications, individual-level predictors account for compositional differences across counties; county-level Census covariates anchor the geographic variation in outcomes to observable structural features of each county; and post-stratification weights derived from the survey's quota sampling design are incorporated directly into the estimation process to ensure that the results reflect the population. The county-level random intercept in both models is the mechanism through which partial pooling operates, and its posterior distribution, when passed through the post-estimation prediction step, produces the county-level estimates and accompanying credible intervals reported throughout this analysis.



ui.charlotte.edu