Psychological Methods

The Big, the Bad, and the Ugly: Geographic Estimation With Flawed Psychological Data

Joe Hoover and Morteza Dehghani

Online First Publication, October 24, 2019. http://dx.doi.org/10.1037/met0000240

CITATION

Hoover, J., & Dehghani, M. (2019, October 24). The Big, the Bad, and the Ugly: Geographic Estimation With Flawed Psychological Data. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000240

American Psychological Association

Psychological Methods

© 2019 American Psychological Association
ISSN: 1082-989X

http://dx.doi.org/10.1037/met0000240

The Big, the Bad, and the Ugly: Geographic Estimation With Flawed Psychological Data

Joe Hoover and Morteza Dehghani
University of Southern California

Abstract

The geographic distribution of psychological constructs has long been an area of focus for psychological researchers. Recently, however, there has been increased interest in investigations of the so-called subnational distribution of psychological variables, which focus on localized groupings of individuals within spatial units, such as counties or states. By estimating the subnational distribution of a given outcome (e.g., estimating its state- or county-level means), researchers have been able to address questions about the spatial variation of a variety of psychological constructs and investigate the regional association between psychological phenomena and real-world outcomes, such as health outcomes, prosocial behavior, and racial inequity. Unfortunately, however, there are many challenges to estimating a construct's subnational distribution, such as those raised by response biases and subnational sparsity. To help psychological researchers address these issues, we provide a comprehensive discussion of subnational estimation and introduce multilevel regression and poststratification (MrP), a method that is widely considered to be the gold standard for subnational estimation with random samples. As psychologists often do not have access to large, national random samples, we also report 3 studies evaluating MrP's performance under simulated and real-world conditions of sample biases. Ultimately, we find that MrP is likely to outperform the subnational estimation methods that psychological researchers currently use. Based on this, we suggest that psychologists interested in understanding how psychological phenomena vary below the nation level use MrP to conduct these investigations. To help facilitate this, we have made all code and data used for the reported studies publicly available.

Translational Abstract

The geographic distribution of psychological constructs has attracted increasing interest among psychological researchers. Relying on these and other data, psychologists have been able to not only address novel questions about the spatial variation of psychological constructs but also investigate the regional association between psychological phenomena and real-world outcomes, such as outcomes associated with health, prosocial behavior, and racial inequity. Unfortunately, there are many challenges to estimating a construct's regional distribution—so-called subnational estimation—and these challenges are exacerbated by issues of nonrepresentativeness and geographic sparsity. In this work, we provide a comprehensive discussion of major obstacles for subnational estimation and introduce readers to state-of-the-art approaches that rely on multilevel regression and poststratification (MrP) to deal with these obstacles. We also present a novel evaluation of MrP and extensions of MrP under conditions of sample size and response bias via simulations (Study 1) and application to real-world data obtained from a large convenience sample (Study 2). Finally, we investigate how estimated associations between an estimated county-level outcome—racial bias—and a secondary outcome—Barack Obama's 2008 county-level Presidential vote share—vary depending on the method used for subnational estimation (Study 3). In addition to offering a comprehensive introduction to cutting-edge methods for subnational estimation, this work provides strong evidence for the necessity of incorporating more sophisticated techniques for subnational estimation into studies of the geographic distribution of psychological phenomena.

Keywords: subnational estimation, geographic psychology, multilevel regression and poststratification, response bias, project implicit

Supplemental materials: http://dx.doi.org/10.1037/met0000240.supp

Joe Hoover, Department of Psychology, University of Southern California; Morteza Dehghani, Department of Psychology and Department of Computer Science, University of Southern California.

This research was sponsored in part by the Army Research Lab. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. An early version of this work, including Study 1, was posted on PsyArxiv (Hoover & Dehghani, 2018). Studies 2 and 3 have not been publicly disseminated. Many thanks to Lucas Leeman and Andrew Gelman for their thoughtful input on some of the conceptual and computational challenges faced during this work. We also thank Project Implicit and Pew Research Center for sharing their data. All data, code, and estimates are available at https://osf.io/8javp/.

Correspondence concerning this article should be addressed to Joe Hoover, who is now at Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208. E-mail: jehoover@usc.edu

The subnational distributions of psychological constructs are attracting increasing interest in the psychological literature where, for example, outcomes such as well-being or racial bias are being studied within smaller units such as states or counties. Such research relies on what is referred to as subnational estimation, which involves estimating the population distribution of a construct across a set of subnational units using samples of data drawn from those units. While subnational estimation is a relatively new approach to psychological research, it is relevant to any psychologist who is interested in working with estimates at smaller, more localized levels like the state-, county-, or city-level, as opposed to larger national or international levels.

By studying a construct's subnational variation, researchers can learn about its stability, relationships with covariates, and responses to naturally occurring perturbations. For example, a growing body of literature has identified systematic subnational geographic covariance among personality traits (Allik et al., 2009; Rentfrow, Gosling, & Potter, 2008; Rentfrow, Jokela, & Lamb, 2015) and between personality and other outcomes, such as life-satisfaction (Jokela, Bleidorn, Lamb, Gosling, & Rentfrow, 2015), liberalism (Rentfrow et al., 2015; Rentfrow, Gosling, et al., 2013), cancer (McCann, 2017b), volunteering (McCann, 2017a), work satisfaction (McCann, 2018), and economic resilience (Obschonka et al., 2016). Recent research has also provided evidence that the congruence between a person's personality and the dominant personality traits in their region is associated with their subjective well-being (Götz, Ebert, & Rentfrow, 2018). In other work, researchers have begun exploring the county-level distribution of moral values in the United States (Hoover, Zhao, & Dehghani, 2018).

Another burgeoning line of work has focused on the subnational distribution of racial bias and its association with indicators of racial inequity. Studies in this area have identified links between county-level implicit bias against Blacks and the Black-White infant mortality gap (Orchard & Price, 2017), Black's death-rates (Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016a, 2016b), exposure to racial out-groups (Rae, Newheiser, & Olson, 2015), disproportionate use of lethal force against Blacks in policing (Hehman, Flake, & Calanchini, 2017), and racial disparities in school-based disciplinary actions (Riddle & Sinclair, 2019). While researchers have long speculated that such associations exist, they have remained difficult to assess quantitatively. However, by focusing on subnational variation in target outcomes, researchers have been able gain novel insight into the relationships between psychological phenomena and real-world outcomes.

Unfortunately, some of the approaches to subnational estimation that are most widely employed in the psychological literature do not adequately address the methodological challenges of subnational estimation. At worst, these approaches can yield completely invalid estimates and inferences. Specifically, the methods most widely used either inadequately address or wholly neglect issues of subnational sparsity and representativeness. A sample exhibits subnational sparsity when, for some subnational units, data is missing or *N*s are very small. Similarly, a sample exhibits subnational nonrepresentativeness when the data representing some subnational units is not representative. If these issues are not addressed, subnational estimates may be unreliable, biased, and (or) completely invalid.

In this work, we review these issues and discuss methods that have been developed to address them. While some of these methods, such as poststratification (Gelman & Little, 1997; Little, 1993; Lohr, 2009) have been used in the psychological literature (Leemann & Wasserfallen, 2017; Leitner et al., 2016a; Obschonka et al., 2016; Orchard & Price, 2017), others, such as raking (Deville, Särndal, & Sautory, 1993; Kalton & Flores-Cervantes, 2003), multilevel regression and poststratification (MrP; Gelman & Little, 1997; Park, Gelman, & Bafumi, 2004), and multilevel regression and synthetic poststratification (MrsP; Leemann & Wasserfallen, 2017) are not as well-known to psychological researchers. Each of these methods constitute an approach to survey adjustment that can be used to address subnational sparsity and nonrepresentativeness. We provide an overview of these approaches and discuss their strengths and weaknesses.

More specifically, however, we propose that MrP and its more recent variants will be particularly useful for psychologists interested in subnational investigations of psychological phenomena. MrP offers a model-based approach to obtaining subnational estimates for a given outcome, such as state-level estimates of public opinion (Krimmel, Lax, & Phillips, 2016) and voter behavior (Gelman, 2014), county-level estimates of racial bias (Riddle & Sinclair, 2019), or city-level estimates of health outcomes (Y. Wang et al., 2018). In contrast to methods like poststratification and raking (see below for discussion of these methods), MrP relies on a hierarchical response model which helps improve estimation accuracy via partial-pooling or smoothing (Park et al., 2004). Accordingly, a researcher interested in studying racial bias, for example, could apply MrP to data from Project Implicit in order to derive estimates of state- or county-level racial bias. Through the application of MrP, these estimates would be stabilized due to partial-pooling as well as adjusted for response biases via the application of poststratification. MrP has become increasingly popular and is now considered the gold standard for estimating subnational political preferences (Caughey & Warshaw, 2019; Leemann & Wasserfallen, 2017; Selb & Munzert, 2011). Recent work has also demonstrated that MrP can even generate surprisingly accurate subnational estimates from nonrandom and nonrepresentative data (W. Wang, Rothschild, Goel, & Gelman, 2015). Further, it has been shown to outperform the methods more commonly used in psychological research, such as and disaggregation (Erikson, Wright, & McIver, 1993)—merely calculating region-specific sample means—and poststratification (Park et al., 2004).

However, previous comparative evaluations of MrP have found that it offers diminishing returns as sample sizes increase (Buttice & Highton, 2013; Hanretty, Lauderdale, & Vivyan, 2016; Lax & Phillips, 2009), suggesting that when enough data is available, more simple approaches like disaggregation may perform comparably. These evaluations, however, were conducted with randomly sampled, nationally representative data and thus cannot necessarily be generalized to the kinds of large, but also nonrandom and biased data (e.g., data collected via Project Implicit, MyPersonality, or YourMorals.org) that psychological researchers often work with today.

Accordingly, in addition to providing a detailed introduction to MrP and some of its recent modifications, we also report results from three new studies investigating its comparative performance under conditions similar to those faced by psychological researchers. Specifically, these studies address the following questions:

1. Under simulated conditions of sampling bias caused by unrepresentative sampling, how does MrP perform (Study 1)?

2. Given a large, unrepresentative, nonrandom sample, how does MrP perform compared to other methods of subnational estimation (Study 2)?

3. Given a large, unrepresentative, nonrandom sample, do downstream inferences about the relationship between subnational estimates and a secondary construct vary depending on the method used to obtain subnational estimates (Study 3)?

In Study 1, we address the first question via a large-scale Monte-Carlo simulation that we use to estimate the accuracy and bias of subnational MrP estimates under varying levels of nonrepresentativeness and sample size. While simulation necessarily requires making simplifying assumptions about data generating processes, this study provides new information about MrP's performance under conditions of varying bias and sample size.

Next, in order to better understand how MrP performs under these conditions when applied to real data, we rely on large-scale data obtained from Project Implicit (Xu, Lofaro, Nosek, & Greenwald, 2013) to generate county-level estimates of the rate of Catholic adherence using MrP as well as a range of other methods. While the county-level rate of Catholic adherence may not be of particular psychological interest, focusing on this variable allows us to directly evaluate estimation accuracy and bias, as a reasonable approximation of "ground-truth" (the true rate of Catholic adherence) is available via the 2010 U.S. Religious Census (Grammich, 2012).

Finally, in Study 3, we investigate how inferences about the relationship between Barack Obama's 2008 General Election county-level vote share and county-level White racial bias against Blacks vary depending on the method used to estimate county-level racial bias. Previous research has found a negative association between intent to vote for Obama and both explicit and implicit racial bias (Greenwald, Smith, Sriram, Bar-Anan, & Nosek, 2009). Given this, a question of presumable interest might be whether this association exists at the county-level. Importantly, however, our goal in this study is not to provide evidence for or against such an association, but rather to investigate how inferences vary depending on the method used to obtain estimates of county-level racial bias. That is, in this study, we sought to determine whether the method of estimation—in this particular context—had substantive implications for the kind of downstream analyses psychologists might be interested in conducting.

Overall, our aim in this work is to introduce psychologists to subnational estimation, highlight its challenges, and provide actionable information regarding how these challenges can and should be addressed. In our empirical work, we provide evidence via simulation and analysis of real data that, under conditions of subnational sparsity and (or) nonrepresentativeness, MrP can improve the accuracy of subnational estimates, regardless of sample size. Further, we also demonstrate that downstream inferences about the relationship between county-level estimates and a secondary county-level outcome can vary substantively depending on the method of estimation. Finally, we also provide all of the code and data used for these studies at https://osf.io/8javp/ that readers can more easily apply these methods or use our estimates in their own research.

Subnational Estimation

Overview

Subnational estimation of a variable involves obtaining estimates of population parameters, such as means or medians, for subnational areas that fall below the nation level, such as states, provinces, counties, or districts. For example, the problem of estimating state-level means for extroversion, explicit racial bias, or well-being are all problems of subnational estimation. Subnational estimation is neither inherently difficult nor complicated. As is the case with many problems of estimation, access to sufficient data renders the problem trivial. For instance, estimating U.S. state-level explicit racial bias would be simple if one had a sufficiently large random sample of racial bias measurements drawn from each state. With such data, subnational estimates of explicit racial bias could simply be obtained by calculating the distribution of means for each state.

Unfortunately, researchers rarely have access to such data due to the cost and difficulty of collecting sufficiently large random samples from multiple subnational areas. Accordingly, various methods are used in order to facilitate the derivation of subnational estimates from less-than-ideal data. In the psychological literature, the methods most frequently used for subnational estimation are disaggregation (Erikson et al., 1993) and poststratification (Gelman & Little, 1997; Lohr, 2009). Below, we review these approaches to subnational estimation and discuss two other approaches that are less well-known to psychological researchers, raking (Deville et al., 1993; Kalton & Flores-Cervantes, 2003) and MrP (Park et al., 2004).

Disaggregation

As noted above, subnational estimation of a variable, such as explicit racial bias, is trivial when a sufficiently large random sample of the variable is available for each subnational unit. With such data, population estimates of the target variable's subnational means can simply be estimated via the subnational sample means, a procedure often referred to as "disaggregation." Further, while such data is rarely directly available, it can, in some cases, be approximated by combining data from multiple nationally representative surveys into a single data-set and then segmenting or "disaggregating" the data into the desired level of analysis (Erikson et al., 1993). Population estimates of the target variable's subnational means can then be simply estimated via the disaggregated sample means.

This approach hinges on the premise that combining multiple random and nationally representative samples will eventually produce a supersample that is sufficiently representative at the targeted subnational level. However, while it is asymptotically valid, in many instances it is not a viable option. While it may be possible to construct a sufficient supersample for a small set of constructs for which data is frequently collected, this is often not the case for constructs excluded from that set, such as personality inventories and measures of explicit and implicit attitudes. Further, depending on the level of geographic analysis, it may not be possible to assemble a supersample for even the most widely collected variables. Consider, for example, that moving from a U.S. state-level analysis to a county-level analysis increases the number of spatial units by a factor of approximately 60; thus, data deemed sufficient for a disaggregated state-level analysis would need to be expanded by roughly the same factor to provide comparable coverage for a county-level analysis.

An even more pressing problem for disaggregation is its inability to address response biases and failures in randomization. If certain segments of the target population over- or underrespond, disaggregated estimates will be biased (Holt & Smith, 1979; Lax & Phillips, 2009; Little, 1993) even if they are derived from an infinite sample (Pew Research Center, 2018).

Poststratification

To address issues of response bias or nonrepresentativeness, researchers employ a range of techniques that aim to adjust a sample so that it reflects known population characteristics. For instance, the proportion of people in a sample who fall in certain age bracket, report a given sex, or perhaps are characterized by some combination of these variables may not match the population proportions for these demographic characteristics. One way to account for this mismatch between sample demographic proportions and population demographic proportions is to calculate sample weights that can be used to weight respondents so that the weighted sample demographic proportions match the population demographic proportions.

One approach to calculating sample weights is "poststratification" (Gelman & Little, 1997; Lohr, 2009), which, in the psychological literature, has most frequently been used to adjust for age and gender (e.g., see Leemann & Wasserfallen, 2017; Leitner et al., 2016a; Obschonka et al., 2016; Orchard & Price, 2017). Poststratification is generally implemented as follows. The first step is to select a set of demographic variables, often referred to as auxiliary variables, for which adjustments will be made. Generally, auxiliary variables should be selected depending on whether the target variable (i.e., the variable for which subnational estimation is conducted) varies over their levels. For instance, age and sex might be selected as auxiliary variables for the subnational estimation of well-being. Conceptually, these auxiliary variables are used to "poststratify" sample respondents into a set of demographic categories or cross-classifications. That is, the auxiliary variables age and sex can be used to poststratify sample respondents into discrete demographic bins that each represent a unique combination of age and sex. By convention, we refer to these as demographic cross-classifications or "poststrata."

Finally, the population estimate of a target variable, such as well-being, within a given subnational area can be estimated as the weighted mean of the poststrata sample means $\pi_{u[l],j}$, where the weights reflect the demographic population proportions corresponding to the poststrata within the subnational unit. Here, $\pi_{u[l],j}$ refers to the poststrata sample mean π for poststratum j located within subnational area l of upper-level area u. For instance, under this notation convention, j might refer to the poststratum combination of age and gender and $u[l]$ might index counties nested in states.

Note that under this approach, the poststratified mean for a given subnational unit is a function of the poststrata sample means *within that subnational unit*. That is, the poststratified subnational estimate for a given subnational area is based exclusively on the data sampled from that subnational area. Accordingly, this approach minimally requires $n_{u[l],j} > 0$, where $n_{u[l],j}$ represents the sample size n for poststratum j located in subnational area l in upper-level area u. Accordingly, $n_{u[l],j} > 0$ simply states that there must be at least one sample respondent for each sample poststrata within each subnational area. However, it is generally preferable to have larger sample sizes, such as $n_{u[l],j} \geq 50$, in order to minimize the effects of sampling error.

To summarize, subnational estimates of a target variable $Y_{u[l]}$ can be obtained via poststratification by selecting a set of auxiliary variables; calculating the means $\pi_{u[l],j}$ of the poststrata j within each subnational area $u[l]$; and finally calculating the weighted mean of $\pi_{u[l],j}$, where weights $p_{u[l],j}$ represent the population proportion of poststratum j in subnational area $u[l]$:

$$Y_{u[l]} = \sum_{j=1}^{j=J} p_{u[l],j}\pi_{u[l],j}, \tag{1}$$

for each poststratum $j \in j = 1, \ldots, j = J$.

Importantly, poststratification adjustment procedures vary substantially in complexity. For instance, it is often desirable to select multiple auxiliary variables, with age, gender, race, and education being the most used set. However, adding auxiliary variables can dramatically increase the number of demographic cross-classifications, particularly considering that they are crossed with subnational units. For example, poststratifying on three-level age and education and two-level gender would produce 18 demographic cross-classifications, which themselves are nested in subnational units. For a state- or county-level analysis, this approach would yield approximately 18 × 50 = 900 or 18 × 3,007 = 54,126 distinct participant cross-classifications and adhering to a ≥ 50 rule would require sample sizes of approximately 45,000 or 2.7 million, respectively.

To mitigate such exploding sample size requirements, poststratification can be reformulated so that estimates of the poststratum means are *pooled* across subnational units. That is, rather than estimating the mean for each poststratum *within* each subnational unit—the no pooling approach—the poststratum means can be estimated across all subnational units (Gelman & Little, 1997). However, while unpooled poststratification risks high-standard errors and inflated between-unit variation, pooled stratification risks homogeneity and suppressed between-unit variation.

Raking

In addition to issues of sparsity within poststrata cells, another challenge that often complicates poststratification is the difficulty of obtaining population estimates for the cross-classification of the auxiliary variables. In response to this issue, methods such as raking (Deville et al., 1993; Kalton & Flores-Cervantes, 2003) are often substituted for poststratification. Whereas poststratification operates on the joint distribution of the auxiliary variables, raking operates on their marginal distributions, such that sample weights are derived by iteratively adjusting the marginal distributions of the auxiliary *sam-*

ple variables to match the population marginal distributions. For example, raking over age and education at the state-level would involve weighting respondents within each state so that the weighted distribution of their ages matches the known marginal state-level distribution of age. The same procedure would then be applied to education and if this reweighting interferes with the age alignment, it would be reapplied to age. This iterative reweighting process would be repeated until the marginal distributions of the auxiliary variables match their known marginal distributions within some a priori range of error.

Raking can considerably expand the pool of viable auxiliary variables, compared with poststratification. However, the fine-grained information encoded by the full join-distribution is lost and this can negatively impact estimates. To mitigate this loss of information, raking can also be conducted over some subset of the cross-classifications of the auxiliary variables. However, as with poststratification, this introduces additional data requirements: the distribution for the chosen cross-classifications must be known and sufficient sample data must be available for each cross-classification category, otherwise estimates may be wildly inaccurate (Gelman, 2007).

Multilevel Regression and Poststratification

While poststratification and raking remain viable approaches to survey weighting, a more recently developed method, multilevel regression and poststratification (MrP; Gelman & Little, 1997; Park et al., 2004), has become increasingly popular and is now considered the gold standard for estimating subnational political preferences (Leemann & Wasserfallen, 2017; Selb & Munzert, 2011). For example, it has been used in subnational studies on legislative responsiveness to constituent opinion (Kastellec, Lax, Malecki, & Phillips, 2015; Krimmel et al., 2016), regional variations in environmental opinions (Fowler, 2016; Howe, Mildenberger, Marlon, & Leiserowitz, 2015), and the relationship between income and political preferences (Gelman, 2014). In contrast to conventional poststratification, in which sample weights are applied directly to the sample means for each poststratum, MrP involves applying sample weights to estimates of poststratum means derived from a hierarchical model fit to individual-level data (Lax & Phillips, 2009; Park et al., 2004). Subnational means are then estimated as the population-weighted mean of these predicted poststratum means.

The primary advantage conferred by MrP arises from how poststratum means for a given outcome are estimated. First, individual-level responses are modeled as a hierarchical, or multilevel, function of demographic auxiliary variables, subnational geographic indicators, and contextual factors (Lax & Phillips, 2009; Park et al., 2004). For example, an individual i's response y_i on a measure of explicit racial bias, could be estimated as a function of their age, level of education, and subnational unit (SNU; e.g., county), and contextual factors $\boldsymbol{X}$ (e.g., associated) with their subnational unit (e.g., county-level Democratic vote proportion, median income, proportion of population living in poverty, etc.):

$$\begin{aligned} &y_i = \beta_0 + \alpha_{a[i]} + \alpha_{e[i]} + \alpha_{c[i]} \\ &\alpha_a \sim \mathcal{N}(0, \sigma_a^2), \text{for } a = 1, \ldots, A \\ &\alpha_e \sim \mathcal{N}(0, \sigma_e^2), \text{for } e = 1, \ldots, E \\ &\alpha_{SNU} \sim \mathcal{N}(\alpha_{ULU[SNU]} + \beta \boldsymbol{X_c}, \sigma_{SNU}^2), \text{for } SNU = 1, \ldots, SNU \\ &\alpha_{ULU} \sim \mathcal{N}(0, \sigma_{ULU}^2), \text{for } ULU = 1, \ldots, ULU. \end{aligned} \tag{2}$$

In the above model, the effects of the auxiliary variables age, $\alpha_{a[i]}$, and education, $\alpha_{e[i]}$, are modeled as random effects (Gelman & Hill, 2006; Raudenbush & Bryk, 2002; Steenbergen & Jones, 2002), such that they are assumed to be generated from a normal distribution with $\mu = 0$ and variance σ^2. Further, the effect of subnational unit, α_{SNU}, is modeled as a normally distributed random effect, conditional on SNU-level contextual factors and the upper-level unit that it is nested in. Finally, the effect of upper-level unit, α_{ULU}, is modeled as an unconditional random effect.

After modeling individual-level measurements of the target construct, the next step in MrP is to use the trained model to generate predictions π_{aeSNU} for each cross-classification of the auxiliary variables age and education, conditional on subnational location and contextual factors. Thus, in the case of this example, for *each* subnational unit, a prediction is made for each combination of the levels of age and education. That is, the model is used to estimate the average response for a person of age a with level of education e who lives in subnational unit *SNU*, for all combinations of $a = 1, \ldots, A$, $e = 1, \ldots, E$, and $c = 1, \ldots, C$.

Finally, poststratification proceeds similar to as discussed above: subnational means $\bar{Y}_c^{MrP}$ are estimated by summing over the products of the predicted means and population proportions for each cross-classification of age and education:

$$\bar{Y}_c^{MrP} = \sum_a^A \sum_e^E P_{aeSNU} \pi_{aeSNU} \tag{3}$$

where, P_{aeSNU} is the proportion of people in *subnational unit* $=$ *SNU* of *age* $= a$ with *education* $= e$ and π_{aeSNU} is the predicted outcome for the same cross-classified group.

All together, this approach helps address issues driven by data sparsity with regard to both subnational units and poststrata cells. Even if there are no observations for a particular county, its effects can still be estimated as a linear combination of demographic effects, its contextual variable scores, and the effects for the other counties in its state. Further, if a given poststratum cell contains few observations (e.g., if the data happens to contain few measurements for women who are over 65 years of age and have not attended college), hierarchical smoothing helps stabilize the estimates for this poststratum cell. This robustness to sparsity makes it possible to include more relevant auxiliary variables, which can further improve estimates (Gelman, 2007).

Another notable benefit of MrP is that it is easy to expand the predictive model to exploit known or expected effects. For example, interactions between auxiliary variables can also be estimated and/or the effects of auxiliary variables can be permitted to vary across spatial units, such as regions, which would constitute a so-called random slopes model (Raudenbush & Bryk, 2002).

However, while MrP offers notable advantages, evaluations of its performance highlight that designing a MrP model requires careful thought (Buttice & Highton, 2013; Hanretty et al., 2016; Lax & Phillips, 2009), as its capacity to capture regional variation in an outcome depends on the variance that the predictive model explains. Both Buttice and Highton (2013) and Hanretty, Lauderdale, and Vivyan (2016) emphasize that well-chosen contextual variables are essential for MrP estimation, finding that MrP models with poor or unrelated contextual variables may not perform well. Further, it should be noted that the variables traditionally used in MrP models—such as presidential vote share—may not offer be as predictive when modeling psychological outcomes. Accordingly, it is important that contextual variables are not chosen based on convention but rather for their association with the target variable.

Similarly, poststratification variables should not be chosen arbitrarily, but rather with attention to the goal of explaining as much variance in the outcome as possible. That said, it is worth noting that we are not aware of any evidence suggesting that a weak MrP model will yield worse estimates than disaggregation or poststratification.

Multilevel Regression and Synthetic Poststratification

While demographic poststratification variables should be chosen in order to maximize explained variance in the outcome, there is a strong constraint on whether a variable is eligible for being chosen: the joint distribution of the poststratification variables' cross-classifications must be known. This has been a major obstacle for MrP estimation, because subnational joint distributions are not available for many combinations of variables (Leemann & Wasserfallen, 2017).

To address this issue, Leemann and Wasserfallen (2017) recently introduced a procedure for conducting MrP with an estimated, or as they call it *synthetic*, poststratification joint distribution, which they derive from marginal distributions. They provide evidence that their method, multilevel regression and synthetic poststratification (MrsP; Leemann & Wasserfallen, 2017), performs better than raking, and comparably to or better than MrP (Leemann & Wasserfallen, 2017), depending on the predictive value of the added auxiliary variable (see Appendix or Leemann & Wasserfallen, 2017 for a detailed discussion of simple and adjusted MrsP).

To generate synthetic poststratification joint distributions, Leemann and Wasserfallen (2017) propose two approaches, which they refer to as "simple MrsP" and "MrsP with adjusted synthetic joint distributions." Under simple MrsP, synthetic joint distributions are calculated merely as the product of the poststratification variables' marginal distributions. For example, if only the county-level marginal distribution of age and education is known, their county-level simple synthetic joint distribution would be estimated as the product of their county-level marginal distributions. Importantly, the same approach can be used to extend a known demographic joint distribution to include other demographic variables for which only the marginal distribution is known. For example, if both the county-level joint distribution of age and gender and the county-level marginal distribution of education are known, their synthetic joint distribution could be estimated as the product of the joint and marginal distributions.

However, a notable short-coming of simple MrsP is that the estimated joint distribution will only be correct when the auxiliary variables are independent. As they diverge from independence, the synthetic joint distribution becomes a less accurate (Leemann & Wasserfallen, 2017). Accordingly, while Leemann and Wasserfallen (2017) find that errors in the synthetic joint distribution do not necessarily induce errors in poststratified, subnational estimates, they also propose a procedure for adjusting synthetic joint distributions. The goal of this adjustment procedure, or "adjusted MrsP," is to encode any available knowledge about the true joint distribution in the synthetic joint distribution. That is, rather than simply estimating the synthetic joint distribution of age, gender, and education as the product of the joint distribution of Age × Gender and the marginal distribution of education, adjusted MrsP would involve using external data (e.g., from a nationally representative survey) to adjust the simple synthetic joint distribution to reflect known correlations between age, gender, and education (see Appendix or Leemann & Wasserfallen, 2017 for a more detailed discussion of simple and adjusted MrsP).

Regardless of whether simple or adjusted synthetic joint distributions are used, after calculating the synthetic joint distribution, MrsP follows the same procedure as MrP. That is, a hierarchical model predicting individual-level responses is estimated, this model is used to generate predictions for each poststratum cell, and these predictions are weighted by their corresponding population weights. The only difference is that the population weights represent a synthetic joint distribution.

Study 1

Previous evaluations of MrP have found that it offers diminishing returns, compared with disaggregation, as sample sizes increase. For instance, Lax and Phillips (2009) found that disaggregation and MrP performed comparably with a sample size of approximately 14,000 in a state-level analysis with 49 states. However, this convergence in performance is contingent on the sample being representative, because, as sample size is increased, a representative sample systematically approaches the population.

In contrast, when a sample is not representative (e.g., due response biases), increasing the sample size does not necessarily yield a more precise approximation of the population. For example, if a particular population segment is not represented in a sample, estimates for subnational units populated by that segment may be biased and the degree of that bias will partly depend on the population portion of that segment. Under such conditions, the performance of disaggregation and MrP will still move toward convergence as sample sizes increase as, after all, an exhaustive sample would be equivalent to the population. However, this convergence will be inhibited by the degree to which the sample departs from representativeness.

To better understand this process, in this study, we investigate the results of Monte Carlo simulations (Mooney, 1997) in which disaggregation and MrP estimates are compared under different conditions of sample size and response bias. Here, our focus is primarily on the performance of MrP as both sample size and bias increase. Specifically, we focus on simulated sample sizes of 1,000, 10,000, 50,000, and 100,000 that are drawn from a subnational area containing 400 regions. These sample sizes were selected so that we could evaluate the relative performance of MrP as a function of a sample size, where sample size ranges from what would be considered a small sample for subnational estimation to a size that would be considered relatively large. Notably, some recent subnational investigations of psychological phenomena have used samples an order of magnitude larger than our largest simulated sample size. However, these investigations were conducted at the county-level and thus focused on a subnational area containing more than 3,000 subnational units. To roughly approximate the ratio of these orders of magnitude while also maintaining computational feasibility, we selected 100,000 as our largest simulated sample size. Finally, in this simulation, we include disaggregation estimates as a performance baseline, but do not include other estimation methods, such as disaggregation combined with poststratification or raked weights. Instead, we report a more comprehensive evaluation of relative performance on real-world data in Study 2.

Method

Our Monte Carlo simulation is structured as follows. First, a population of respondents is generated over a grid of 400 subnational units. Specifically, over the grid of units, marginal distributions of demographic characteristics are sampled for three demographic variables: one two-level, γ^1 variable and two three-level variables, γ^2 and γ^3. Then, constrained by these marginal distributions, demographic characteristics are assigned to simulated respondents within each subnational unit. Thus, each respondent is associated with a specific level of each γ^d, for d in $d = 1, 2, 3$.

Then, given the generated population, a set of population weights are randomly drawn. These weights consist of linear effects (i.e., model parameters) for the demographic and contextual factors, a random subnational unit effect, and individual-level error. These population weights are used to generate values for the response variable $Y_{l[i]}$.

Next, samples of sizes $S = (1{,}000, 10{,}000, 50{,}000, 100{,}000)$ are drawn from the population for each of three degrees of response bias. To simulate response bias, respondents are drawn with specified probabilities for each level of one of the three-level demographic variables. Specifically, three different degrees of bias are examined: $p = (1/3,1/3,1/3)$, $p = (0.4, 0.35, 0.25)$, and $p = (0.70, 0.2, 0.1)$. Further, to simulate random residual response-bias at the subnational unit level, response probabilities are randomly assigned to each subnational unit, which makes responses from some units more likely than others. Finally, given a drawn sample, disaggregation and MrP estimates of unit means are obtained and root mean squared error and bias are calculated.

For this study, three populations (Ns = 10,000,000) were simulated. Then, for each population, 50 sets of population weights were sampled. Finally, for each combination of the four sample sizes and three levels of response bias, 100 samples were drawn. This yielded a total of 180,000 iterations, or 15,000 iterations for each combination of sample size and response bias. These settings were chosen in order to minimize uncertainty while also maintaining reasonable computational cost. For a detailed description of the data generating process, please see Study 1 Data Generating Process in the online supplemental material.

Results

For each sample drawn within our simulation framework, disaggregation and MrP estimates were obtained and used to calculate root mean square error (RMSE) and bias. To evaluate the performance of these methods as a function sample size and response bias, we estimated the mean RMSE and bias for each method within each combination of sample size and response bias across samples, population weights, and populations.

As expected, per previous findings, under conditions of low response bias, the performance of disaggregation quickly converges with that of MrP (see Table 1 and Figure 1). For example, with samples of 100, disaggregation's expected RMSE ($\bar{X} = 0.59$, $\hat{\sigma} = 0.04$) was 0.22 higher than MrP's ($\bar{X} = 0.37$, $\hat{\sigma} = 0.04$). However, increasing the sample size to only 1,000 considerably reduced this gap, such that disaggregation's expected RMSE ($\bar{X} = 0.18$, $\hat{\sigma} = 0.01$) was 0.05 higher than MrP's ($\bar{X} = 0.13$, $\hat{\sigma} = 0.01$). Further, with samples of 10,000, disaggregation's expected RMSE ($\bar{X} = 0.06$, $\hat{\sigma} < 0.01$) was only about 0.015 higher than MrP's ($\bar{X} = 0.04$, $\hat{\sigma} < 0.01$).

However, while the convergence of disaggregation and MrP's expected RMSE only decreases slightly under medium response bias, under high response bias the convergence is attenuated considerably. Further, our simulations suggest that as response bias increases, disaggregation's RMSE becomes more variable. Notably, this increase in variance is not observed for MrP.

Regarding the estimation bias of disaggregation and MrP, neither method showed strong mean bias under any conditions (see Table 1 and Figure 1). However, the variances of their estimates of bias

Table 1
Mean RMSE and Bias by Sample Size and Bias

		Mean RMSE		Mean bias	
Sample size	Bias	Disaggregation	MrP	Disaggregation	MrP
1,000	Low	.594 (.04)	.368 (.04)	.001 (.03)	0 (.02)
1,000	Medium	.595 (.04)	.369 (.04)	.009 (.04)	.001 (.02)
1,000	High	.61 (.04)	.38 (.04)	.022 (.11)	0 (.03)
10,000	Low	.184 (.01)	.129 (.01)	0 (<.01)	0 (<.01)
10,000	Medium	.19 (.02)	.13 (.01)	.011 (.04)	0 (<.01)
10,000	High	.245 (.05)	.14 (.02)	.031 (.13)	0 (<.01)
50,000	Low	.08 (<.01)	.058 (<.01)	0 (<.01)	0 (<.01)
50,000	Medium	.091 (.01)	.058 (<.01)	.011 (.04)	0 (<.01)
50,000	High	.16 (.06)	.062 (<.01)	.031 (.13)	0 (<.01)
100,000	Low	.056 (<.01)	.041 (<.01)	0 (<.01)	0 (<.01)
100,000	Medium	.071 (.02)	.041 (<.01)	.011 (.04)	0 (<.01)
100,000	High	.146 (.07)	.044 (<.01)	.031 (.13)	0 (<.01)

Note. MrP = multilevel regression and poststratification; RMSE = root mean square error.

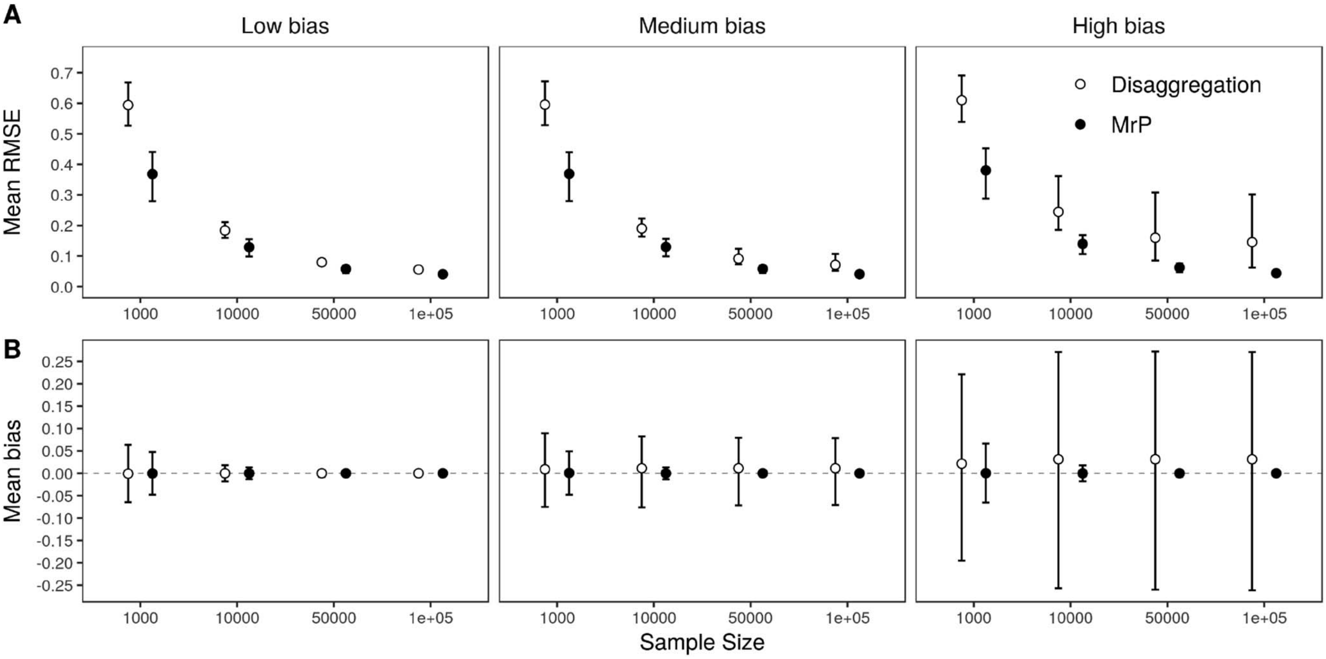

Figure 1. RMSE (A) and mean bias (B) as function of sample size and bias. Error bars represent 2.5th and 97.5th percentiles in simulation distribution. MrP = multilevel regression and poststratification; RMSE = root mean square error.

demonstrated starkly different patterns. Specifically, while, in the low bias condition, the variance of both disaggregation and MrP's bias shrunk toward zero as sample size increased, this remained true only for MrP as response bias was introduced. That is, as expected, under conditions of response bias, increasing the sample size had virtually no effect on the estimation bias of disaggregation.

Discussion

Previous evaluations of MrP suggested that the performance disaggregation and MrP converges as sample size increases. However, these evaluations were conducted with representative samples. In this study, we show that this convergence is inhibited as response biases are introduced to the sampling mechanism. Specifically, we found that under conditions of response bias, MrP considerably outperformed disaggregation in terms of error and bias, regardless of sample size.

However, it is important to interpret MrP's superior performance in these simulations in context. In any situation, MrP's performance will depend on the association between the outcome and the variables selected for poststratification, the contextual factor(s), and the modeled residual hierarchical variance (Buttice & Highton, 2013). Accordingly, the degree to which MrP outperforms disaggregation is dependent on strength and comprehensiveness of the MrP model. If a MrP analysis does not include a strong contextual factor or it poststratifies on demographic variables that explain very little variance in the outcome, the estimates generated by the analysis will be worse than those generated by a stronger MrP analysis. This means that a weak MrP analysis may not substantively outperform disaggregation, as performance depends on the quality of the MrP model. We emphasize this not to suggest that disaggregation is a viable alternative to MrP, but rather to highlight the importance of building a strong MrP model.

However, this raises the question of whether MrP will perform when applied to real-world psychological data. Further, while the current study provides evidence that MrP performs well under response bias, it does not address MrP's performance relative to alternative approaches to survey adjustment, such as poststratification and raking. We address these issues in the next study by comparing real-world ground-truth to county-level estimates obtained via application of MrP, MrsP, disaggregation, and raking to real-world data.

Study 2

In the current study, we use disaggregation, poststratification, poststratification with raking, MrP, and MrsP to obtain county-level estimates of Catholic Adherence from data collected by Project Implicit (Xu et al., 2013). We then compare these estimates to ground-truth, which has been obtained from the 2010 U.S. Religious Census (Grammich, 2012).

While we could have selected data from other sources for this study, we chose to focus on Project Implicit data—specifically, their Public IAT Racial Bias data—for two reasons. First, it exemplifies the kind of large-scale data that can be collected via online, opt-in collection strategies. Over 17 years of operation, Project Implicit has collected millions of responses and it offers an unprecedented opportunity to examine subnational variations in racial bias and associations between racial bias and secondary outcomes. Second, because of these characteristics, these data have been increasingly used to estimate subnational racial bias and we expect that such applications will only become more popular in the future. As such, we believe that it is particularly important to evaluate subnational estimates based on these data using a variety of procedures.

To this end, we evaluate estimates obtained from five different estimation procedures: (a) disaggregation, (b) disaggregation with rates obtained via raking, (c) disaggregation with weights obtained via poststratification and raking, (d) MrP, and (e) MrsP. Specifically, these methods are used to estimate the county-level rate of Catholicism. We then evaluate estimate accuracy and bias via RMSE and mean average bias (MAB).

Data

Project implicit data. The primary data used for this study were responses to an item measuring religious affiliation which was administered to participants in Project Implicit's (Xu et al., 2013) racial bias IAT survey from 2002–2013. This item, along with items measuring participants' age, level of education, sex, race, and county were obtained from the 2002–2018 Public Racial Bias IAT Open Science Foundation repository.[1]

2010 U.S. Religious Census. Ground-truth estimates of Catholic adherents for 3,105 counties located in the contiguous U.S. was obtained from the county-level 2010 U.S. Religious Census (Grammich et al., 2010) data, which was downloaded from the Association of Religion Data Archives http://www.thearda.com/Archive/.

2011–2015 American Community Survey 5-year estimates. To estimate county-level joint distributions, we rely on U.S. Census data obtained from the 2011–2015 American Community Survey (ACS) 5-year estimates and 2010 decennial U.S. Census data. We also use these data to estimate county-level proportion of Blacks, Latinos, people below the poverty line, proportion of population living in an urban area, and population density. Census data was accessed using the "tidycensus" (Walker, 2019) and "acs" R packages.

MIT election data. To estimate county-level 2016 Democratic vote proportion, we use data obtained from the MIT Election Data and Science Lab (MIT Election Data & Science Lab, 2018).

Geographic data. County names, Federal Information Processing Standard (FIPS) codes, and locations were determined using data accessed via the "USAboundaries" R package (Mullen & Bratt, 2017), which provides access to the U.S. Census Bureau's geographic database (U.S. Census Bureau, 2015).

Method

In this study, we estimate county-level Catholic adherence rates using disaggregation, poststratification, poststratification and raking, MrP, and MrsP. For our standard MrP estimates, age, gender, and race are selected as poststratification variables, as these are directly available from the U.S. Census. In contrast, for our MrsP estimates, we extend the poststratification variables to also include level of education by calculating its adjusted synthetic joint distribution with age, gender and race (Leemann & Wasserfallen, 2017).

Primary data preparation. Participants in the IAT Implicit Race survey who reported religious affiliation and who were matched to counties located in the contiguous U.S. (3,018) were selected for analysis, $N = 3{,}014{,}859$. This yielded a data set with coverage of 3,088 counties (county N summary statistics: $Mean = 976$, $Median = 114$, $SD = 3, 440$). For the selected participants, race, age, sex, and education was coded as follows: race = (Black, Hispanic, Other, White); age = (18–29, 30–44, 45–64, 65+); sex = (female, male); education = (high school graduate or less; some college through bachelor's degree; graduate degree).[2] We selected these demographic categories with the goal of minimizing the risk of sparsity while maintaining as much demographic distinction as possible. That is, selecting a more fine-grained set of races would allow for more demographic distinction, but it would also increase sparsity as other races were far less frequent in our sample. Finally, participants' Catholic affiliation was represented as a binary indicator where "1" indicates self-reported Catholic affiliation.

Secondary data preparation. Four contextual variables were selected for the MrP and MrsP models: county-level Democratic vote share for the 2016 Presidential election and the proportion of Blacks, Latinos, people below the poverty line, population living in an urban area, and population density. These variables were selected based on a priori expectations about their potential association with racial bias against Blacks. That is, we expected racial bias in a given area to be partially dependent on the areas Democratic vote share, the racial composition of the area, the number of people living below the poverty line, and the urban/rural status of the area. These variables were all standardized prior to inclusion in Mr(s)P models.

Disaggregation. We obtained disaggregated estimates of D score $y_{u[l]}$ for county l located in region u simply by calculating the sample mean for observations from county l:

$$\bar{y}^{1}_{u[l]} = \frac{\sum_{i}^{N_l} y_{u[l],i}}{N_l} \tag{4}$$

where $y_{u[l],i}$ represents response i from county $u[l]$, N_l represents the sample size for county $u[l]$, and $\bar{y}^{1}_{u[l]}$ is the estimated mean for county $u[l]$ obtained via Method 1, disaggregation.

Poststratification and raking. In addition to simple disaggregation, we also estimate two sets of poststratification weights and use these to perform weighted disaggregation. The first set of weights (Rake 1) were calculated via raking across the county-level marginal distributions of age, sex, and race using the same demographic levels as in the Mr(s)P models. To address issues of poststratum cell sparsity, to generate the second set of weights (Rake 2), we collapsed age into two levels (below/above 30 years of age), race into three levels (White, Black, Other), and education into two levels (no college/at least some college). These demographic collapses were selected in order to minimize sparsity—which helps stabilize sample estimates—while also maximizing demographic variety. Notably, the necessity of collapsing levels and choosing which levels to collapse is one of the major obstacles for poststratification and raking as there are no established guidelines for making these choices. We then performed a second raking procedure across the joint distribution of the collapsed age and race variables and the marginal distributions of gender and education. For each set of weights, county-level estimates of Catholic adherence were then generated via weighted disaggregation.

Raking was performed for each county based on the data present in that county. Thus, counties with insufficient demographic cov-

[1] https://osf.io/yn2g7/.

[2] We did not discriminate between professional and non-professional secondary degrees.

erage (e.g., counties for which data was fully missing for a demographic cell) were dropped from analysis. Iterations were limited to 1,000 and counties for which the raking procedure did not converge were discarded. In this context, convergence failure is typically caused by demographic sparsity and it indicates that stable sample weights could not be derived. When raking is used for nation-level analyses, it is common to identify the source of the sparsity and remove it by collapsing additional demographic variables. Unfortunately, this is often not practical when raking is applied to subnational estimation as this procedure would need to be repeated for each subnational area for which convergence was not reached. As an alternative, we exclude areas that did not reach convergence. Importantly, this approach could introduce bias into the distribution of subnational estimates, as counties for which weights do not converge could be systematically different from counties that do. However, for subnational estimation procedures that involve thousands of subnational areas, conducting boutique adjustments for each area is simply not practical. Raking was implemented using the "survey" R package. (Lumley, 2004).

MrP. To estimate county-level Catholic adherence via MrP, we modeled individual Catholic affiliation using a hierarchical generalized linear model with a logit link estimated using the "lme4" R package (Bates, Mächler, Bolker, & Walker, 2015). In this model, age, race, and sex were included as demographic variables, county, state, and division were included as geographic levels, and county-level Democratic vote share for the 2016 Presidential Election and the proportion of Blacks, Latinos, people below the poverty line, population living in an urban area, and population density were included as contextual factors.

These demographic variables were selected based on data available information from the U.S. Census. Further, we included county, state, and division as geographic random effects in order to maximize the benefits of partial pooling. In this design, the random effect for a county with few respondents will be shifted toward the intercept for the state containing the county. Similarly, state means are assumed to be distributed around a random region intercept. Accordingly, this three-level structure allows the model to reflect potentially complex regional patterns across the U.S. Finally, we selected our contextual factors based on a priori expectations about their potential association with Catholicism. That is, we expected the prevalence of Catholicism in a given area to be partially dependent on the area's Democratic vote share, the racial composition of the area, the number of people living below the poverty line, the urban/rural status of the area, and the area's density.

These variables were all standardized prior to inclusion in Mr(s)P models. Specifically, we estimated the following model:

$$\begin{aligned}
&P(Y=1) = \text{logistic}(\eta)\\
&\eta = \beta_0 + \alpha_{division:race[i]} + \alpha_{division:sex:age[i]} + \alpha_{state[i]} + \alpha_{county[i]}\\
&\alpha_{division:race[i]} \sim \mathcal{N}(0, \sigma_a^2), \text{ for}\\
&\qquad division:race = 1, \ldots, Division \times Race\\
&\alpha_{division:sex:age[i]} \sim \mathcal{N}(0, \sigma_e^2), \text{ for}\\
&\qquad division:sex:age = , \ldots, Division \times Sex \times Age\\
&\alpha_{county} \sim \mathcal{N}(\alpha_{state[county]} + \beta \boldsymbol{X_c}, \sigma_{county}^2), \text{ for}\\
&\qquad county = 1, \ldots, County\\
&\alpha_{state} \sim \mathcal{N}(0, \sigma_{state}^2), \text{ for } state = 1, \ldots, State.
\end{aligned} \tag{5}$$

Specified using "lme4" this model would be:

$$\begin{aligned}
\text{glmer}(y = 1 + \ldots +\\
&(1\,|\,\text{county})+\\
&(1\,|\,\text{state})+\\
&(1\,|\,\text{Division}:\text{race})+\\
&(1\,|\,\text{Division}:\text{sex}:\text{age}))
\end{aligned}$$

where ". . ." includes fixed effects for each contextual factor. That is, we estimated random intercepts at both the county ($N = 3{,}088$) and state ($N = 48$) levels. Further, demographic effects were estimated as random intercepts crossed with division. Specifically, a random intercept was estimated for each level of race within each of the nine U.S. divisions. Similarly, the interaction of sex with age was also crossed with division. Initially, we did not cross the demographic effects with division; however, models estimated on the full data set with this specification did not reach convergence after many iterations. In contrast, we found that models that crossed demographic effects with division converged relatively quickly.

This model was then used to make predictions $\pi_{county,j}$ for each cross-classification j of race, age, and gender within each county. Finally, the poststratification step was implemented using the county-level population joint distribution for race, age, and gender estimated by the U.S. Census:

$$\bar{Y}_{county} = \sum_j P_{county,j}\pi_{county,j}. \tag{6}$$

MrsP. MrsP estimates were obtained following exactly the same procedure as for MrP. However, an additional random effect for education was estimated. As in the MrP model, the random effect for education was crossed with division.

To obtain poststratified estimates, we calculated the adjusted synthetic joint distribution between the county-level joint distribution of race, age, and gender and the county-level marginal distribution of education. To inform the adjustment procedure, we relied on national-level estimates of the full joint distribution of these variables obtained from the U.S. Census. Using the adjusted synthetic county-level joint distribution, poststratification was implemented following the same procedure used for MrP.

Results

To evaluate the relative performance of each estimation method, we used the rate of Catholic adherents reported by the 2010 U.S. Religious Census (Grammich et al., 2010) to calculate RMSE and MAB for each set of estimates (see Table 2). The results indicate that both MrP (0.09) and MrsP (0.09) slightly outperform the estimates obtained via disaggregation (0.12), raking over the marginal distributions of age, sex, and race (Rake 1; 0.10), and collapsing the levels of the demographic variables and raking over the joint distribution of age and race and marginal distributions of sex and education (Rake 2; 0.10). Regarding the average bias of the estimates, all estimates were slightly negatively biased, but the MrP and MrsP estimates were slightly less biased than the estimates obtained via disaggregation and the first raking procedure.

However, it is important to interpret these results relative to county coverage. For example, while MrP and MrsP perform only slightly better than the other methods, they offer complete coverage of the 3,105 counties for which ground-truth was obtained.

Table 2
Performance Metrics for Estimates of County-Level Catholic Adherence

Method	RMSE	MAB	*N* Counties
Disaggregation	.12	−.02	3,076
MrP	.09	−.01	3,105
MrsP	.09	−.01	3,105
Rake 1	.10	−.02	1,468
Rake 2	.10	−.01	2,155

Note. MAB = mean average bias; MrP = multilevel regression and poststratification; MrsP = multilevel regression and synthetic poststratification; RMSE = root mean square error.

While disaggregation also offered almost complete coverage, it showed the worst performance in terms of RMSE and MAB. Further, the raking procedure that preserved the original coding of the demographic variables offered coverage of only 1,468 counties. While the procedure that collapsed the demographic variables and included the joint distribution of age and race offered better coverage, it still missed nearly 1,000 counties. Notably, if RMSE is calculated for MrP and MrsP over only the counties covered by the raking procedures, it drops to 0.08.

Finally, to evaluate the overall associations between ground-truth and each set of estimates, we calculated their correlations and plotted their lines of linear fit (see Figure 2). Notably, MrP (r = .73) and MrsP (r = .74) were substantially more strongly correlated with ground-truth, compared with the estimates obtained via disaggregation (r = .59) and both raking procedures, Raking 1 r = .62 and Raking 2 r = .65.

Discussion

These results suggest that even with an extremely large sample, Mr(s)P can be used to obtain subnational estimates that are clearly superior to methods that have been more widely used in the psychological literature. Notably, while the Mr(s)P estimates only slightly reduced error and bias, these improvements were achieved over the full set of counties. In contrast, the other methods were only able to generate estimates for a subset of counties. Further, estimates obtained via Mr(s)P also showed substantially stronger correlations with ground-truth.

That said, it should be noted that we only compared Mr(s)P estimates to estimates derived from two different raking/poststratification procedures. Thus, it is certainly possible that, out of the universe of possible raking configurations, a better performing configuration may exist. Nonetheless, for both configurations we sought to include as much information as possible while also minimizing issues caused by demographic sparsity.

It is also notable that MrsP offered virtually no improvement in performance, relative to MrP. This, of course, is a function of the conditional relationship between the demographic variable added for MrsP (education) and the outcome (Catholic affiliation). Simply put, in this case, adding education to the MrP model did not improve its accuracy. Accordingly, in our view, depending on the research context, researchers should still consider the possible benefits of extending the poststratification joint distribution.

Ultimately, these results provide evidence that MrP should be preferred for obtaining subnational estimates from large-scale convenience data. While other methods performed only slightly worse in terms of error and bias, MrP offered better coverage of subnational units and stronger correlations with ground-truth. Further, it is worth noting that, in our experience, estimating a single MrP model over a set of subnational units is considerably simpler and allows far fewer researcher degrees of freedom than obtaining sample weights via poststratification or raking because MrP does not require the arbitrary collapsing of demographic categories.

Study 3

Results from the previous study indicated that Mr(s)P offered both superior performance and better coverage of subnational units, relative to unweighted disaggregation and weighted disaggregation. However, in instances where prediction accuracy is not the central focus, these results might raise the question of whether it matters which estimation method is used. For example, in psychology, researchers are often primarily interested in obtaining county-level estimates of a given construct and then drawing inferences about the association between this construct and a second county-level outcome. In such situations, to what extent might it matter which estimation procedure researchers use?

In this study, we address this question by estimating the association between county-level implicit and explicit racial bias and Barack Obama's Presidential vote share in 2008. In addressing this question, our goal is not necessarily to provide evidence for or against an association between these constructs. Rather, we are interested in how conclusions about this association might vary

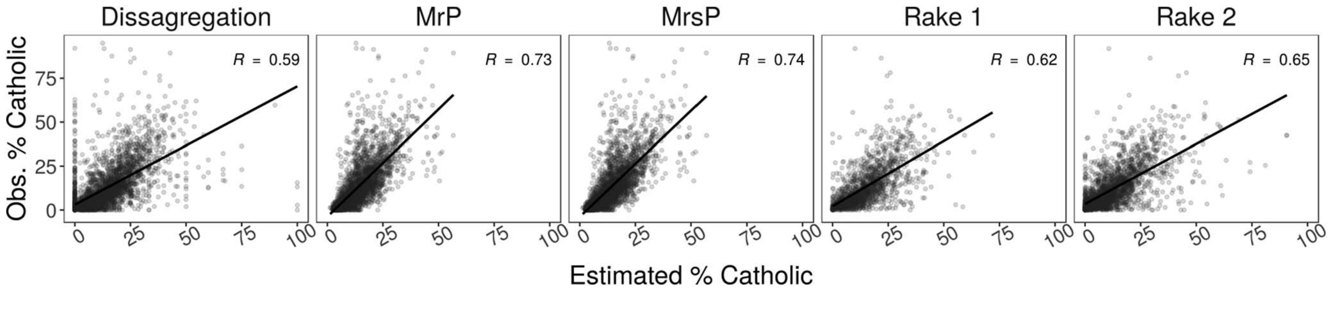

Figure 2. Observed (y-axis) versus estimated (x-axis) county-level % Catholic for each estimation method. Points represent counties. Correlation coefficient for predicted versus observed values shown in top right of each panel. MrP = multilevel regression and poststratification; MrsP = multilevel regression and synthetic poststratification.

depending on how county-level racial bias is estimated. Nonetheless, for the purposes of this study, we sought to test the following hypothesis:

> *Hypothesis 1:* Controlling for 2004 county-level Democratic vote share, White's implicit and explicit county-level racial bias against Blacks should be negatively associated with Barack Obama's 2008 county-level Presidential vote share.

Data

The primary data used for this study were responses to the race implicit association test (IAT) obtained from Project Implicit (Xu et al., 2013) and collected between 2002 and 2017. The IAT relies on a timed dual-categorization task that requires respondents to evaluate pairings of White and Black faces and words referring to "good" and "bad" things. An indication of racial bias (against Blacks) occurs when a respondent more quickly categorizes words representing "bad" things as bad when they are paired with a Black face, compared with a White face, and when they are able to more quickly categorize "good" words when they are paired with a White face (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). Through repeated measures sampling across categorization trails, the IAT permits the estimation of the so called D score, which represents the difference in response latency. D scores range from −2.0 to 2.0, where scores above 0 indicate a positive bias toward White faces and a negative bias toward Black faces. Participants who completed the Race IAT and who were located in a county in the contiguous U.S. were retained for analysis, N = 1,704,789.

Explicit racial bias toward Blacks was measured −10 to 10 via single item reflecting participants' "warmth toward Blacks," N = 1,091,841.

All other data sources were identical to those reported in Study 2.

Method

As in Study 2, we generated county-level estimates using unweighted disaggregation, MrP, MrsP, and two variations of weighted disaggregation, where weights were calculated using a combination of raking and poststratification. The MrP and MrsP models were identical to those reported in Study 2, with the exception that for this study racial bias was modeled as a continuous random variable. For each method, estimates of both implicit and explicit racial bias were obtained.

After obtaining estimates of county-level implicit and explicit racial bias using each estimation procedure, we estimated separate linear regression models in which Barack Obama's 2008 county-level Presidential vote share was regressed on either county-level implicit or explicit racial bias. As controls, we also included John Kerry's 2004 county-level Presidential vote share as well as the county-level proportion of Blacks, Latinos, people living in urban areas, people living below the federal poverty line, density. Controlling for these variables is essential as they are used in obtaining the Mr(s)P estimates. All independent variables were standardized in all models. We then examined the estimated coefficient for either measure of racial bias across estimation methods.

Results

As expected, MrP and MrsP provided better county coverage (see Table 3). Further, the smoothing effects of the hierarchical model are evident in the reduced variance and more reasonably minimum and maximum values of the Mr(s)P estimates, relative to the other methods.

However, despite the reduced variance in the Mr(s)P models, for both implicit and explicit racial bias the estimated association between Barack Obama's 2008 Presidential vote share was substantially stronger for the models that relied on the Mr(s)P estimates (see Tables 4 and 5). For example, the association between implicit racial bias and Barack Obama's 2008 Presidential vote share was estimated as $b = -0.05\%$, $SE < 0.001$, 95% CI [−0.05, −0.042], indicating that a one SD increase in implicit racial bias—as estimated via MrP—was associated with an expected 5% decrease in Obama's county vote share. In contrast, estimates obtained from other methods were substantially weaker, though still statistically significant.

Across all models, the estimated effects for the control variables were equivalent within rounding error at three decimal places (see Table 6 for these estimates).

Discussion

Overall, these results clearly demonstrate that downstream inferences based on county-level estimates can vary dramatically depending on the method of estimation. Using the Mr(s)P estimates, we observed a stronger association between racial bias and Barack Obama's, 2008 Presidential vote share; however, using estimates obtained from the other methods, this association attenuated. Given our results in Studies 1 and 2 as well as other literature on MrP, we would be generally more inclined to trust inferences derived from MrP estimates as these can be reasonably expected to be the most accurate. Importantly, in some cases, such as when a large random sample is available, MrP may perform no better than disaggregation. However, when such a sample is not available or, further, when the available sample is subject to various sources of sampling and response bias, MrP can provide a more robust approach to obtaining subnational estimates, compared to other approaches like disaggregation, raking, or poststratification.

Table 3
Summary of Subnational Estimation of Implicit Racial Bias for Each Method

Method	N Counties	Mean	SD	Min	Max
Implicit bias–Disaggregation	3,086	.33	.10	−.62	.99
Implicit bias–MrP	3,105	.36	.07	−.03	.45
Implicit bias–MrsP	3,105	.36	.07	−.03	.45
Implicit bias–Rake 1	1,612	.35	.11	−.37	.78
Implicit bias–Rake 2	2,270	.34	.14	−.48	1.02
Explicit bias–Disaggregation	3,073	.53	.63	−10.00	5.00
Explicit bias–MrP	3,105	.55	.34	−1.37	1.20
Explicit bias–MrsP	3,105	.56	.34	−1.39	1.19
Explicit bias–Rake 1	1,612	.52	.58	−4.38	5.97
Explicit bias–Rake 2	2,269	.47	.81	−5.83	6.09

Note. MrP = multilevel regression and poststratification; MrsP = multilevel regression and synthetic poststratification.

Table 4
Estimated Conditional Association Between County-Level Implicit Racial Bias and Obama's 2008 Presidential Vote Share

Method	Estimate	*SE*	95% CI
Disaggregation	−.006	.001	[−.008, −.004]
MrP	−.047	.001	[−.051, −.042]
MrsP	−.048	.001	[−.053, −.044]
Raking 1	−.003	.001	[−.005, −.001]
Raking 2	−.003	.001	[−.005, −.001]

Note. MrP = multilevel regression and poststratification; MrsP = multilevel regression and synthetic poststratification.

Table 6
Estimated Effects for Control Variables

Parameter	Estimate	Std. error	*t* value	Pr(>|t|)
Intercept	.418	.001	547.000	0
Kerry vote share std.	.120	.001	127.000	0
% Black std.	−.029	.002	−17.200	0
% Urban std.	.009	.001	9.690	0
% below poverty std.	−.015	.001	−16.500	0
Density std.	−.001	.001	−1.670	.095
% Latino std.	0	.001	−.449	.653

However, it should also be noted that using MrP estimates to estimate associations with secondary variables raises unique challenges. For example, given that the estimates are a function of the contextual predictors in the MrP model, it is important to consider if and to what extent an association between MrP estimates and a secondary variable might be driven by masked associations with the contextual variables. Here we attempt to account for this possibility by including the variables used in the MrP model. However, researchers who use MrP in such instances should carefully consider these issues. Finally, it should be noted that MrP does not resolve issues of causal direction (Caughey & Warshaw, 2019). That said, given the gains in accuracy and robustness to sampling bias, in our view MrP is still likely one of the best methods for obtaining subnational estimates which will then be used in secondary analyses.

General Discussion

One of the primary difficulties for psychological research is establishing connections between hypothesized constructs and real-world phenomena. While psychologists are experts at simulating phenomena in laboratory settings and developing indirect (e.g., survey based) measures of target phenomena, establishing external validity remains one of the primary challenges for psychological research. While geographic studies of psychological phenomena raise their own challenges, they also directly supplement conventional approaches to psychological research. More specifically, geographic approaches to psychological research offer an opportunity to directly investigate the association between psychological constructs and real-world outcomes (Rentfrow & Jokela, 2016).

Table 5
Estimated Conditional Association Between County-Level Explicit Racial Bias and Obama's 2008 Presidential Vote Share

Method	Estimate	*SE*	95% CI
Disaggregation	.004	.001	[−.006, −.003]
MrP	.043	.001	[−.047, −.04]
MrsP	.040	.001	[−.044, −.036]
Raking 1	.003	.001	[−.006, −.001]
Raking 2	.004	.001	[−.006, −.002]

Note. MrP = multilevel regression and poststratification; MrsP = multilevel regression and synthetic poststratification.

This is, of course, not a revelation, as cultural psychology pioneered cross-cultural (e.g., international) studies decades ago. However, early investigations of the geographic study of psychological outcomes were limited by focus on nation-level variation. Recently, however, researchers have begun focusing psychological constructs' subnational variation and associations with target outcomes. This work has yielded a number of important findings, such as new evidence for the deleterious effects of racial prejudice (Hehman et al., 2017; Leitner et al., 2016a; Orchard & Price, 2017; Rae et al., 2015) and associations between personality and various target outcomes (Götz et al., 2018; Jokela et al., 2015; McCann, 2017a, 2017b, 2018; Rentfrow et al., 2015), as well as insight into the spatially structured patterning of personality (Jokela et al., 2015; Rentfrow et al., 2015, 2008, 2013) and moral values (Hoover et al., 2018).

However, while the subnational study of psychological constructs affords a range of exciting new opportunities for research, it also raises a number of challenges that are not commonly encountered or addressed in psychological research. In this work, our goal was to offer a comprehensive introduction to state-of-the-art methods, Mr(s)P, for addressing these challenges that have been developed in other fields. Beyond this, we sought to evaluate the performance of these methods under conditions similar to those in which many subnational studies of psychological phenomena have been conducted. Specifically, the methods reviewed in this work were both designed for and have been most often applied to relatively small, randomized, and representative samples. In contrast, psychologists today often find themselves working with large, nonrandom, and nonrepresentative samples with nonuniform subnational sparsity. Accordingly, in order to provide researchers with an informed introduction to these methods, we addressed several questions regarding the optimal approach to subnational estimation under such conditions: (a) whether MrP offers improvements in accuracy, compared with other methods, when applied to very large samples; and (b) whether estimates obtained via MrP yield different conclusions about associations with secondary variables, compared with estimates obtained via other methods.

Specifically, we found that MrP outperforms other commonly used methods, including disaggregation, raking, and poststratification, when applied to samples with response biases, even when those samples contain 100,000 (Study 1) or more than three million (Study 2) responses. In Study 1, we evaluated the differential performance of disaggregation—a simple but widely used approach to small-area estimation—and MrP under varying conditions of sample size and response bias. This study provided strong evidence that under conditions of response bias, MrP dramatically outperforms disaggregation, regardless of sample size. Importantly, previous research has shown that with even modestly large *random* samples (e.g., $N = 10{,}000$) disaggregation performs comparably to MrP. In contrast, our results suggest that in both simulated and real-world data, MrP outperforms disaggregation under conditions of response bias. Further, in Study 2, we provide evidence that even with a very large convenience sample, Mr(s)P is a better estimator than not only disaggregation, but also raking and poststratification. We also show that it also offers substantially better coverage of spatial units even with a sample size of more three million respondents.

Finally, in Study 3, we show that the associations between an estimated county-level construct and a secondary outcome can vary substantially depending on the estimation procedure used. Notably, the estimates obtained via Mr(s)P were consistent with previous literature; however, using Mr(s)P estimates for secondary analysis raises issues of potential contamination caused by the contextual variables included in the MrP model. To address this issue, we suggest that researchers who use MrP estimates in secondary analyses conduct sensitivity analyses by controlling for the contextual factors in the MrP model (e.g., in a regression model that includes MrP estimates as an independent variable) and, ideally, compare inferences across multiple estimation strategies (e.g., against estimations obtained via poststratification).

Overall, given the relative ease of implementing MrP and the improvements in estimation accuracy and stability that it offers, we see little reason for this method to not be more widely applied to subnational geographic studies of psychological phenomena. While today's large, online, opt-in samples have opened many new opportunities for studying the geographic distribution of psychological constructs, it is important that limitations of these samples are accounted for as well as is possible. In our view, MrP is a useful tool that can help psychological researchers work toward this goal.

References

Allik, J., Realo, A., Mõttus, R., Pullmann, H., Trifonova, A., McCrae, R. R., & 56 Members of the Russian Character and Personality Survey. (2009). Personality traits of Russians from the observer's perspective. *European Journal of Personality, 23,* 567–588. http://dx.doi.org/10.1002/per.721

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Buttice, M. K., & Highton, B. (2013). How does multilevel regression and poststratification perform with conventional national surveys? *Political Analysis, 21,* 449–467. http://dx.doi.org/10.1093/pan/mpt017

Caughey, D., & Warshaw, C. (2019). Public opinion in subnational politics. *The Journal of Politics, 81,* 352–363. http://dx.doi.org/10.1086/700723

Claassen, C., & Traunmüller, R. (2018). Improving and validating survey estimates of religious demography using Bayesian multilevel models and poststratification. *Sociological Methods & Research*. Advance online publication. http://dx.doi.org/10.1177/0049124118769086

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association, 88,* 1013–1020. http://dx.doi.org/10.2307/2290793

Erikson, R. S., Wright, G. C., & McIver, J. P. (1993). *Statehouse democracy: Public opinion and policy in the American States*. Cambridge, UK: Cambridge University Press.

Fowler, L. (2016). The states of public opinion on the environment. *Environmental Politics, 25,* 315–337. http://dx.doi.org/10.1080/09644016.2015.1102351

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22,* 153–164. http://dx.doi.org/10.1214/088342306000000691

Gelman, A. (2014). How Bayesian analysis cracked the Red-State, Blue-State problem. *Statistical Science, 29,* 26–35. https://www.jstor.org/stable/43288447

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23,* 127–135. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.5270

Götz, F. M., Ebert, T., & Rentfrow, P. J. (2018). Regional cultures and the psychological geography of Switzerland: Person–environment–fit in personality predicts subjective wellbeing. *Frontiers in psychology, 9,* 517–532.

Grammich, C. A. (2012). *2010 U.S. Religion Census: Religious congregations & membership study: An enumeration by nation, state, and county based on data reported for 236 religious groups*. Lenexa, KS: Association of Statisticians of American Religious Bodies.

Grammich, C., Hadaway, K., Houseal, R., Jones, D. E., Krindatch, A., Stanley, R., & Taylor, R. H. (2010). *U.S. Religion Census: Religious congregations & membership study*. Lenexa, KS: Association of Statisticians of American Religious Bodies.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17–41.

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 U.S. presidential election. *Analyses of Social Issues and Public Policy, 9,* 241–253.

Hanretty, C., Lauderdale, B. E., & Vivyan, N. (2016). Comparing strategies for estimating constituency opinion from national survey samples. *Political Science Research and Methods, 6,* 1–21. http://dx.doi.org/10.1017/psrm.2015.79

Hehman, E., Flake, J. K., & Calanchini, J. (2017). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*. Advance online publication.

Holt, D., & Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A (General), 142,* 33–46. http://dx.doi.org/10.2307/2344652

Hoover, J., & Dehghani, M. (2018). The big, the bad, and the ugly: Geographic estimation with flawed psychological data. *PsyArXiv*. Retrieved from psyarxiv.com/bthqc http://dx.doi.org/10.31234/osf.io/bthqc

Hoover, J., Zhao, C., & Dehghani, M. (2018). *MapYourMorals*. Retrieved from https://mapyourmorals.usc.edu/#/

Howe, P. D., Mildenberger, M., Marlon, J. R., & Leiserowitz, A. (2015). Geographic variation in opinions on climate change at state and local scales in the USA. *Nature Climate Change, 5,* 596–603. http://dx.doi.org/10.1038/nclimate2583

Jokela, M., Bleidorn, W., Lamb, M. E., Gosling, S. D., & Rentfrow, P. J. (2015). Geographically varying associations between personality and life satisfaction in the London metropolitan area. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 725–730. http://dx.doi.org/10.1073/pnas.1415800112

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics, 19,* 81–97. http://www.jos.nu/Articles/abstract.asp?article=192081

Kastellec, J. P., Lax, J. R., Malecki, M., & Phillips, J. H. (2015). Polarizing the electoral connection: Partisan representation in supreme court confirmation politics. *The Journal of Politics, 77,* 787–804. http://dx.doi.org/10.1086/681261

Krimmel, K., Lax, J. R., & Phillips, J. H. (2016). Gay rights in congress: Public opinion and (mis) representation. *Public Opinion Quarterly, 80,* 888–913.

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the states? *American Journal of Political Science, 53,* 107–121. http://dx.doi.org/10.1111/j.1540-5907.2008.00360.x

Leemann, L., & Wasserfallen, F. (2017). Extending the use and prediction precision of subnational public opinion estimation. *American Journal of Political Science, 61,* 1003–1022. http://dx.doi.org/10.1111/ajps.12319

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016a). Blacks' death rate due to circulatory diseases is positively related to whites' explicit racial bias: A nationwide investigation using project implicit. *Psychological Science, 27,* 1299–1311. http://dx.doi.org/10.1177/0956797616658450

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016b). Racial bias is associated with ingroup death rate for blacks and whites: Insights from project implicit. *Social Science & Medicine, 170,* 220–227. http://dx.doi.org/10.1016/j.socscimed.2016.10.007

Little, R. J. A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association, 88,* 1001–1012. http://dx.doi.org/10.1080/01621459.1993.10476368

Lohr, S. L. (2009). *Sampling: Design and analysis*. Boston, MA: Nelson Education.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9,* 1–19.

McCann, S. J. H. (2017a). Higher USA state resident neuroticism is associated with lower state volunteering rates. *Personality & Social Psychology Bulletin, 43,* 1659–1674. http://dx.doi.org/10.1177/0146167217724802

McCann, S. J. H. (2017b). The relation of state resident neuroticism levels to state cancer incidence in the USA. *Current Psychology*. Advance online publication. http://dx.doi.org/10.1007/s12144-017-9774-6

McCann, S. J. H. (2018). U.S. state resident big five personality and work satisfaction: The importance of neuroticism. *Cross-Cultural Research, 52,* 155–191. http://dx.doi.org/10.1177/1069397117723607

MIT Election Data and Science Lab. (2018). County presidential election returns 2000–2016. *Harvard Dataverse*. Retrieved from http://dx.doi.org/10.7910/DVN/VOQCHQ

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: SAGE Publications.

Mullen, L., & Bratt, J. (2017). *USA boundaries: Historical and contemporary boundaries of the united states of America* (R package version 0.3.0) [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=USAboundaries

Obschonka, M., Stuetzer, M., Audretsch, D. B., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2016). Macropsychological factors predict regional economic resilience during a major economic crisis. *Social Psychological and Personality Science, 7,* 95–104. http://dx.doi.org/10.1177/1948550615608402

Orchard, J., & Price, J. (2017). County-level racial prejudice and the black-white gap in infant health outcomes. *Social Science & Medicine, 181,* 191–198. http://dx.doi.org/10.1016/j.socscimed.2017.03.036

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-Level estimates from national polls. *Political Analysis, 12,* 375–385. http://dx.doi.org/10.1093/pan/mph024

Pew Research Center. (2018). *Internet/broadband fact sheet*. Retrieved from http://www.pewinternet.org/fact-sheet/internet-broadband/

Rae, J. R., Newheiser, A.-K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the united states. *Social Psychological and Personality Science, 6,* 535–543. http://dx.doi.org/10.1177/1948550614567357

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Rentfrow, P. J., Gosling, S. D., Jokela, M., Stillwell, D. J., Kosinski, M., & Potter, J. (2013). Divided we stand: Three psychological regions of the united states and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology, 105,* 996–1012. http://dx.doi.org/10.1037/a0034434

Rentfrow, P. J., Gosling, S. D., & Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science, 3,* 339–369. http://dx.doi.org/10.1111/j.1745-6924.2008.00084.x

Rentfrow, P. J., & Jokela, M. (2016). Geographical psychology: The spatial organization of psychological phenomena. *Current Directions in Psychological Science, 25,* 393–398. http://dx.doi.org/10.1177/0963721416658446

Rentfrow, P. J., Jokela, M., & Lamb, M. E. (2015). Regional personality differences in Great Britain. *PLoS ONE, 10,* 1–20. http://dx.doi.org/10.1371/journal.pone.0122245

Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. *Proceedings of the National Academy of Sciences of the United States of America, 116,* 8255–8260. http://dx.doi.org/10.1073/pnas.1808307116

Selb, P., & Munzert, S. (2011). Estimating constituency preferences from sparse survey data using auxiliary geographic information. *Political Analysis, 19,* 455–470. http://dx.doi.org/10.1093/pan/mpr034

Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science, 46,* 218–237. http://dx.doi.org/10.2307/3088424

U.S. Census Bureau. (2015). *2015 Census Bureau's MAF/TIGER geographic database*. Retrieved from https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html

Walker, K. (2019). *tidycensus: Load U.S. Census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames* (R package version 0.9.2) [Computer software manual]. Retrieved from https://github.com/walkerke/tidycensus

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31,* 980–991. http://dx.doi.org/10.1016/j.ijforecast.2014.06.001

Wang, Y., Holt, J. B., Xu, F., Zhang, X., Dooley, D. P., Lu, H., & Croft, J. B. (2018). Using 3 health surveys to compare multilevel models for small area estimation for chronic diseases and health behaviors. *Preventing Chronic Disease, 15,* E133. http://dx.doi.org/10.5888/pcd15.180313

Xu, F. K., Lofaro, N., Nosek, B. A., & Greenwald, A. G. (2013). *Race IAT 2002–2017*. Retrieved from https://osf.io/52qxl/

(*Appendix follows*)

Appendix

Simple MrsP

To generate synthetic post-stratification joint distributions, Leemann & Wasserfallen (2017) propose two approaches, which they refer to as ‘simple MrsP’ and ‘MrsP with adjusted synthetic joint distributions’. Under simple MrsP, synthetic joint distributions are calculated merely as the product of the post-stratification variables’ marginal distributions. Thus, for a given set of post-stratification variables, the joint distributions for each sub-national unit *k* is estimated as the product of the post-stratification variables’ marginal distributions within each sub-national unit. For example, for an arbitrary sub-national unit *k*, the simple synthetic joint distribution of 3-level age and education P_{aek} can be estimated as the products of their marginal distributions, as shown in Table A1.

Table A1
Example of Simple MrsP Synthetic Joint Distribution for Age and Education

	age = 1	age = 2	age = 3	
edu = 1	0.01	0.05	0.04	0.1
edu = 2	0.08	0.4	0.32	0.8
edu = 3	0.01	0.05	0.04	0.1
	0.1	0.5	0.4	

Note. MrP = multilevel regression and synthetic poststratification. The marginal probabilities for each level of age and education are shown in the row and column margins, respectively. The simple synthetic joint distribution is shown in the interior cells.

It is also worth noting that simple MrsP can be used to extend a known joint distribution to include an additional variable for which only the marginal distribution is known. For example, the joint distribution of age and gender could be extended to include education via the same procedure.

After generating synthetic joint probabilities via MrsP, the estimation procedure is identical to estimation with MrP. That is, sub-national means are calculated as the population-weighted mean of the model predictions for each post-stratification cross-classification. Accordingly, substituting MrsP estimation for MrP estimation is relatively easy and requires little additional domain expertise.

However, a notable short-coming of simple MrsP estimation is that the estimated joint distribution will only be correct when the auxiliary variables are independent. As they diverge from independence, the synthetic joint distribution becomes a less accurate (Leemann & Wasserfallen, 2017). Given that complete independence is rarely observed, this means that simple synthetic joint estimates will almost always be wrong. However, Leemann & Wasserfallen (2017) find that errors in the synthetic joint distribution do not necessarily induce errors in post-stratified, sub-national estimates.

Specifically, sub-national estimates remain constant, regardless of the synthetic joint distribution, as long as the auxiliary variables are modeled with constant marginal effects. Under these conditions, MrP and simple MrsP yield identical sub-national estimates, even if the synthetic joint distribution estimated in MrsP differs substantially from the true joint distribution used in MrP. However, a biased synthetic joint distribution will yield divergent MrsP estimates when the condition of constant marginal effects is violated.

Thus, for instance, Leemann & Wasserfallen (2017) note that MrsP estimates derived from a probit or logistic response model—which both have non-constant marginal effects—will diverge from MrP estimates. However, even with complete dependence among auxiliary variables, the magnitude of this divergence is generally negligible. Importantly, divergences can also occur with linear regression response models, but only when the marginal effects of the auxiliary variables are non-constant, which occurs when interaction effects are estimated. In this case, the degree of divergence between MrsP and MrP will vary as a function of the magnitude of the interactions.

Adjusted MrsP

However, this *does not* mean that researchers should avoid estimating interactions among auxiliary variables, as ignoring a meaningful interaction will also inhibit estimation accuracy. Instead, when auxiliary variables are correlated and their marginal effects are non-constant, researchers can either accept that some degree of bias will affect simple MrsP’s sub-national estimates, or they can rely on the second approach to estimating synthetic joint distributions, which employs an adjustment procedure to refine the synthetic joint distribution.

The goal of this adjustment procedure is to encode any available knowledge about the true joint distribution in the synthetic joint distribution. For example, while the joint distribution of a set of auxiliary variables may not be known at the *county*-level—thus making MrsP necessary for county-level estimation—in many cases it can be estimated at the national or even state level. Such higher-level estimates of the joint distribution can then be used as a baseline or template for estimating the synthetic joint distribution for each sub-national unit.

To generate an adjusted synthetic joint distribution for a given level of sub-national analysis, the following data is requisite:

(*Appendix continues*)

1. First, the upper-level, auxiliary variable cross-classification population counts $N_{u,j}$ must be gathered or estimated, where j indexes the cross-classifications of the auxiliary variables and u indexes the upper-level units in $u = 1, \ldots, U$. In some cases, this joint distribution may be available via census data; however, surveys can also be used to estimate it. For example, a nationally representative survey could be used to estimate the joint distribution for a set of auxiliary variables, such as gender, age, and education. However, the quality of adjusted MrsP estimates will depend on the accuracy of the auxiliary correlations encoded in this data. Accordingly, researchers must be careful in deciding whether a given survey is sufficiently reliable, as using unreliable data to adjust MrsP can produce estimates that are inferior to simple MrsP.

2. Next, at the targeted lower level of analysis, if the cross-classification population proportion of any subset of the auxiliary variables is available, this should be obtained. We represent this subset of the joint distribution as $P_{u[l],k}$, where $u[l]$ indexes the lower-level units $l = 1, \ldots, L$ in upper-level unit u and k indexes the cross-classifications of the auxiliary variable subset. For example, the county-level joint distribution of age × gender × education could be the desired joint distribution; but perhaps only the county-level joint distribution of age × gender is available. In this case, the county-level joint distribution of age × gender $P_{u[l],k}$—which represents the proportion population proportion of people who fall into age × gender cross-classification k for county l, which is nested in upper-level unit u—should be obtained. If no joint distribution data is available for the targeted level of analysis, $P_{u[l],k}$ will be a marginal distribution for one of the variables in the desired joint distribution. For example, if the county-level joint distribution of age × gender is not available, $P_{u[l],k}$ can be the marginal distribution of age or gender in unit $u[l]$.

3. Finally, at the targeted sub-national level of analysis, the *marginal* population proportions $P_{u[l],m}$ for each variable m not represented in $P_{u[l],k}$ must be obtained. Thus, if age × gender × education is the desired joint distribution, but only the joint distribution of age × gender is available, the *marginal* population proportions of education must be obtained in order to construct the synthetic joint distribution.

Given these data, several things can be treated as known. First, the overall or general relationship among the auxiliary variables is encoded in $N_{u,j}$. Second, if $P_{u[l],k}$ is available, the sub-national joint distribution for some subset of the auxiliary variables is known. Finally, for those variables not included in subset in $P_{u[l],k}$, the distributions $P_{u[l],m}$ tell us their marginal population proportions at the targeted sub-national level. The purpose of adjusted MrsP is then to generate a sub-national synthetic joint distribution $P^{MrsPA}_{u[l],j}$ that accounts for the correlational information encoded in $N_{u,j}$.

To accomplish this, $P_{u[l],m}$ is first used to adjust the marginal distribution of m, the variable being added to the joint distribution, in $N_{u,j}$ for each unit $u[l]$ so that it matches $P_{u[l],m}$. In our ongoing example, this means that the marginal distribution of education in $N_{u,j}$ is adjusted to match the known marginal distribution of education in unit $u[l]$, $P_{u[l],m}$. This adjustment is accomplished by transforming $P_{u,j}$ with a correction factor:

$$cf_{u[l]m[i]} = \frac{P_{u[l]m[i]}}{P_{um[i]}} \qquad (7)$$

where $P_{u[l],m[i]}$ is the true marginal population proportion of people that fall into level i of variable m within sub-national unit $u[l]$ and $P_{u,m[i]}$ is an estimate of the same marginal proportion derived from $N_{u,j}$. For example, if $m = \textit{education}$, $P_{u,m[i]}$ is the proportion of people with education level i observed in the national-level data; and $P_{u[l],m[i]}$ is the proportion of people with that education level in sub-national unit $u[l]$. Thus, $cf_{u[l],m[i]}$ is simply the ratio of the true proportion of people with $\textit{education} = i$ to the proportion estimated from the national data.

$N_{u,j}$ is then transformed as follows:

$$N^{adj}_{u[l],k,m[i]} = N_{u,k,m[i]} \times cf_{u[l],m[i]} \qquad (8)$$

where $cf_{u[l],m[i]}$ is the correction factor for level i of variable m, $N_{u,k,m[i]}$ is the set of cross-classification population counts in unit $u[l]$ for which $m = i$, and $N^{adj}_{u[l],k,m[i]}$ is the cross-classification population counts for unit $u[l]$ that have been adjusted so that the margin of $m[i]$ is the same as the observed margin in $P_{u[l],m[i]}$. This means that, in our example, $N^{adj}_{u[l],k,m[i]}$ is the adjusted population count for the cross-classification of age × gender × education in unit $u[l]$; the adjustment ensures that the marginal distribution of education matches the known marginal distribution of education in county $u[l]$.

Finally, the adjusted synthetic joint distribution is generated by using $N^{adj}_{u[l],k,m[i]}$ to extend $P_{u[l],k}$, the known cross-classification population proportions:

$$P^{MrsPA}_{u[l],j} = P_{u[l],k} \times \frac{N^{adj}_{u[l],k,m[i]}}{\sum_{i=1}^{I} N^{adj}_{u[l],k,m[i]}} \qquad (9)$$

where the second right-hand term is the relative weight of m's levels for each cross-classification of the k, the subset of auxiliary variables for which the joint distribution is known. Specifically, the numerator is simply the adjusted population count of people who fall into cross-classification j (i.e. cross-classification $k, m = i$) in sub-national unit $u[l]$; and the denominator is the adjusted population count of people who fall into cross-classification j, summed across each level of m. This yields $P^{MrsPA}_{u[l],j}$, the estimated proportion of people in sub-national unit $u[l]$ who fall into cross-classification j.

(*Appendix continues*)

For example, the numerator of the second term in EQ 9 represents the adjusted population count for the cross-classification gender × age × education. The denominator, on the other hand, represents the adjusted population count of the cross-classification of gender × age summed across each level of education. $P_{u[l],k}$ represents the proportion of people in unit $u[l]$ who fall into cross-classification k of age × gender. Finally, $P^{MrsPA}_{u[l],j}$ represents the estimated proportion of people in unit $u[l]$ who fall into cross-classification j of age × gender × education. Accordingly, in contrast to simple MrsP estimates, adjusted MrsP estimates are enhanced by information about the correlational relationship between the auxiliary variables. To generate estimates for an outcome Y in unit $u[l]$, $P^{MrsPA}_{u[l],j}$ can be substituted for $P_u[s[c]], j$ in EQ 3.

This procedure can be repeated multiple times to further extend $P^{MrsPA}_{u[l],j}$. For example, it could be applied to the adjusted synthetic joint distribution of age × gender × education and the marginal distribution of another variable in order to produce a 4-dimensional joint distribution. However, it some instances there may be no available data on the marginal distribution for a desired variable. In such cases, it is still possible to estimate a synthetic joint distribution. However, rather than relying on marginal distributions to implement adjustments, the marginal distribution for the target variable can be estimated via the full adjusted MrsP procedure. That is, a multinomial response model estimating the proportions of the target model can be used to make predictions for each level of the variable and then these predictions can be post-stratified, yielding an estimated marginal distribution for the target variable (Claassen & Traunmüller, 2018; Kastellec et al., 2015). Then, adjusted MrsP can proceed as above. However, while this approach is feasible, its benefit should be weighted against its cost: if a reasonable response model cannot be estimated, then the predicted marginal distribution will be inaccurate and this will negatively effect the accuracy of the final estimation procedure. Accordingly, researchers should carefully evaluate the importance of including a particular variable, keeping in mind that estimation of that variable's marginal distribution may be biased.

Received October 12, 2018
Revision received July 19, 2019
Accepted July 19, 2019 ■