access/download;Manuscript;KennedyGelman_manuscript.pdf

# Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample

Lauren Kennedy*
Econometrics and Business Statistics
Monash University

Andrew Gelman
Department of Statistics and Department of Political Science,
Columbia University, New York

Abstract

Psychology research often focuses on interactions, and this has deep implications for inference from non-representative samples. For the goal of estimating average treatment effects, we propose to fit a model allowing treatment to interact with background variables and then average over the distribution of these variables in the population. This can be seen as an extension of multilevel regression and poststratification (MRP), a method used in political science and other areas of survey research, where researchers wish to generalize from a sparse and possibly non-representative sample to the general population. In this paper, we discuss areas where this method can be used in the psychological sciences. We use our method to estimate the norming distribution for the Big Five Personality Scale using open source data. We argue that large open data sources like this and other collaborative data sources can potentially be combined with MRP to help resolve current challenges of generalizability and replication in psychology.

*Keywords:* Bayesian statistics, generalization, multilevel models, post-stratification, surveys

---

We thank the U.S. Office of Naval Research, National Science Foundation, and Institute of Education Sciences for their partial support of this work. We also thank Jessica O'Rielly and Matthijs Vákár for their thoughtful comments on this manuscript.
Preprints of this work have been posted on arxiv.org, and ideas around this work have been presented at the Australian Mathematical Psychology meeting (2019) and the 2019 Society of Experimental Social Psychology.
Corresponding author. Email: Lauren.Kennedy1@monash.com

Psychology is all about people, and because people are so wonderfully heterogeneous, generalizing psychology research has to be all about interactions. Not even randomization can save us from heterogeneity. Even in studies that only claim to investigate whether an effect "exists," the expected heterogeneity of the effect should be considered when interpreting the results. At the same time, some studies are concerned with effects that hold in some broader population. If our sample isn't representative (as many psychological research samples are not), and the effect is heterogeneous, how can we estimate this effect in the population?

This challenge is not new to statisticians. Traditionally, survey weighting has been employed to account for differences between sample and intended population from design and nonresponse. However, it is uncommon for participants in a psychology experiment to be chosen using any formal sampling design. Convenience samples (or non-probability samples) dominate the field, which makes it difficult to construct classic design-based weights. Even a random sample from a known population is unlikely to be truly random because of nonresponse patterns. Without random sampling, this problem grows even more difficult. Often in psychology we rely on convenience samples, such as first year undergraduates, kind community members, or (more recently) Amazon Mechanical Turk workers and other crowdsourcing alternatives. These convenience samples rarely represent the population that we are interested in, and they can differ in important ways from underlying populations of interest.

Throughout this paper we argue that the statistical technique known as multilevel regression and poststratification (MRP; Gelman and Little, 1997; Little, 1993; Park, Gelman, and Bafumi, 2004) could be applied to convenience samples in psychology. This method allows the researcher to infer quantities in the *population* from a sparse and possibly nonrepresentative *sample*, combining two ideas in the survey research literature: small-area estimation and nonresponse adjustment. MRP is popular and is a continuing subject of research within the political science literature (see, for example, Ghitza and Gelman, 2013; Lax and Phillips, 2009b; Si, Trangucci, Gabry, and Gelman, 2017) and has also been introduced in public health applications (Downes et al., 2018). Wang, Rothschild, Goel, and Gelman (2015) demonstrate the effectiveness of MRP for a large non-probability political poll.

In psychology we are well trained in experimental design. In this paper we are not discussing an alternative to randomization, nor are we considering the challenge of generalizing to new experimental conditions not in the existing study. Instead we focus on generalizing the effect of an experimental intervention would have if it were applied to a wider population beyond people in the sample. While this is a relatively uncommon adjustment within psychology, examples within political science demonstrate the importance of this adjustment in survey experiments (Green and Kern, 2012). One reason for this is that the population of interest (e.g., voters or the general adult population) is more clearly defined in political applications. The other is that political science tends to use design based surveys more frequently.

To extrapolate from sample to population we make two types of assumptions. First, we make statistical model assumptions in terms of variables included, priors (if any) used, and the type or form of the model. In particular, if we are interested in extrapolations of treatment effects, it is important to include interactions between the treatment and the

person-level variables that capture key differences between sample and population. Second we make assumptions of equivalence—that the people unobserved are the same as the observed once we have adjusted sufficiently. If we adjust on age group and gender, equivalence means that people within a specific age x gender group would have the same expected difference given an intervention (with some random variation).

MRP and other survey adjustments are not widely used to analyze experimental data in psychology. Randomization of treatment assignment is thought to allow us to estimate the average treatment effect. However, in the presence of interactions between demographic characteristics and the quantities of interest (which are typically the object of study in psychology experiments), the average treatment effect is uninterpretable without reference to a population, hence adjustment for non-representativeness of the sample again becomes necessary. Even in experiments that are only concerned with whether an effect exists (the "what can" argument of Mook, 1983), heterogeneity can explain when a study doesn't find an effect (even when there is one for some groups), or why an effect didn't replicate after being observed once, a point highlighted by Henry (2008).

To encourage the intuition behind this, we recommend that readers reflect on known or suspected moderators of an intervention effect. Some moderators are reflective of various decisions that can be made regarding the experimental design (such as the duration a stimulus is displayed), but others can be attributed to person-wise heterogeneity (such as the socialization of different genders). Our concern is on the latter - if there is between person moderators present or suspected, then differences between the sample and the population can be of concern.

In explaining how MRP has potential to be useful for generalizing research findings in psychology, we first discuss in high-level language what we mean by multilevel modeling and poststratification, and how the two methods combine to be such a useful tool. Then we describe some caveats with MRP, before using an open data set measuring scores on the Big Five personality scale to demonstrate an application of MRP. We also further demonstrate the idea of moderation and randomized control trials with a simulated example. We conclude with a discussion of the limitations of the method and active research currently being conducted in this area.

## What is MRP?

Multilevel regression and poststratification combines two statistical techniques to (a) quantify the relationship between some outcome variable of interest and a number of predictors, and (b) obtain generalizable inferences by adjusting for known discrepancies between sample and population. Similar approaches use alternative models with poststratification (e.g., Caughey & Warshaw, 2019; Yuxiang, Kennedy, Simpson, & Gelman, 2019). The important point is that the model uses some sort of regularization or partial pooling to obtain stable estimates from relatively small samples. Here we focus on mixed effects models, as one of the most familiar technique to psychologists to explain how this regularization works.

A mixed effects model is similar to a traditional regression (where some outcome variable $y$ is modeled as a function of a set of predictors $x_1, x_2, x_3, \ldots, x_m$), but a mixed effects model breaks these predictors into two sets; constant and varying effects. We avoid the terms "fixed" and "random" here because they are given different meanings in different contexts; see Gelman (2005).

In the case of MRP, the technique advocates for using varying effects for person-descriptive predictors such as education, race/ethnicity, state, and age group that take on multiple levels in the data. We do not restrict them to be used for multiple observations per individual (as in the traditional use of multilevel models in psychology). We demonstrate how multilevel modeling differs from classical least-squares regression or ANOVA with a simple hypothetical example. For a more detailed description, we recommend Sorensen, Hohenstein, and Vasishth (2016). We will also build on the notation of Gelman and Hill (2007), which is commonly used in the MRP literature. For reader ease, we begin with a hypothetical example where multilevel models have often been used. Say you have test scores from a sample of students each belonging to one of $K$ schools, and you are interested in predicting scores $y$ from school $k$. How is multilevel regression, with varying intercepts for school, different from least-squares regression including school indicators?

The classic model setup for including school effects would be to create $K$ binary variables, denoted $d_k$. (An alternative parameterization is to create $K-1$ indicator variables with the $K$th replaced by the intercept. We formulate the model with $K$ predictors as it allows easier formulation as a varying effect.) Each variable indicates whether the student belongs to school $k$. We could then fit the following non-multilevel model:

$$\begin{gathered} y = \beta_1 * d_1 + \beta_2 * d_2 + \beta_3 * d_3 + \cdots + \beta_K * d_K + \epsilon, \\ \epsilon \sim \text{normal}(0, \sigma_y) \end{gathered} \tag{1}$$

If a student is in school 7, for example, then $d_7 = 1$ and all other $d_{k;k\neq 7} = 0$. This means that the above equation would simplify to:

$$y = \beta_7 + \epsilon, \tag{2}$$

for which the estimate would simply be the mean of school 7.

In multilevel regression, we would model the intercept for the schools as $\beta_k, k = 1, \ldots, K$, and then apply a probabilistic or 'soft' constraint to the set of $\beta_k$'s such that they are distributed with mean $\mu$ and variance $\sigma$.

$$y = \beta_k + \epsilon. \tag{3}$$

$$\beta_k \sim \text{normal}(\mu_k, \sigma_\beta). \tag{4}$$

The two models are similar in that each school is modeled as having a different mean level of scholastic ability. The difference is the amount of information that is shared between the levels. In the first formulation, the test scores in each school are modeled independently of other schools. In the second formulation, the test score component for each school uses information from observed test scores at other schools. With multilevel modeling, the amount of shared information forms a continuum, ranging from no pooling (Equation (1)) to full pooling, which would correspond to a model with an intercept that is the same for each school (equation below). Gelman and Pardoe (2006) describe this continuum more formally.

$$y = \beta_{int} + \epsilon. \tag{5}$$

Multilevel modeling allows us to fit the amount of pooling (through the size of $\sigma_\beta$) with the other parameters in the model. The amount of pooling is also akin to the amount of regularization. More pooling indicates more regularization, less pooling indicates less regularization. Moreover, it provides an avenue to make predictions about new populations or samples. One example of this is in Weber et al. (2018). To do this we might need to use a strong prior about the relationship of the observed sample to the sample or population that we would like to generalize to.

This leads us to the second component of MRP, poststratification. For the school example, we would need a poststratification table that contains the total number of students in each school. We would use the formula obtained from the regression analysis to predict the test scores for each school. To obtain an estimate for the total population of students, we would multiply each school estimate by the number of students in that school, add these all up, and then divide by the total number of students in the population. Mathematically if the school estimate for the $k^{th}$ school is referred to as $\theta_k$, this would be expressed as,

$$\theta_{\texttt{POP}} = \frac{\sum_{k \in K} N_k \theta_k}{\sum_{k \in K} N_k}.$$

The steps of MRP are as follows:

1. Measure key demographic features in sample during survey collection.
2. Identify the poststratification table: estimate population counts for each possible combination of these demographic features (each combination is a cell in the table).
3. Measure some key quantity in the sample. This is what you would like to estimate in the population.
4. To estimate this quantity of interest in the population, use multilevel modeling to predict this quantity using the observed demographic features in the sample.
5. Estimate the outcome variable in each cell of the poststratification table.
6. Aggregate over cells of the poststratification cells (using the cell size) to obtain population level estimates.

**Conditions of data necessary for generalization**

Not all situations are suitable for generalization through MRP. The method is designed to be used in examples where we expect heterogeneity, the heterogeneity is expected to interact with the outcome or manipulation, and the sample is not representative of the population. If all three conditions are met, then there is potential for this approach to be beneficial. Even when we do have an example that meets these conditions, the data that we have collected might not be sufficient to model it in this way. First, to model heterogeneity, the data actually need to contain heterogeneity. If the sample is an undergraduate population and the effect is expected to differ between young adults and the elderly, then this method will not be appropriate to estimate a population effect but might be appropriate to estimate an undergraduate effect. If the sample is an undergraduate population and we don't expect there to be heterogeneity across age or previous research demonstrates there

is not, then perhaps the sample is suitable. This is discussed partially in Smith and Little (2018) in relation to small-N vision studies.

Although this might seem counterintuitive, in practice no sample, even one obtained by random sampling, can be truly representative on all possible covariates. Although MRP traditionally follows the survey weighting literature to adjust for individual demographics and sampling design; different sets of adjustment variables might be more appropriate when generalizing in psychology. This raises challenges because these adjustment variables may not be known in the population. It is our hope that by applying MRP to fields like psychology there will also give an opportunity for the field of survey research to improve the methods currently in place for non-probability studies.

In survey research the term "nationally representative" is used to mean a sample that was drawn from a frame that covered the entirety of (generally the US) a country. To be representative more generally is difficult because non-response patterns and frame issues mean that a sample is rarely (if ever) truly representative. Sometimes a weighted sample is said to be "adjusted to representative" of a specific population by several important covariates (that generally represent heterogeneity or historic systematic exclusion in surveys). The covariates differ by country and by survey context. While a number of variables collected in large surveys like the American Community Survey (education, race/ethnicity, gender and age) seem like promising adjustment variables, psychology will likely need to explore and reflect on what factors will determine whether a sample is sufficiently representative.

Another necessary condition for MRP is sufficient data to fit these hierarchical models. It's difficult to give broad recommendations for the required sample size, but we doubt that much would be gained if this method were used with a between-person design with a small sample such as from 50 individuals. If MRP were to be used with this sample, we expect that the estimates for new groups would be very uncertain, or else inferences would depend strongly on any priors used.

In addition, we recognize that not all research in psychology is about estimating population level effects. In some fields the research question is simply whether an effect is observed or not. In this case MRP is not directly relevant. That said, if effects are heterogeneous, then not observing an effect in a given study does not mean that this effect doesn't exist in the wider population. Similarly, a "statistically significant" effect could be observed in a particular sample but still be difficult to replicate in different samples.

## What should we adjust for?

In political science the variables that we expect to adjust for are fairly consistent. Basic demographics such as age, sex, race/ethnicity, education, along with geographic factors such as state and urban/rural/suburban classification. But what adjustment variables should we use in psychology applications? Basic demographics should be a good starting point, but applying this method in any particular example would benefit from understanding where heterogeniety is expected in terms of the impact of a given intervention.

This knowledge is typically ingrained in the expertise of the researchers conducting a study. For example, Sears (1986) needed to have a good understanding of the college student population and the social phenomena he was studying to discuss the implications of a college student sample on the phenomena. However, recently Simons, Shoda, and Lindsay (2017) advocate for the inclusion of a constraints of generality (COG) statement in

all empirical work. We believe that such a statement will help the researcher communicate knowledge about expected heterogeneity of effect and provide clues of what to adjust for. Indeed, in their paper they argued that such a statement will make clear where findings are expected to replicate, and encourage other researchers to explore outside of the proposed boundary conditions.

The constraints on generality statement provides necessary information to move beyond a general statement about replicability to a statistical approach leading to quantitative conclusions (and, as appropriate, large uncertainties) about particular replications or generalizations of interest. Additionally, the COG statement has been subjected to peer review. Although peer review is not infallible, it does provide some suggestion the COG statement reflects the knowledge and experience of researchers in the field. We discuss the importance of this later in this article.

## Incentives to use MRP

Having discussed how MRP is formalized, in this section we discuss why such a technique can be so useful. The use of MRP aligns with the contents of an effective COG statement. That is, MRP is useful when we as researchers have formalized a population of interest, have identified key variables that are believed or theorized to impact the outcome variable of interest, and have distinguished differences between the sample and the population using these key variables. By our interpretation of Simons et al. (2017)'s paper, the COG provides a structure to do exactly that. Unsurprisingly given the close relationship between the two, the incentives of MRP mirror those of the COG.

The COG statement gives the researcher an avenue to consider the population of interest and to consider whether findings from the sample should generalize to the population as a whole. MRP uses population level information to *directly* estimate the variable of interest in said population, including an estimate of uncertainty. Without the COG, which involves the researcher considering what population they hope to generalize to, and the differences between the sample and population that might impact generalizations from the sample to the population, we would not know which variables to include in our multilevel model, nor would we be able to define the population well enough to generalize to it. The COG statement provides this information, so MRP can build upon it to infer quantities in the population.

Likewise, while the COG statement provides some basis for the researcher to guess how likely it is that the findings will replicate based on differences between previous sample populations and the current sample, MRP provides a way to estimate the variable of interest in a new sample directly. While dissonance between the MRP estimate and the observed value in a new sample doesn't immediately signal a failure to replicate, it does provide a tool for further research to explore whether there are additional differences between the two samples that might cause failure to replicate.

Lastly while the COG statement uses researcher intuition and domain-specific knowledge of the field, the multilevel component part of MRP provides an avenue to test and quantify these beliefs. While it might be intuitive that a specific demographic variable might be related, multilevel regression helps to quantify the size of the relationship, leading to better population level predictions.

By conducting MRP and finding heterogeneity between demographic subgroups, we can also find inspiration for future research. For example, say we wish to estimate mathematical reasoning in the population and find that in our sample gender is a good predictor. The current research project might poststratify using gender to obtain population level estimates of mathematical reasoning, while future research might focus on exploring this relationship in more depth.

All of these incentives require the researcher to be able to formally state and describe the population that they wish to generalize to, which to us is one of the biggest benefits of the COG statement. In the next section we consider how the COG and MRP might work together in practice.

## Example 1: Tutorial with real data

In this section we use an open source dataset to demonstrate the mechanics of an MRP approach. Say you are developing some new scale (such as a personality scale) for use in the general public. After validating that it measures what it is intended to, your next step would be to estimate the distribution of this scale in the general population so that an individual's score can be meaningful relative to the greater population.

To see how this works in practice, we apply this technique to a large database of responses to a 50 item IPIP (Goldberg et al., 1999; Goldberg et al., 2006) version of Goldberg (1992)'s 5 factor model of personality collected through the Open Source Psychometrics Project, 2019. The full scale and scoring is described as an example scale on the International Personality Item Pool Project (n.d.) website. This scale measures five facets of personality; Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each subscale is measured with 10 items, each measured on a likert scale with items scored from 1 to 5 (with some items scored in reverse) so that total scores on each subscale range from 10 to 50. This tutorial is accompanied with a Rmarkdown document (for Openness subscale) and .R file (for all subscales), found in the supplementary materials.

This dataset contains a convenience sample of 19 719 individuals who completed the scale online. Following the scale, participants also were asked to provide basic demographic information, with location information derived through technical information, which we used to subset to US participants specifically. A total of 8 665 US participants provided all of the requested information.

One way to interpret the Big Five is to compare an individual's score in relation to the wider population distribution. To do this, we need a distribution of scores on a representative sample. This is particularly important as there are substantial individual differences in personality scores, for example across age (Donnellan & Lucas, 2008) and gender (Weisberg, DeYoung, & Hirsh, 2011).

A convenience sample is rarely representative. In this particular case, our sample was much less like to be male ($M = 34\%$ or 2 939) when compared to the wider US population as estimated from the 2012 ACS ($M = 49\%$). The proportion of the sample aged 13–25 is 60% in our sample but only 20% in the ACS. Although we do not know the who decided to participate in this study, we can guess from the other demographics (predominantly young women) that at least some portion were undergraduate psychology students. Ideally we would also be able to adjust for education level of our sample but this covariate is not available in our dataset. This is a major limitation of this analysis as it means we are

assuming that either our sample does not differ from the population on education level, or that education level is not related to the Big Five. We are also limited in that the survey we are using does not measure race using the same categories as those in the American census. Ideally we would create a mapping between the two measurements, but for the purposes of this tutorial we can still adjust the sample distribution of each of the facets of the Big Five on gender and age group. The accompanying Rmarkdown document demonstrates this analysis for the Openness subscale, which is repeated for the four other subscales in the included R script.

**Step 1: Model the outcome in terms of the adjustment variables**

After downloading the data and reverse coding the necessary items, we sum each of the 10 items to get a total score on each subscale of the Big Five. These subscales are the outcomes that we would like to estimate in the population. To do this we need to fit a multilevel model with age and sex as the adjustment variables (predictors). The dataset measures gender (male, female, or other), while the ACS measures sex (male or female). For simplicity, we remove all cases where gender is not stated as male or female. We hope future research will work on more appropriate ways to poststratify gender to the census. Age is broken into six uneven categories; under 18 ($N = 1\,903$), 18–24 ($N = 3\,285$), 25–34 ($N = 1\,507$), 35–44 ($N = 847$), 45–65 ($N = 890$), and 65+ ($N = 233$). The dataset also measures race/ethnicity, but does not use categories that map easily to those used in the US census so we do not adjust by race/ethnicity.

For each outcome (O, C, E, A, and N) we fit a model in R using brms (Bürkner, 2017), a package that allows the user to fit fully Bayesian models using standard R formula notation and with enough flexibilty that our model can account for truncation of the outcome variable between 10 and 50. It is possible to perform multilevel modeling without being fully Bayesian, but we find that a Bayesian approach is natural, especially for accounting for different sources of inferential uncertainty when making predictions.

For the purpose of readability, we describe the process using one outcome variable—scores on the Openness subscale—but adjust all five subscales in the accompanying code. With the following code we fit a regression model with upper and lower bounds (`ub=50` and `lb=10`) with `O` as the outcome variable, gender as an indicator for female, and `age_group` as a varying effect. We specify the data as `data_us`. The remaining input specifies computational details, namely that 4 chains of MCMC will be run, that there are 4 available cores, and the step size (`adapt_delta`) that should be used. More on these control settings can be found at https://mc-stan.org/misc/warnings.html. There may be a small number (<5) divergent transitions when running this model, which in this case can be solved with an adapt_delta closer to 1.

```
m_O <- brm(O | trunc(lb=10, ub = 50) ~ female + (1|age_group),
  data=data_us, chains=4, cores=4, control=list(adapt_delta=.80))
```

In mathematical notation this can be written as

$$y_i \sim \text{normal}(\beta_0 + \beta_{\text{male}} X_{\text{female[i]}} + \alpha_{age[i]}, \sigma) \tag{6}$$

$$\alpha_{age} \sim \text{normal}(0, \sigma_{age}) \tag{7}$$

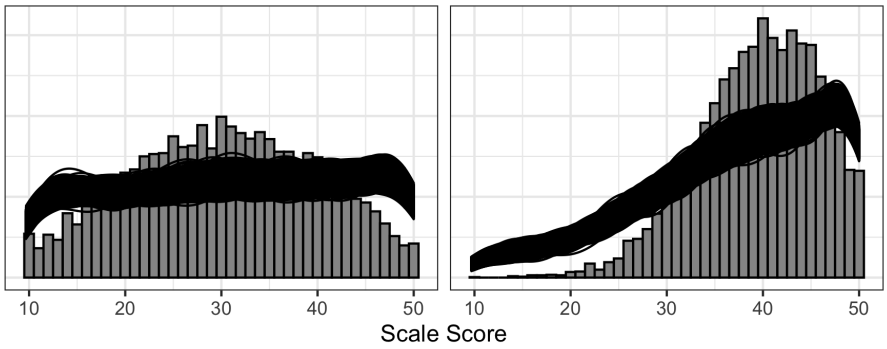


*Figure 1*. Observed sample histogram for each of the five personality subscales. For each we display the posterior estimates for the sample (black lines) and population (red), which give an indication of uncertainty of our estimates.

One important feature of the Bayesian workflow is the selection of priors. By default brms normalizes and rescales the data and sets priors that reflect this transformation. We can change the default prior choices using the `prior` argument to the `brm` call.

One way of understanding the choice of priors is using prior predictive checks (Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019). Using an additional argument we can sample from the prior only. One thing to note is that the default prior in brms for a $\beta$ parameter is unconstrained, which is difficult to sample from. We specify a wide $N(0, 10)$ prior to enable us to do a prior predictive check, but to not advocate for this prior necessarily in all situations.

```
    m_O <- brm(O | trunc(lb=10, ub = 50) ~ female + (1|age_group),
      data=data_us, chains=4, cores=4, control=list(adapt_delta=.99),
      prior = set_prior("normal(0,10)", class = "b"), sample_prior =
"only")
```

In Figure 1 we plot the expected distribution for the Openness and Extraversion given the model. These priors are not updated by the data, but because they are created relative to normalized data, they have different effects given different outcomes. For a more thorough description of Bayesian workflow in psychological examples, refer to Schad, Betancourt, and Vasishth (2019).

This model fits well with no warnings. The focus of this manuscript is not on how to test good fitting in Bayesian models so we do not discuss this further here. We direct readers towards Gabry et al. (2019) for more tools on effective model checking and diagnostics. The takeaway from this step is that we have used our sample to fit an estimate of O scores for different gender and age groups. We plot the estimate for each in Table 1. Other models could also have been used for this.

**Step 2: Adjust the sample to the population**

Next we need an estimate for the population distribution of the adjustment variables, in this case age and gender. We would like to generalize to the population of U.S. residents

Table 1
*Parameter estimates for the intercept if male, females, and six different age categories.*

| | Posterior mean | Posterior sd | Quantile 2.5 | Quantile 97.5 |
|---|---|---|---|---|
| intercept | 43.1 | 0.4 | 42.4 | 43.9 |
| female | -2.7 | 0.2 | -3.1 | -2.3 |
| <18 | -0.6 | 0.4 | -1.5 | 0.1 |
| 18-24 | -0.5 | 0.4 | -1.3 | 0.2 |
| 25-34 | 0.4 | 0.4 | -0.3 | 1.2 |
| 35-45 | 0.3 | 0.4 | -0.5 | 1.1 |
| 45-64 | 0.4 | 0.4 | -0.4 | 1.2 |
| 65+ | 0.0 | 0.5 | -0.9 | 1.0 |

Table 2
*Population counts of each combination of demographics as estimated using the ACS, where N is the number of Americans in that category.*

| female | age_group | N |
|---|---|---|
| 0 | 1 | 10 713 479 |
| 0 | 2 | 15 974 402 |
| 0 | 3 | 22 216 888 |
| 0 | 4 | 20 279 699 |
| 0 | 5 | 31 659 960 |
| 0 | 6 | 30 275 386 |
| 1 | 1 | 10 193 764 |
| 1 | 2 | 15 166 108 |
| 1 | 3 | 21 758 629 |
| 1 | 4 | 20 455 441 |
| 1 | 5 | 32 907 478 |
| 1 | 6 | 36 732 643 |

aged 13 and over (the youngest participant in the survey is aged 13). We get the population distribution of age × gender from the American Community Survey (Bureau, 2012, ACS), a large representative survey of the US, which we can use with the provided weights to approximate census level information. We use ACS estimates from 2012, the year when most of our sample data were collected.

After downloading and merging the files (the ACS is released in four datafiles), we subset down to the age and gender variables. Using the age variable, we create the same age categories as we used in the sample. We can then use the ACS survey weights to estimate the number of people in each combination of age group and gender. We use the package dplyr for this, and print the resulting poststratification matrix in Table 2.

After fitting this model for the openness score, we simulate a random sample of size 10 000 from the population, proportional to the estimated population cell sizes. Any sample size could be used, depending on the desired precision.

```
sample_pop <- sample(1:12, 10000,prob=acs_ps$N, replace=TRUE)
```

```
sample <- acs_ps[sample_pop,1:2]
```

We then use a function from the brms package to predict the Opennness scores we would have observed in this simulated sample. We use the following code to estimate five possible Openness scores for each person in the sample. More could be taken, but we do this get a sense of posterior variance.

```
PPC_O <- posterior_predict(m_O, newdata = sample)
```

We use a similar line of code to predict for the observed data to compare the predicted distributions of the model given the sample. In Figure 2 we plot the sample (histogram), sample estimates (black lines) and population estimates (red lines). There are multiple lines to represent each posterior predictive estimate, giving an indication of uncertainty of our estimates. We can see that this MRP adjustment makes a considerable adjustment for some subscales (such as conscientiousness and neuroticism), a small amount of difference for others (openness and agreeableness) and negligible difference for extraversion.

### Constraints on this analysis

Using multilevel regression and poststratification in this analysis we used a non-probability convenience sample to estimate the population distributions of different psychometric scales. In our analysis we adjusted for age and gender, noting that the sample differed considerably from the population on these two demographics. This analysis is limited in that we did not adjust for education (as it was not measured), and we are concerned that education could be related to various personality factors and (judging from the dominant age/gender of the sample) we suspect that many of the respondents where psychology undergraduates. We also did not adjust for race/ethnicity (which is a common adjustment variable in political science) because of substantial differences in measurement between the sample and the US census. Lastly we focused only on individuals who responded as either male or female due to data constraints in the census. The purpose of our analysis is to provide an open data tutorial and proof of conscience rather than using these curves as a gold standard going forward, but we encourage others who seek to use this method to consider these variables.

## Example 2: Simulated experimental data

Here we present a fictional but plausible example to extend this idea to experimental psychology. Say you would like to estimate impact of an intervention on maths anxiety.

For convenience, you ask students from your first year psychology class to take a survey that measures maths anxiety as well as a selection of demographic such as like age, gender, and major field of study. You also post flyers inviting participants from other faculties to participate, but your sample is not representative of the distribution of degree or gender at the university. Following the initial survey, participants are allocated to an intervention designed to reduce maths anxiety, or a control task. After they have completed the intervention, they are again scored using the maths anxiety survey.

Writing a COG statement you acknowledge that while all students were members of the population, the sample was *not* representative of the population of interest (the body of undergraduate students at your university). Furthermore, given that there might be

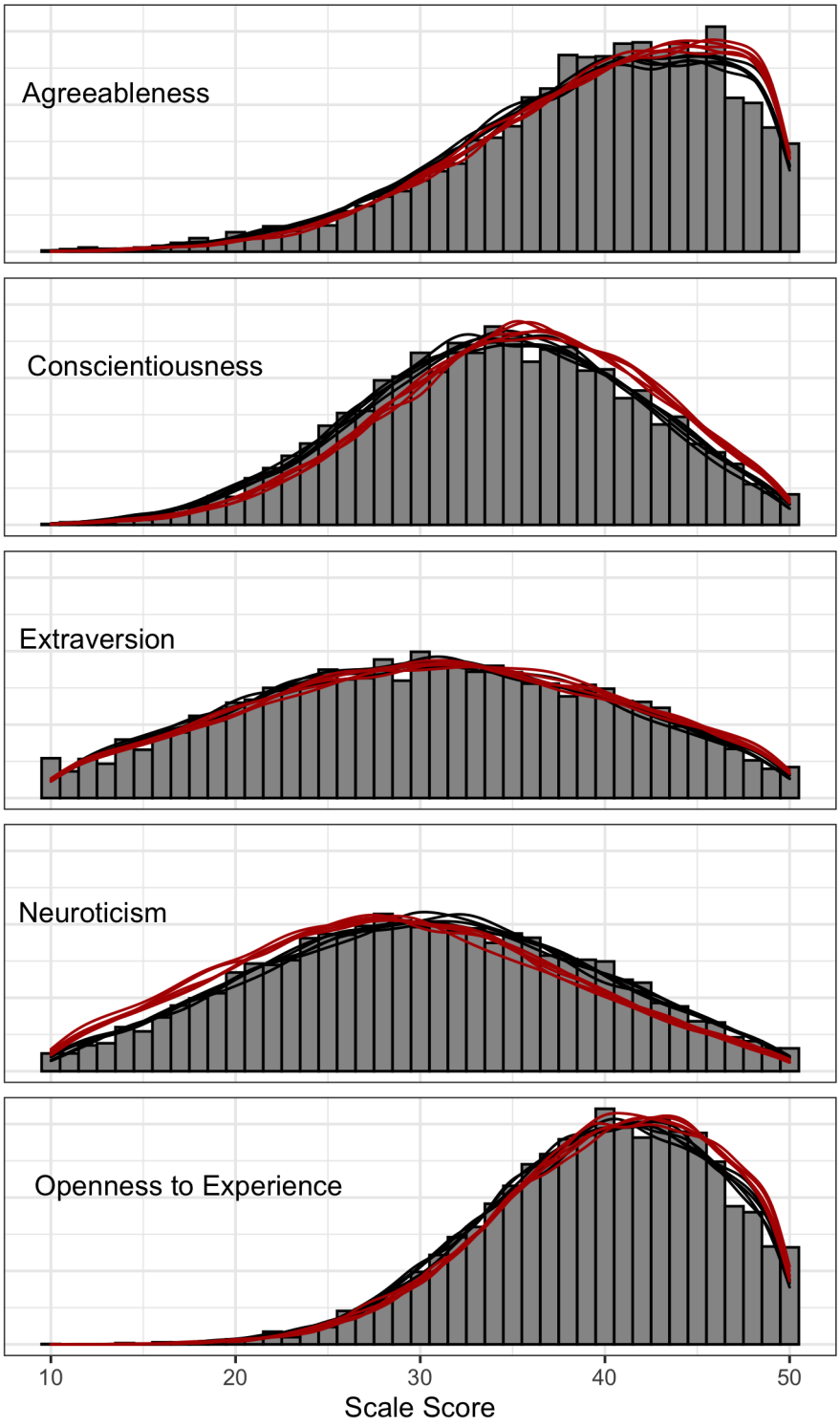


*Figure 2*. Observed sample histogram for each of the five personality subscales. For each we display the posterior estimates for the sample (black lines) and population (red), which give an indication of uncertainty of our estimates.

interactions of gender or major with maths anxiety (maths majors might be less likely to be maths anxious than a major like psychology), you declare that the total maths anxiety estimate from your sample might not be representative of the undergraduate population as a whole. In addition, you declare that gender or major might interact with the efficacy of the intervention, and so the estimate of the effect of the intervention might not be representative of the intervention's effects of the undergraduate population.

This is a similar requirement to considering moderators to an intervention effect. Here we focus on person specific moderators (such as gender). The reason we focus on moderators that are person specific attributes is because if this type of moderator exist, then we have to consider whether the sample is representative of the population on these moderators when interpreting the results of any statistical analysis.

There are multiple possible aims for generalizing this study. One aim might be to estimate the degree of maths anxiety that exists in the university. Another might be to estimate the impact of the maths anxiety intervention if it were implemented across the university. A third aim might be to replicate the study's results with a new sample from the same university. We address each of these aims in turn to explore some of the potential for MRP in psychology.

**Estimating maths anxiety in the university**

The COG provides the framework to identify key areas where the sample differs from the population and how this might impact the results. What it doesn't do is provide a way of estimating maths anxiety in the actual population of interest. MRP partnered with the COG statement, however, provides a way to estimate maths anxiety in the full undergraduate population from the sample, without additional data collection. The procedure to do so would be as follows:

1. Measure gender and degree major in the initial survey. In the simulated data that accompanies this tutorial we assume the initial sample is a generous n=300.

2. Obtain demographic data about the full population of undergraduate students at your university. This may or may not be easy, but we assume that undergraduate demographic data are published by or available from your university. Use the demographic data from the population to construct a poststratification table. This table counts how many people in each possible demographic category (i.e., the number of women studying for an engineering degree, the number of men studying for an economics degree, etc.). The table should look something like the following, with the N column summing to the total number in the undergraduate population. It should contain all possible combinations of the demographic categories, but some may be empty.

| Gender | Major | N |
|---|---|---|
| F | Engineering | 982 |
| F | Law | 1392 |
| ⋮ | ⋮ | ⋮ |
| M | Liberal Arts | 672 |
| M | Liberal Arts | 342 |
| ⋮ | ⋮ | ⋮ |

3. Identify the outcome variable that you're interested in; here, it is baseline maths anxiety at a university level.

4. Using the sample, create a multilevel model with the demographic variables (especially gender and major) as predictors and and maths anxiety as the outcome variable. In our case we provide simulated data assuming the maths anxiety scale ranges from 10 (low) to 50 (high) so that we can use the same priors in brms as before. The only slight difference is that because this is simulated data we can include move beyond a binary gender variable if the data show a need for such an analysis. We fit the model using

```
brm(mathsanxiety_t1 | trunc(lb=10, ub = 50) ~ (1|gender) +
  (1|major), data=dat_s1, chains=4, cores=4,
  control=list(adapt_delta=.99),
  prior = set_prior("normal(0,10)", class = "b"))
```

5. Use the model from 4 to predict the degree of maths anxiety using the poststratification table from 2. As with the first example, we do this for both the sample and the population, taking numerous posterior draws to compare the noise of the estimates. Unlike the previous example, we have much less data and so the estimates of distribution are much noisier. Also in contrast to the first example, our simulated population is much smaller as well, only 4222 in this university. This means we can predict pre-intervention maths anxiety for each individual in the population. We plot the sample and population in Figure 3.

6. Aggregate over these estimates using the size of the cell to estimate population or subpopulation values. In this case we calculate the average maths anxiety in the sample as M = 28.4 we estimate math anxiety in the university as M = 25.9. As this is a simulated dataset, we know the true mean of maths anxiety in that university is 26.7.

**Estimating the impact of the maths anxiety intervention in the university**

Often the primary aim of psychological research is not simply to estimate a quantity but instead to estimate the impact of an intervention or manipulation. To do so we often rely on random assignment to intervention and control groups. However, if the sample is different to the population, even randomization doesn't guarantee that the effect estimated in the sample will generalize to the wider population. Here we extend our MRP analysis

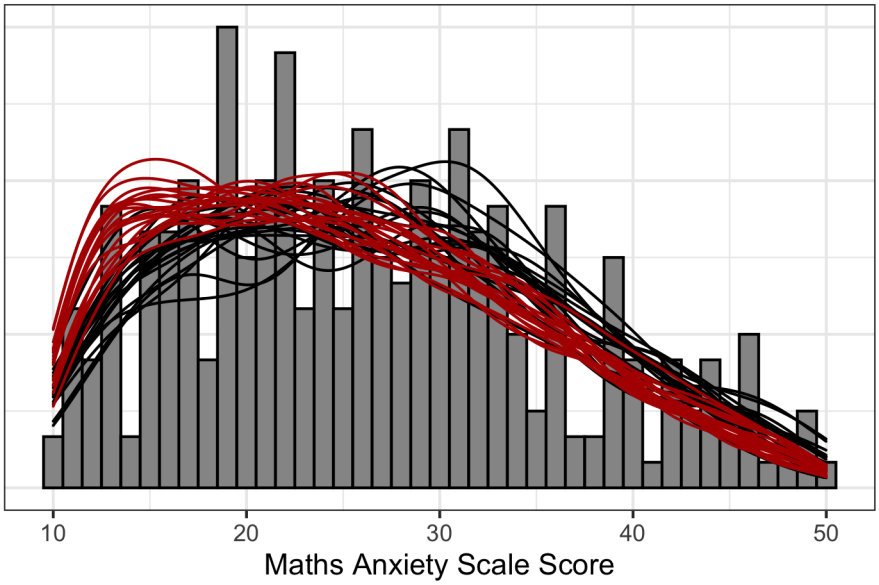


*Figure 3*. Observed sample histogram for the simulated sample maths anxiety scale score. For each we display multiple posterior estimates for the sample (black lines) and population (red), which give an indication of uncertainty of our estimates.

to include randomized control trials. Building off the analysis presented in the previous section, we start at step 3.

3. We now use the sample to predict the post-intervention (Z where Z=0 if control group and Z=1 if treated) maths anxiety given the pre-intervention anxiety level, gender and major of the participant. This means that we are simply changing item 3 in the previous item to be the difference between pre and post maths anxiety scores.

4. Now using the sample we model the difference between pre and post intervention maths anxiety for the control and intervention groups. If we were able to model maths anxiety using a linear model, then we would be able to model the before-afer difference directly as the difference between two normal distributions is normally distributed. As we are using a truncated regression to model maths anxiety, we instead model preintervention anxiety given gender and major

```
brm(mathsanxiety_t1 | trunc(lb=10, ub = 50) ~ (1|gender) +
  (1|major), data=dat_s1, chains=4, cores=4,
  control=list(adapt_delta=.99),
  prior = set_prior("normal(0,10)", class = "b"))
```

and then post treatment anxiety given gender and major and pre treatment anxiety

```
brm(mathsanxiety_t2 | trunc(lb=10, ub = 50) ~ mathsanxiety_t1 +
  (Z|gender) + (Z|major), data=dat_s1,
  chains=4, cores=4, control=list(adapt_delta=.99),
  prior = set_prior("normal(0,10)", class = "b"))
```

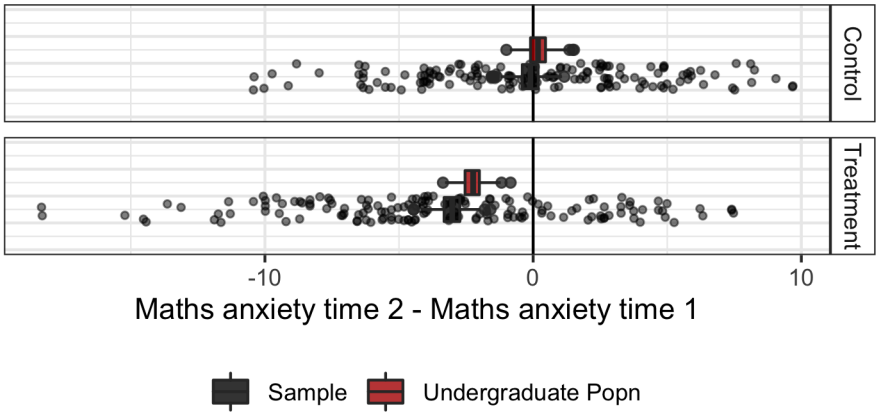


*Figure 4*. Expected post-pre difference in maths anxiety estimated in both the population (red) and the sample (black). Points represent observed differences in the sample.

5. Then we use the first model to predict the degree of pre-intervention maths anxiety for each undergraduate in the university (taking 20 posterior samples to maintain uncertainty) and then for each posterior predicted estimate for math anxiety before treatment we can predict the post intervention predicted estimate for math anxiety as if each individual was allocated to either treatment or control.

6. Using these two estimates, we can calculate the expected difference between pre and post maths anxiety given treatment and control intervention for both sample and population. We plot the estimated difference in both sample and undergraduate population in Figure 4. The mean post-pre difference in the sample is -3.36 for the intervention group and -0.47 for the control group. We estimate it in the undergraduate population as -4.66 for the intervention group and 0.11 for the control condition. As this is simulated data we also know the true effect in the population is -4.04 in the intervention group and 0.16 in the control.

**Generalize the impact of the maths anxiety intervention in a new sample in the university**

Using the COG statement to explicitly define the population in terms of several key demographic features provided us the opportunity to make estimates for the population. However, Simons et al. (2017) noted that the purpose of the COG statement was more than simply describing the population. It also provides an avenue for future researchers to estimate the degree to which they ought to replicate the findings with a new sample based on the features of the current sample.

Say you are interested in the difference between pre and post treatment for an intervention. In your sample, you find a mean difference of $c$. Another researcher attempts to replicate your intervention with the same population, but finds their estimate of the difference to be $d$, where $d$ is of the opposite sign to $c$. However, the two samples differ on a number of demographic variables. The question is whether you ought to expect to see a difference $d$ in the sample given that you saw a difference $c$ in the original sample.

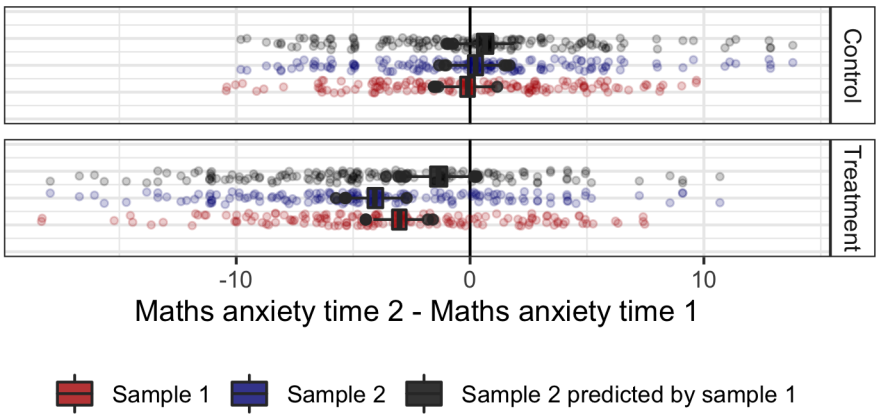


*Figure 5*. Expected post-pre difference in maths anxiety estimated in sample 1 (red), sample 2 (blue) and sample 1 predicting sample 2 (black). Points represent observed differences in the sample (sample 2 in the first two rows in each figure, sample 1 in the third).

Following on from our previous example of a maths anxiety intervention, we now consider using our first sample, which happens to contain mostly social science students, to predict the expected impact of the intervention on another sample now mostly engineering and science students. If we expect there to be heterogeneity in the expected effect of the intervention, it is possible that the effect observed in sample 1 will be different than the raw effect observed in sample 2.

2. In this case we are going to treat sample 2 as the population. We could summarize it as a poststratification table, but provided it is relatively small (in this case we simulate a second sample of 300) we can simply use it as individual data. We can repeat steps three and 4 from the previous section using sample 1.

5. Now we can predict the baseline maths anxiety and post maths anxiety in sample 2 pretending as though each participant was given both treatment and control. This is similar to step 5 in the previous section, except the population is now the second sample.

6. We can repeat Figure 4 but instead of predicting the difference between treatments in the undergraduate population, we predict the difference in treatment and control in sample 2 using sample 1. We present these estimates in Figure 5.

**Generalize to other universities**

In the data accompanying this tutorial, we simulate not just one university but multiple universities, subsetting down to one for simplicity's sake in the previous sections. If we wanted to generalize to other universities we would need to have samples from multiple universities and model university, and university characteristics as another random effect in the model. If we didn't do this then we would be essentially ignoring university level heterogeneity. An example of a sampling design that aimed to account for heterogeneity between schools, see Yeager et al. (2019). For simplicity we do not go through this here;

the approach would likely be similar to approaches by Lax and Phillips (2009a) to model state.

## Active research areas

At this point you may notice that all of these examples of MRP's possibilities share certain features. Absence of these features correspond to some of the limitations of our method. MRP is widely used throughout the political science literature but is relatively new to psychology. The provided examples demonstrate that it already can be a valuable tool, but research need to be done to modify this method to specifically suit psychology's aims.

One of the main challenges of MRP as set up above is that it is designed to estimate a parameter in the population given some demographic characteristics. In a pre-post design, the difference between pre and post can be treated as the outcome and implemented similarly or completed in a two-stage process as demonstrated in this tutorial. However, psychology is a science that considers complex relationships. For instance, consider the example used by Simons et al. (2017) for the article by Whitsett and Shoda (2014) investigating the relationship between support seekers distress and willingness to provide support, mediated by high and low personal distress. In their COG statement, Simons et al. (2017) note that the sample was "a large number of different undergraduates sampled from the subject pool at the University of Washington" and that they "believe the results will be reproducible with students from similar subject pools serving as participants."

From this information, we can infer that new undergraduates sampled from the University of Washington would be expected to show a similar relationship between participant willingness to support and support seeker's degree of distress. But further modeling would be needed to formalize this in an MRP context, which would allow the relationship to change given demographic characteristics. We expect the approach would be similar to Hill (2011).

Indeed, all the examples in the paper by Simons et al. (2017) involve the generalization of an observed relationship in a sample to a wider population or a different sample. When considering Simons (2013), a study investigating the Dunning-Kroger effect with competitive bridge players, Simons et al. (2017) suggest a COG that includes

> A direct replication would test bridge players in sessions that include players with skill levels ranging from relative novice to expert in the context of their regular bridge game (p. 1126 Simons et al., 2017)

suggesting that the effect in player groups would replicate with a more diverse skill level, provided they still regularly compete. Although we present a simulated example considering the heterogeneous effect of an experimental manipulation, we suspect that further research needs to be done on the choice of sensible priors to induce regularization in a reasonable way.

## The impact on COG statements

The method we propose here wouldn't be possible without the proposal to include COG statements in psychological research. However for these statements to be maximally

useful, they will need to be as specific as possible. We have proposed some additional guidelines to keep in mind when writing COG statements.

Moderators of an effect or potential individual differences should be clearly listed. For example, in our hypothetical example of maths anxiety, we expected there to be both individual differences in gender and university major. We also hypothesized potential moderators for the effect of the intervention. Ideally these moderators/individual differences should be identified before collection and measured in the main data collection phase.

Then the population the results are intended to generalize to should be clearly stated. With some exceptions, a researcher should start with the population of individuals who had potential to be in the study (i.e., psychology undergraduates). This may be the only population that the results can generalize to. However in some cases we can assume that the results generalize further (e.g., to all undergraduate students in the university, to all undergraduates in the country, or even all adults in the country). This generalization is untestable without observing a wider sample, and so it should be clearly stated that this is an assumption that relies on either there being no differences between the sample and the population, or that the differences between the sample and the population are unrelated to the outcome of interest in the study.

## Limitations

As always in statistics, our claims are only as good as our models. MRP (or, more generally, regularized regression and poststratification) relies on a model or procedure to predict the outcome variable given some set of demographics. This model can fail to make good predictions for several reasons, including insufficient data, the lack of some demographic predictor, misspecification of some important part of the model, or insufficient regularization. Partial pooling in multilevel modeling uses data efficiently to mitigate some of these concerns, and a solid COG statement helps to provide some focus on the others.

Other limitations to this method arise because not only do we need to collect demographic variables in the sample (arguably relatively easy to do with some forethought), but we also need estimates of these same demographic variables in the population. These data are often available through government and census data, but not always and not always in the desired form. Some creativity may be needed to coerce available data into the desired form. For example, in political surveys it can be helpful to poststratify on party identification, which is not in the census and so one must use other surveys to estimate its distribution conditional on the relevant demographic and geographic predictors.

## Conclusion

We argue here that one of the important benefits of the COG statement is that it paves the way for statistical methods like MRP to be used, and we encourage other researchers to join us in considering how and when this technique might benefit the field.

Psychology has developed methods of estimating and evaluating internal validity through our rich and rigorous training in experimental design. However, we must not let a stellar job of accounting for internal validity distract us from also considering external validity. Inferences from convenience and snowball samples have serious threats to external validity once we consider heterogeneity of effects.

We have demonstrated how psychology can use MRP to estimate average treatment effects in defined populations, a particularly relevant task when working with non-probability samples. However, MRP will not always be a perfect solution. MRP is useful for adjusting a non-representative sample to a larger population. It is not, however, designed for situations where there are no individuals in a particular sub-population present in the sample (for example, using data from a WEIRD sample to generalize to the larger population of the world). In this case, we must either rely on strong assumptions or broaden our data pool through collaborations across the world—which is perhaps one of the most encouraging possibilities for MRP.

## References

Bureau, U. S. C. (2012). American Community Survey (ACS): PUMS. Retrieved January 29, 2019, from https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t#

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Caughey, D. & Warshaw, C. (2019). Public opinion in subnational politics. University of Chicago Press Chicago, IL.

Donnellan, M. B. & Lucas, R. E. (2008). Age differences in the big five across the life span: Evidence from two national samples. *Psychology and aging*, *23*(3), 558.

Downes, M., Gurrin, L. C., English, D. R., Pirkis, J., Currier, D., Spittal, M. J., & Carlin, J. B. (2018). Multilevel regression and poststratification: A modelling approach to estimating population quantities from highly selected survey samples. *American Journal of Epidemiology*.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A. & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression.

Gelman, A. & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*(2), 241–251.

Ghitza, Y. & Gelman, A. (2013). Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, *57*(3), 762–776.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, *4*(1), 26.

Goldberg, L. R. et al. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, *7*(1), 7–28.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, *40*(1), 84–96.

Green, D. P. & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, *76*(3), 491–511.

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, *19*(2), 49–71.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

International Personality Item Pool Project. (n.d.). *50-item sample questionnaire*. Retrieved from https://ipip.ori.org/new_ipip-50-item-scale.htm

Lax, J. R. & Phillips, J. H. (2009a). Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review*, *103*(3), 367–386.

Lax, J. R. & Phillips, J. H. (2009b). How should we estimate public opinion in the states? *American Journal of Political Science*, *53*(1), 107–121.

Little, R. J. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, *88*(423), 1001–1012.

Mook, D. G. (1983). In defense of external invalidity. *American psychologist*, *38*(4), 379.

Open Source Psychometrics Project. (2019). Open psychology data: Raw data from online personality tests. Retrieved April 30, 2019, from https://openpsychometrics.org/_rawdata/

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with post-stratification: State-level estimates from national polls. *Political Analysis*, *12*(4), 375–385.

Schad, D. J., Betancourt, M., & Vasishth, S. (2019). Toward a principled bayesian workflow in cognitive science. arXiv: 1904.12765 [stat.ME]

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, *51*(3), 515.

Si, Y., Trangucci, R., Gabry, J. S., & Gelman, A. (2017). Bayesian hierarchical weighting adjustment and survey inference. *arXiv preprint*, *1707.08220*.

Simons, D. J. (2013). Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin & Review*, *20*(3), 601–607.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128.

Smith, P. L. & Little, D. R. (2018). Small is beautiful: In defense of the small-n design. *Psychonomic bulletin & review*, *25*(6), 2083–2101.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, *12*(3), 175–200.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991.

Weber, S., Gelman, A., Lee, D., Betancourt, M., Vehtari, A., & Racine-Poon, A. (2018). Bayesian aggregation of average data: An application in drug development. *The Annals of Applied Statistics*.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, *2*, 178.

Whitsett, D. D. & Shoda, Y. (2014). An approach to test for individual differences in the effects of situations without using moderator variables. *Journal of Experimental Social Psychology*, *50*, 94–104.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., ... Hinojosa, C. P., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369.

Yuxiang, G., Kennedy, L., Simpson, D., & Gelman, A. (2019). Improving multilevel regression and poststratification with structured priors. arXiv: 1908.06716 `[stat.ME]`