Multi-level regression and post-stratification for discrete choice modelling and stated preference research

Patrick Lloyd-Smith*

February 17, 2026

Obtaining valid answers from respondents has been a central concern of stated preference (SP) studies. In contrast, sample representativeness is the main focus of current public polling accuracy debates, largely due to declining survey response rates. We introduce the multi-level regression and post-stratification (MRP) modelling approach to welfare estimation using discrete choice models. Through monte carlo simulations and two water resource valuation surveys, we demonstrate how MRP can help researchers to (i) generate population-relevant welfare measures from non-representative samples, (ii) estimate preference heterogeneity and distributional impacts across people, and (iii) evaluate the impacts of removing potentially invalid responses using a consistent target population frame. We propose MRP as a complementary modelling approach for non-market valuation data and discuss its opportunities and limitations.

*Patrick Lloyd-Smith is an Associate Professor in the Department of Agricultural and Resource Economics at the University of Saskatchewan (Room 3D34, Agriculture Building, 51 Campus Drive, Saskatoon, SK S7N5A8 Canada; patrick.lloydsmith@usask.ca). This research was supported by the Social Science and Humanities Research Council (SSHRC), the Global Water Futures research programme, and Smart Prosperity. I thank the editor Klaus Moeltner and two anonymous reviewers, seminar participants at Oregon State University, University of Alberta, and University of Saskatchewan as well as participants of the 2024 Canadian Resource and Environmental Economics Association conference for helpful comments on this research. I have no conflict of interest or financial discloures to declare.

Introduction

Stated preference (SP) methods are the dominant approach in environmental valuation. The largest global database of environmental valuation studies contains twice as many SP primary valuation studies as all other primary valuation methods combined.[1] In many settings, particularly when valuing non-use benefits or assessing prospective environmental changes, SP methods are the only feasible means of estimating welfare effects. While practitioners often state a preference for revealed preference methods, they demonstrate a *revealed preference* for SP in applied work.[2]

Despite their prevalence, SP methods have long faced scrutiny. Concerns about hypothetical bias, strategic behavior, scope sensitivity and other survey biases intensified following high-profile natural resource damage cases such as Exxon Valdez and Deepwater Horizon, where SP estimates, particularly of non-use values, were subject to intense criticism (Maas and Svorenčík 2017; Banzhaf 2017; McFadden and Train 2017; Hausman 2012; Kling, Phaneuf, and Zhao 2012). In response, a substantial literature and set of best practices has developed to strengthen SP research (Johnston et al. 2017), drawing on mechanism design theory to improve incentive compatibility and response validity (Vossler, Doyon, and Rondeau 2012; Carson, Groves, and List 2014). Much of this work has focused on the internal validity of responses among those who complete surveys.

Yet an equally important, and often underemphasized, challenge lies upstream: who is responding in the first place. Contemporary SP practice increasingly relies on opt-in, non-probability internet panels because they are fast, affordable, and scalable. The use of probability-based sampling strategies in academic SP research is now the exception rather than the norm.[3] However, opt-in panels are susceptible to both known and unknown forms of selection bias, raising

[1] By May 2025, the Environmental Valuation Reference Inventory (EVRI) contained 3,257 SP primary valuation studies compared to 982 revealed preference studies and 617 market price studies. Macaskill and Lloyd-Smith (2022) provide further empirical evidence that SP studies now make up 80% of recent environmental resource valuation studies in Canada.

[2] Part of this preference reflects the constraint that for many research questions, SP is the only method available given the lack of appropriate RP data or the requirement to estimate non-use values.

[3] One notable exception is the well-founded Deepwater Horizon SP study (Bishop et al. 2017) where obtaining a high quality sample and assessing its generalizability to the population was emphasized throughout.

concerns about the representativeness of resulting welfare estimates (Baker et al. 2013; Bailey 2024). Moreover, SP researchers are often institutionally and operationally removed from the sampling process itself, making it easy to overlook coverage and nonresponse issues that can undermine population-level inference.

These concerns mirror a broader crisis in public opinion polling. Well-publicized polling misses, such as the 2016 U.S. presidential election and Brexit, have fueled skepticism about surveys as tools for understanding public preferences (Skibba 2016; Bailey 2024). A primary driver of these credibility issues is the substantial decline in response rates across survey modes (Keeter and Christian 2012; Keeter et al. 2017; Bailey 2024). Response rates can now fall below 1% even for well-regarded pollsters.[4] Such declines intensify concerns about sample representativeness and population inference (Bailey 2024). As Prosser and Mellon (2018) note in their review of polling errors, "By far the most common cause of polling error is unrepresentative samples." While the "death of polling" may be exaggerated (Prosser and Mellon 2018), modern survey researchers are having to substantially revise their methods to keep up with changing respondent habits and technology (Bailey 2024).

If polling organizations struggle to produce accurate insights even with substantial resources, short and relatively simple questionnaires, and clear external benchmarks in the form of concurrent surveys and actual election outcomes, SP research operates under even more demanding inferential conditions. SP surveys impose substantial cognitive demands on respondents, employ complex experimental designs, and operate with limited budgets and limited formal training in survey methodology. Opportunities for criterion validity tests are also rare (Johnston 2006; Kling, Phaneuf, and Zhao 2012). At the same time, policy interest is shifting beyond average willingness-to-pay toward distributional analyses ("OMB" 2023), increasing the importance of accurate subgroup-level inference.

The challenges facing SP researchers can be conceptualized within the total survey error (TSE) framework (Groves et al. 2009). Figure 1 adapts the TSE framework to SP research and

[4]To obtain telephone responses from 4,097 registered voters in 5 US states, the New York Times/Siena College Poll called 410,000 people (Cohn 2024).

highlights the multiple stages at which error can enter the process from a conceptual welfare measure (e.g., compensating variation) to an estimated survey statistic such as estimated willingness-to-pay. While SP scholars have devoted considerable effort to minimizing measurement error among respondents, contemporary survey realities, particularly the widespread reliance on opt-in samples, underscore the need to confront representation error with equal rigour. The experience of public polling underscores that concerns about representativeness are not peripheral; they are central to the policy credibility of SP estimates.

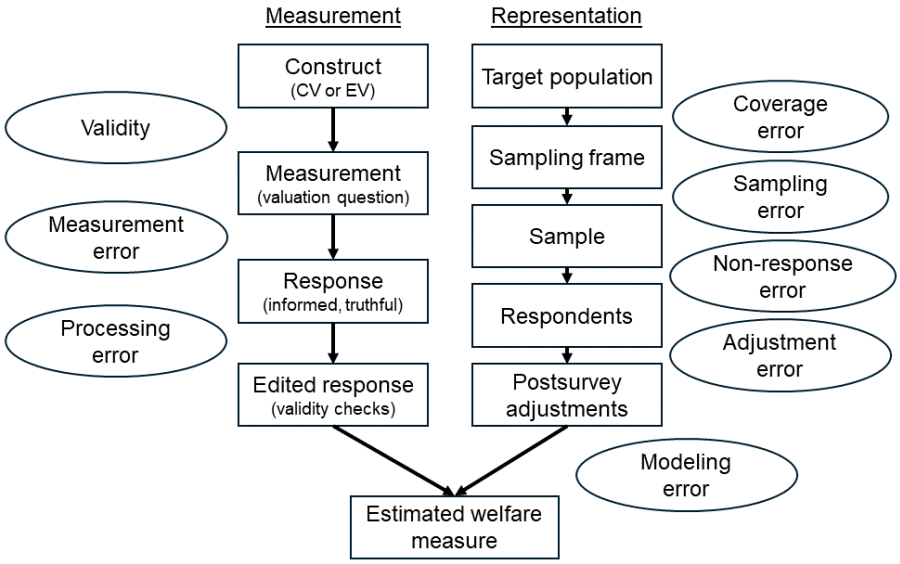

Figure 1: The Total Survey Error (TSE) framework for stated preference welfare estimation (adapted from Groves et al. (2009)).

This paper adapts the multilevel regression with post-stratification (MRP) modelling approach to welfare estimation using discrete choice models. The main use case of MRP is to better generalize population-relevant insights from non-representative samples, a common task in SP research. MRP has been developed in the political science literature (Gelman and Little 1997) and mainly used to predict people's voting behavior by survey companies such as YouGov. The MRP approach consists of three steps. First, estimate a binary or multinomial

logit model with random parameters specified over respondent characteristics such as age and income groups. This specification is in contrast to the typical random parameters logit (RPL) model where heterogeneity is modeled at the individual level. Second, use the group-level heterogeneity modelling to impute welfare measures for all subgroups in the target population where subgroups are defined based on these respondent characteristics. Third, aggregate the subgroup estimates using auxiliary counts data from sources such as the census to generate population-relevant welfare measures.

Using experimental simulations and data from two water resource surveys, we demonstrate how MRP can generate population-relevant estimates from non-representative samples and two additional applications of MRP for SP practitioners. The second use case is that MRP provides a complementary approach to estimating preference heterogeneity and conducting distributional analyses. The MRP approach combines the benefits of knowing the source of the heterogeneity by including respondent characteristics in the model with the partial pooling benefits of random parameters to focus on subgroup level rather than individual level heterogeneity. Specifying random parameters at the respondent characteristic level allows the MRP approach to use this partially pooled information across people. This approach is particularly promising for cases when specific subgroups of interest are poorly represented, or entirely absent, in the survey data. Additionally, the MRP approach also assesses population-level heterogeneity rather than merely sample heterogeneity due to the post-stratification step.

The third MRP use case for SP is to provide a framework for estimating consistent population-relevant welfare measures despite changes in analysis samples. For example, a common task in SP research is assessing the welfare impacts of excluding potentially 'invalid' respondents such as yea-sayers (Johnston et al. 2017; Poe 2016). While such exclusions may enhance the internal validity of the sample and results, they risk making the analysis sample less representative of the population. Due to the post-stratification step, the MRP approach maintains a consistent target population frame for deriving welfare measures, irrespective of the specific analysis sample used in model estimation. Throughout, the empirical demonstrations aim to show how

SP practitioners using typical survey data can benefit from MRP in their work.

While this paper focuses on a model-based solution to addressing sample non-representativeness, an alternative solution is to avoid using opt-in internet panels in favor of higher quality, probability-based samples. However, several challenges limit this solution. First, the high costs of obtaining probability-based sample makes it unfeasible for most SP study budgets. Second, the continued decline in response rates means even high quality probability survey samples require adjustments for non-representativeness, where the MRP approach can be useful. Third, relying solely on better sampling going forward would limit the policy-making insights use in benefit transfer exercises of existing SP survey datasets. The MRP approach introduced in this paper can be applied to older, unrepresentative survey samples, provided contemporaneously comparable auxiliary data for the target population is available.

The paper contributes to the SP literature proposing solutions to the well-known issue of generalizing insights from sample to population. A substantial body of research has explored the effects of survey mode on SP sample composition and response quality, particularly the impact of using the internet for survey administration (Lindhjem and Navrud (2011) Boyle et al. (2016)). A related literature compares the use of probability-based and non-probability, opt-in panel samples (Whitehead et al. 2023; Penn, Petrolia, and Fannin 2023; Sandstrom-Mistry et al. 2023), revealing mixed results regarding differences in sample composition, SP data quality, and welfare estimates. Broader issues of non-response bias in internet surveys have also been investigated and two-stage correction solutions proposed for discrete choice models (Johnston and Abdulrahman 2017; Kolstoe, Naald, and Cohan 2022; Bonnichsen and Olsen 2016). Most SP studies using opt-in panels employ some form of sampling design, such as quota-based sampling, to improve sample representativeness. Representativeness is then assessed by comparing marginal distributions of a select few respondent characteristics with target population benchmarks. If differences are found, researchers often include caveats in their interpretation and/or apply explicit weighting of the sample to match the population (Bonnichsen and Olsen 2016).

MRP shares some similarities and limitations to conventional survey weighting in its use of ob-

served respondent characteristics and auxiliary population information to make adjustments for non-representativeness. While survey weighting tries to address non-representativeness before modelling to estimate population-relevant model parameters, MRP models the unadjusted, non-representative data directly and uses post-stratification to generate population-level insights. As described in Section 2, the main advantage of MRP over weighting and post-stratification with fixed parameters is the combination of group-level random parameters and post-stratification which allows efficient pooling of preference information across people while also using information on the subgroup structure of the target population. In direct comparisons using political science and health data, MRP performs better than weighting and disaggregation, particularly for smaller subgroups of the population (Downes and Carlin 2020b; Lax and Phillips 2009; Warshaw and Rodden 2012). We contribute to this empirical evidence using SP simulations and data.

Like survey weighting, MRP does not address the challenging issue of systematic nonresponse due to unobservable respondent characteristics. However, implementing proper adjustments for observed respondent characteristics is still important to reduce representation errors. For example, in their evaluation of why the 2016 presidential election polls provided misleading insights, the committee set up by the American Association for Public Opinion Research (AAPOR) found one of the two reasons with the most evidence was the pervasive lack of adjustments for educational attainment (Kennedy et al. 2018). Furthermore, there was little evidence for the more difficult to measure social desirability in responses (i.e. 'Shy Trump' voters) causing the polling errors.

This paper also contributes to the SP literature on evaluating different elicitation formats. Single binary choice (SBC) has long been held up as the "gold standard" for validity given its attractive incentive compatibility qualities (Arrow et al. 1993). However, preference information collected from SBC is limited in terms of estimating marginal welfare measures for specific attributes of environmental quality changes and assessing preference heterogeneity (Johnston et al. 2017). Formats that collect more preference information raise additional incentive compatibility concerns. For example, asking multiple binary choice questions requires

that respondents view each choice as independent Vossler et al. (2025) and trinary choice questions commonly implemented in choice experiments are not incentive compatible (Carson and Groves 2007). This paper identifies another potential trade-off in valuation formats when using non-representative samples. While SBC may be preferred for validity, formats that gather more individual-specific preference information may be better suited for non-representative samples as this information can improve the generalizability inference as we move from the sample to the population.

The remainder of the paper proceeds as follows. Section 2 adapts MRP to discrete choice models and welfare measurement. Section 3 uses Monte Carlo simulation experiments to compare the performance of MRP to weighting and post-stratification with fixed parameters. Section 4 describes the two SP survey datasets as well as the post-stratification data and Section 5 applies this data in empirical applications to demonstrates three ways MRP can help SP researchers. Section 6 concludes by describing limitations and opportunities of MRP in non-market valuation applications, including beyond SP settings.

Multilevel regression with post-stratification (MRP) and random utility maximization (RUM) models

We first introduce MRP and then adapt it to random utility maximization (RUM) models and welfare estimation. MRP combines ideas about multilevel models, hierarchical Bayesian estimation, and post-stratification (Gelman and Little 1997). MRP has mainly been used for two related applications that are both relevant for SP research. The first application involves using non-representative survey samples to generate population-relevant estimates. One of the most spectacular demonstration of the MRP's potential is the Xbox election study, where the highly unrepresentative users of the video gaming platform were surveyed (e.g. the sample was 93% males, with 65% of the sample aged 18 to 29). MRP was used to produce reliable forecasts of the 2012 US presidential election (Wang et al. 2015). Given the prevalent use of opt-in internet panels in SP research and the implications for sample representativeness, MRP has similar potential for this use application.

The second MRP application is in small-area estimation, which refers to using national surveys to generate estimates of subnational opinions (e.g. Park, Gelman, and Bafumi (2017); Lax and Phillips (2009)). MRP has been tested in numerous political science and public polling contexts, with evidence suggesting it outperforms alternatives such as disaggregation and weighting (Lax and Phillips 2009; Warshaw and Rodden 2012)) and has emerged as the "gold standard" of using national surveys to estimate local preferences (Selb and Munzert 2011). While the spatial distribution and downscaling of preferences is certainly one important focus area of SP research (Glenk et al. 2020), the broader relevance of small-area estimation techniques for SP research lies in their ability to robustly estimate preference heterogeneity across different groups, including assessing distributional impacts.

Most MRP applications have focused on voting preferences, using binary logit models to estimate voting probabilities and predict election outcomes (Leemann and Wasserfallen 2020). In non-market valuation research, we are typically interested in welfare estimates based in random utility theory. The rest of this section outlines the three main implementation steps in MRP adapted to RUM models. The empirical applications presented later will provide particular implementations of these models for single binary choice and choice experiment data.

1 MRP Step 1: Estimation of a modified random parameters logit model in WTP-space

We assume that the utility that individual i receives from alternative j in choice task t is defined as

$$U^*_{ijt} = \beta^* x_{jt} + \alpha^* p_{jt} + \varepsilon^*_{ijt}, \qquad \varepsilon^*_{ijt} \sim \text{Gumbel}\left(0, \sigma^2 \frac{\pi^2}{6}\right) \tag{1}$$

where x_{jt} is a vector of non-monetary attributes, p_{jt} is the monetary attribute, ε^*_{ijt} is the error

term, which is assumed to follow a Gumbel distribution, and β^* and α^* are the parameters of interest. This general discrete choice model specification is not identified and we need some type of normalization to estimate parameters. Typically, σ^2 is normalized to 1 and the model is estimated in preference-space. We specify utility in WTP-space[5] rather than preference-space because SP research is typically interested in welfare measures and to avoid the additional step of calculating the marginal rates of substitution between non-monetary attribute parameters and the price parameter. To do so, we normalize through the price parameter ($-\alpha^*$) and then multiply both sides of Equation 1 by $\lambda = (-\alpha^*/\sigma)$ to derive

$$U_{ijt} = \lambda(\omega x_{jt} - p_{jt}) + \varepsilon_{ijt}, \qquad \varepsilon_{ijt} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{2}$$

where $\omega = (\beta^*/-\alpha^*)$ can be now interpreted as the marginal WTP for the set of non-monetary attribute.

We relax the assumption that ω is the same for all individuals by characterizing ω_i as a random parameter. Unlike the typical RPL model employed in non-market valuation research, where heterogeneity is specified at the individual-level, here we model heterogeneity at the respondent characteristics level. We specify R random parameters at the respondent characteristic level which can include categorical variables such as age and income group. Thus, we specify ω_i as

$$\omega_i = \beta_0 + \sum_{r=1}^{R} \alpha^r_{c_r[i]} \tag{3}$$

where β_0 is the mean parameter and $\alpha^r_{c_r}$ is the random parameter for a respondent characteristic r with C_r number of categories indexed by c_r. We specify a normal distribution for the random parameters with zero mean and a standard deviation σ^2_r such that $\alpha^r_{c_r} \sim N(0, \sigma^2_r)$ for $c_r = 1, ..., C_r$, where C_r can vary by respondent characteristic. The purpose of this step

[5]For the rest of the paper, we use the term WTP-space as the empirical applications are both WTP contexts, but the modelling can be set-up analogously to be in WTA-space.

is to build a model for predicting subgroup WTP in step 2, not necessarily to interpret these parameters.

The emphasis on using random parameters over observed respondent characteristics to explain heterogeneity distinguishes the MRP approach from other approaches to modelling preference heterogeneity. RPL or latent class (LC) models excel at uncovering unobserved preferences heterogeneity whether continuous or discrete, but the partially unexplained nature of these preference differences makes these models less equipped for adjustments from sample to population. Furthermore, RPL models using normal distributions to model individual-level heterogeneity often leads to implausible individual-level WTP estimates in the tails.[6] While normal distributions are typically intended as approximate representations rather than literal interpretations, understanding the 'tails' of distributions is often important for identifying people who may be disproportionately affected (either positively or negatively) by policy changes. LC logit models also estimate subgroup WTP partially defined by respondent characteristics, but these 'classes' are endogenous determined by the model rather than the MRP approach that pre-defines subgroups based on unique sets of observed individual characteristics. Thus, the MRP approach trades insights into more flexible, unobserved individual-level heterogeneity modeled with a typical RPL or LC model to gain insights into how this heterogeneity varies with respondent characteristics. In this sense, MRP shares the observed heterogeneity focus of fixed parameter logit specifications where respondent characteristics are included as mean shifters, but the MRP approach uses random parameters to minimize overfitting.

Specifying random parameters at the respondent characteristic level provides some benefits relative to fixed parameter specifications that include interaction effects. A challenge with incorporating interaction effects is that selecting respondent characteristics for inclusion is complicated by the large number of parameters that need to be estimated and typically small survey samples, especially across subgroups. In this sense, MRP provides a middle ground between the large number of estimated parameters if all respondent characteristic category

[6]While normal distributions are by far the most popular, researchers are not restricted to them and several alternative distributions that can bound preference parameters such as log-normal are sometimes specified, although with their own limitations and issues.

levels are included as mean shifters (i.e. $C_r - 1$ parameters for each respondent characteristic)[7] and the sparsity of the typical RPL specification where only one individual-level standard deviation parameter for each attribute is estimated. In the MRP specification, we estimate R standard deviation parameters (σ_r^2) for each attribute. One limitation of the MRP approach relative to the fixed parameter is that respondent variables must be categorical and thus any continuous variable requires discretization. The influence of this discretization process is assessed in the simulations.

1.0.1 Bayesian estimation

We implement MRP using Bayesian methods for two main reasons. First, Bayesian methods are helpful in the estimation step as the nonlinearity of the WTP-space specification means there is no guarantee of finding a global maximum if maximum simulated likelihood methods are used (Train 2009). Second, Bayesian methods provide a coherent workflow for propagating parameter uncertainty from the estimation step to the imputation and post-stratification steps as we can use the full predicted posterior distribution of each parameter to calculate WTP.

All models are estimated using Bayesian techniques implemented using the Stan programming language (Carpenter et al. 2017) through the *brms* R package (Bürkner 2017). The models are derived using Dynamic Hamiltonian Monte Carlo (HMC) inference algorithm. For each model, four independent chains are generated with 2,000 iterations each. The first 1,000 draws for each chain are considered warm-up and the remaining 1,000 are considered as valid sample draws for a total of 4,000 across all four chains. HMC typically requires much fewer iterations to produce effective independent samples than random walk samplers such as Gibbs Sampler or Metropolis–Hastings algorithm because HMC uses gradients and the geometry of the posterior to move longer distances in the parameter space. Thus the autocorrelation among iterations is much lower although each iteration is computationally more expensive. Additional details on

[7] For example, including an income variable with 7 levels require 6 estimated parameters in the fixed parameter specification but only a single standard deviation parameter for the random parameter in MRP.

HMC and the specific No-U-turn (NUTS) sampler used is provided by (Hoffman and Gelman 2014) and (Stan Development Team 2024).

Model diagnostics were reviewed to ensure convergence and enough appropriate samples were obtained. Convergence of the models were primarily assessed through the Potential Scale Reduction Factor (R-hat) which compares within-chain variance and between-chain variance of the MCMC samples. All R-hat values were less than 1.01. We assess the efficiency of the samples and ensuring enough samples were run by computing the size of an equivalent independent sample through Bulk Effective Sample Size (ESS) and Tail ESS measures. Other model diagnostics and warnings that were checked to ensure that there were no reported divergent transitions, maximum tree depth exception warnings, or low Energy Bayesian Fraction of Missing Information (E-BFMI) warnings.

Because we are using Bayesian methods, we need to specify priors for all the parameters. Sampling from complex Bayesian models is facilitated by scaling variables so that parameters are close to zero and priors can be specified as standard normals. We first scale the price variable so that we can use standard normal prior for the mean parameter ($\beta_0 \sim N(0,1)$).[8] We specify half-Cauchy priors for the standard deviation parameters ($\sigma_r^2 \sim$ half-Cauchy$(0,1)$) and a half-normal for the scale parameter ($\lambda \sim$ half-$N(0,1)$) to bound these parameters to be positive.[9]

2 MRP Step 2: Impute subgroup WTP

In the second step, we impute willingness-to-pay ($\widehat{WTP}_s$) for each subgroup s using the model specified in Equation 2. Subgroups are defined by relevant combinations of respondent

[8] In the two examples below, the price variable is scaled by 100, but this will be unique to the study context.

[9] We use a half-Cauchy prior for the standard deviation parameters to allow heavy tails in the estimated random effects when supported by the data and based on earlier work showing its performance over other options such as inverse-gamma or a uniform prior (Gelman 2006). To inform the choice of the prior for the scale parameter, we examined the results of a simpler MNL specification where the scale parameter was estimated to be 0.403 (see Table 2). We used a weakly informative prior of $\lambda \sim$ half-$N(0,1)$ to aid convergence as we found convergence issues when the scale parameter was given a less informative prior such as $\lambda \sim$ half-$N(0,10)$.

characteristic levels. The number of subgroups to impute is typically all unique combinations of these levels, although there may be contexts where a category level is included in the model and not imputed (e.g.respondents with missing income information). If we have R respondent characteristics, each with C_r number of categories, then we have a total of $S = \prod_{r=1}^{R} C_r$ subgroups to impute. The number of subgroups to impute can increase quite rapidly as for example if we have five respondent characteristics each with five categories, then we would impute WTP for $5^5 = 3,125$ subgroups.

Because the model is estimated in WTP-space, we can use the model parameters directly to impute $\hat{WTP}_s$ using the estimates for the mean parameters β as well as realizations of the random parameters. For example, if subgroup s is part of the first age group, then $\hat{\alpha}_1^{age}$. The marginal $\hat{WTP}_s$ for a particular attribute can be computed by adding the relevant random parameter realizations such that

$$\hat{WTP}_s = \hat{\beta}_0 + \sum_{r=1}^{R} \hat{\alpha}^r_{c_r[s]}$$

Because we are using Bayesian methods in estimation, we derive the full posterior distribution of $\hat{WTP}_s$ and we draws from this distribution in step 3.

The large number of subgroups to impute is where the benefits of specifying random parameters at the respondent characteristic level in Step 1 becomes apparent as the partial pooling allows us to obtain more robust WTP estimates without overfitting. For example, the WTP for a 18-34 year old female who is a university graduate and lives in region A is influenced by all people in the 18-34 year category, all females, all university graduates, and all people living in region A. Partial pooling also allows us to predict WTP for a subgroup even if no respondents from that subgroup are represented in the survey, as we can borrow preference information from similar individuals.

3 MRP Step 3: Post-stratification

The third step is post-stratification, where we weight the draws from the posterior distribution of $\hat{WTP}_s$ imputed in step 2 by the corresponding subgroup counts (N_s) in the target population using auxiliary data.[10] Thus, we calculate the full posterior distribution of $\hat{WTP}^{MRP}$ as

$$\hat{WTP}^{MRP} = \frac{\sum_{s=1}^{S}(N_s \times \hat{WTP}_s)}{\sum_{s=1}^{S} N_s} \tag{4}$$

For most general population surveys, the most relevant auxiliary data on the target population comes from government census sources. The use of auxiliary data also restricts the types of respondent characteristics we can use to define subgroups as there is a need for a corresponding match between survey and auxiliary information.

The post-stratification step does not need to be a single aggregate population estimate but can be also conducted for any group defined using the respondent characteristics. As we demonstrate in the empirical applications later, we can estimate welfare measures across the education distribution while accounting for the unique characteristics of each educational group.

Monte Carlo Analysis

With the general MRP modelling workflow outlined, we use Monte Carlo evidence to assess the performance of MRP relative to two other survey adjustment methods for estimating population-level WTP: weighting and post-stratification. We first detail the set of generated data experiments and then summarize the results.[11]

[10] Individual-level welfare estimates from a typical RPL model could also be post-stratified to produce a population-relevant estimate, but this process would require all subgroups to be represented in the data which is unlikely to be the case.

[11] All the code for these Monte Carlo experiments are available at https://github.com/plloydsmith/MRPandSP.

4 Experimental setup

Population structure: We generate a synthetic population of N = 1,000,000 individuals that reside in K = 20 regions. The population shares of regions range in size from 1% to 17%. Each individual is characterized by their income (Inc_i^{cont}) and 3 other demographic characteristics (X_{ij}): age, education, and city size. Each demographic characteristics has 8 levels whereas income is a 0 mean centered continuous variable ranging from -0.5 to 0.5. We assume that census information is known and available on the population structure for the survey adjustment models.

True willingness-to-pay model: An individual i's WTP is defined as

$$WTP_i = \beta_0 + \sum_j \beta_j(X_{ij}) + g(Inc_i^{cont}) + \alpha_r(i) + \varepsilon_i$$

where

- $\beta_0 = 100$: Baseline WTP.
- $\beta_j(X_{ij})$: Main effects for the 3 non-income individual characteristic j linearly spaced across levels from -25 to +25.
- $g(Inc_i^{cont}) = 50(Inc^{cont})_i - 20(Inc_i^{cont})^2$ is a quadratic function relating income and WTP which implies WTP increases at all income levels in our data, but at a diminishing rate.[12] This translates into the lowest effect of income on WTP to be -30 and the highest +30 similar to the other demographic variables.
- $\alpha_r(i) \sim N(0, \sigma_r^2)$ with σ_r^2=25: Regional random effects.
- $\varepsilon_i \sim N(0, \sigma^2)$ with σ^2=50: Individual-level-heterogeneity.

[12]We assume nonlinearity between income and WTP to assess how the methods perform with a more complex relationship. A common empirical finding in the environmental valuation literature is an income elasticity below one indicates that WTP increases less than proportionally with income (Kristrom and Riera, 1996). Although this does not imply a quadratic relationship, it does suggest a diminishing marginal effect of income on WTP. Moreover, a concave relationship between income and WTP is often used in conceptual valuation research (Drupp et al., 2018).

After generating individual-level WTP in the population, we normalize each individual's WTP by the population-weighted mean WTP. This normalization implies that the true WTP in the population is 1 and provides a consistent benchmark for comparing estimates across scenarios.

Non-response bias mechanism: We implement multiplicative non-response bias where sampling probability depends on the same set of individual characteristics that also influence WTP.[13] While income is continuous in its effects on WTP, we discretize income into 8 evenly spaced categories (Inc_i^{cat}) for sampling and model estimation to match the other characteristics and assess the affect of this categorization process. Response probabilities are defined for subgroups of people s defined by their unique set of individual characteristics (X_{ij}), income category (Inc_i^{cat}), and region k. The probability of an individual with the unique set of characteristics in subgroup s is

$$P(sample|\text{Region}_k, Inc_i^{cat}, X_{ij}) \propto \ weight_s \times \text{bias}_{inc}(Inc_i^{cat})^{\beta_{\text{strength}}} \times \prod_j [\text{bias}_j(X_{ij})]^{\beta_{\text{strength}}} \tag{5}$$

where

- $weight_s$ is the population weight of subgroup s which the individual belongs to. Subgroups are defined over unqique combinations of individual characteristics and regions.
- bias_{inc} bias_j are the bias weights that ranges evenly from 0.5 to 3 across the 8 levels of each characteristic. For each characteristic, there is a positive correlation between the bias weight and WTP such that WTP is higher at higher levels of non-response bias.[14]
- β_{strength} is a bias strength parameter. If $\beta_{\text{strength}} = 0$, there is no non-response bias and this scenario is equivalent to random sampling (i.e. sampling only based on $weight_s$). As β_{strength} increases, there is stronger differential non-response.

[13] There is no non-response bias across regions but we ensure at least one individual is sampled in each region to ensure all regions are represented which is critical for regional estimation.

[14] For example, level 1 for age has a bias weight of 0.5 and a main effect on WTP of -25, whereas level 8 for age has a bias weight of 3 and a main effect on WTP of 25.

Survey response: Respondents to the survey answer a single binary choice question and respond truthfully comparing their own WTP to a randomly assigned bid level.[15] For each respondent, we collect their yes or no response to the valuation question, the bid level presented, the region where they live, and the characteristic level of respondents.

Survey adjustment methods: For each generated sample, we compare the performance of MRP against three other logit-based approaches.

1. Unweighted Logit Model (Unweighted): Estimates a logit model with a program constant and bid parameters with no adjustment for potential sample biases. The unweighted results provides a baseline and quantifies how the sampling bias translates to differences in sample and population WTP.

2. Weighted Logit Model (Weighted): Estimates a weighted logit model with a program constant and bid parameters and uses raked weights in estimation to match known population marginals across the individual characteristics and regions. The specific algorithm is iterative proportional fitting using the *anesrake* package in R (Pasek 2018).

3. Post-stratification with Fixed Effects Logit Model (Post-stratification): Estimates a logit model with a program constant, a bid parameter, and individual characteristics and regions included as fixed effect interaction terms with the program constant. WTP is then calculated for each subgroup using the relevant model parameters and aggregated to a population-level WTP using population weights following a similar approach as Equation 4.

Experimental scenarios To assess how well the models perform across multiple experimental factors, we vary the non-response bias strength (0, 0.5, 1.0) and change the sample size (500, 1000, 3000). We also consider two model misspecification scenarios including one where we only have access to *coarse income categories* and one involving *omitted variables*.

[15] There are 4 bid levels that are set at population WTP quantiles. Thus we abstract away from issues of bid selection.

We assess performance using two summary metrics: *Mean percent bias*, is the percentage difference between the estimated mean population-level WTP and the true value, and *Root Mean Square Error* is the square root of the difference between the estimated mean population-level WTP and the true value.[16] Because true WTP is always normalized to 1, the RMSE units can be interpreted as percentage terms (e.g. RMSE = 0.20 is equal to 20% average estimation error).

We use 100 Monte Carlo replications per scenario. For each replicate we 1) draw a sample from the population using Equation 5 and collect survey responses, 2) use the survey data to estimate WTP using all four models, and 3) calculate the two performance metrics. All of the non-MRP models are estimated using maximum likelihood estimation (MLE) and 100 bootstrap replications.

5 Experimental results

5.1 Survey bias and sample size

The first set of simulation results in Figure 2 shows how the models perform for various non-response bias levels and sample sizes. The left-hand panels show the results using a sample size of 1,000 and different non-response bias strength levels. Starting with the mean percent bias in the top-left panel, we see that all models provide reasonable estimates of WTP under the random sampling scenario as expected. As we move from left to right, the non-response bias strength increases as illustrated by the rising percent bias of the unweighted model WTP estimates at 41% when $\beta_{\text{strength}} = 0.5$ to 60% when $\beta_{\text{strength}} = 1.0$. Weighting helps mitigate the influence of sampling bias, but the weighted models struggle to recover the true WTP at higher levels of sampling bias and the mean percent bias remains around 25%. The mean percent bias for post-stratification with fixed parameters is 9% at moderate survey bias but

[16] We also calculated RMSE for each Monte Carlo replicate using the draws from the posterior distribution for the MRP model and bootstrapped sample replicates for the MLE models to assess the within replicate model variability. The results are qualitatively similar to the RMSE results across replicates so are not reported.

this increases to 28% at high bias levels ($\beta_{\text{strength}} = 1$) alongside substantial uncertainty around the estimated WTP given the large number of parameters. The MRP approach has the lowest estimated mean bias for all sample bias scenarios rising to 6% in the high bias scenario. The root mean squared error (RMSE) results in the bottom left panel, provides further insights into the variability of the bias and MRP performs well relative to the alternative adjustment models.

The right-hand side of Figure 2 presents the model performance metrics under different sample sizes compare the results across sample sizes of 500, 1000, and 3000 using moderate non-response bias (β_{strength}=0.5). The unweighted models shows that unadjusted sample bias remains similar across these sample sizes and also the variability decreases. The performance of all adjustment models improve as the sample size increases. The post-stratification with fixed parameter models are most affected by changing survey sample sizes with an mean percent bias of 15% with 500 respondents decreasing to 3% with 3000 survey respondents. The average RMSE of the post-stratification using 3000 respondents is 0.05 compared to a MRP average RMSE of 0.02. As sample size increases, the precision of the subgroup imputation by the post-stratification with fixed parameter models improves and the role of partial pooling in the MRP approach is lessened.

5.2 Model misspecification

Figure 3 presents the simulation results for the model misspecifications scenarios using a sample size of 1,000 and high survey bias strength (β_{strength}=1). The correct specification scenario in the middle is provided as the benchmark. The coarse income categories scenario investigates model performance where less detailed income information is collected from respondents and used for adjustments. In particular, the survey only collects income information in 3 broad categories rather than 8 as in the correct specification. Performance degrades for all adjustment models relative to the correct specification, but in different ways. For the MRP model, the performance degradation is through a systematic shift in the mean percent bias from 6% to

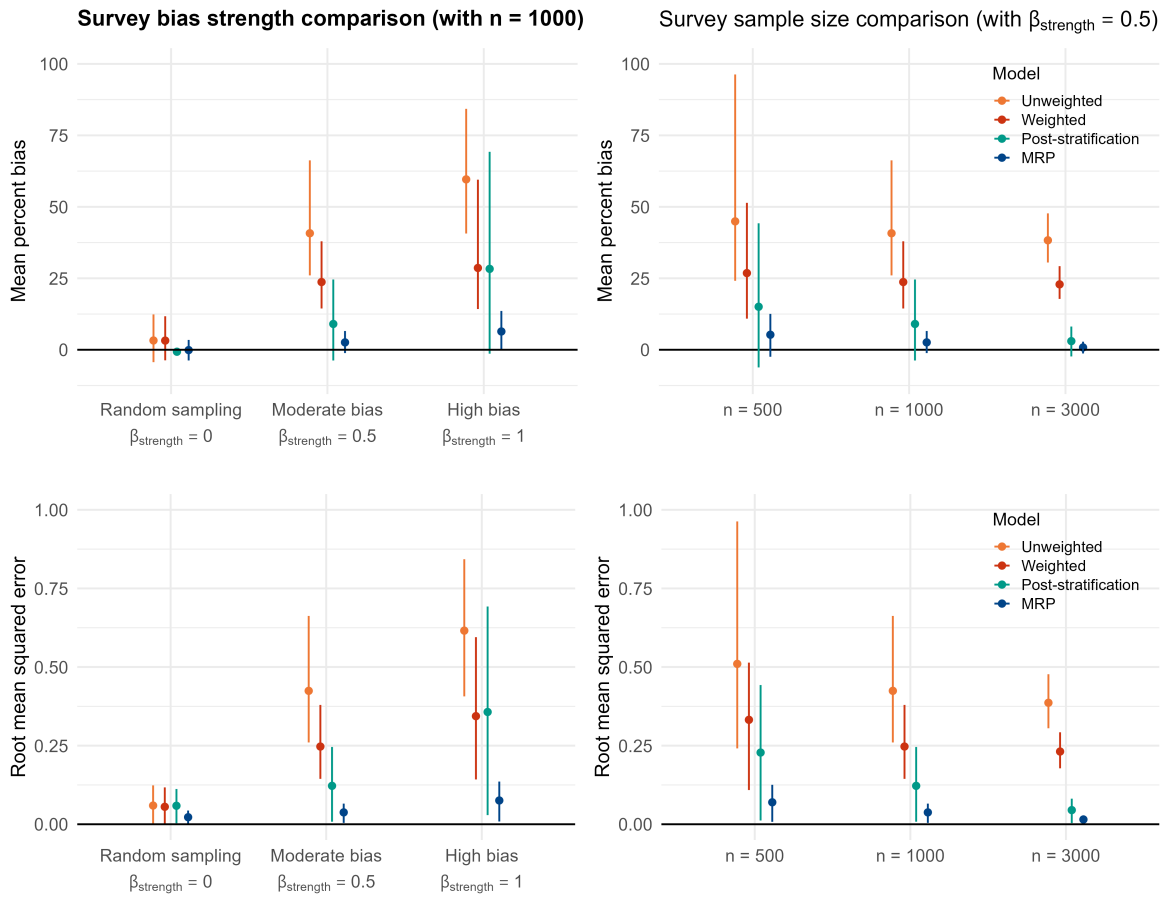

Figure 2: Experimental comparison of model performance under different survey bias and sample size scenarios

17% whereas the mean percent bias stays relatively similar for the post-stratification model at 28% but the variability in the mean percent bias increases dramatically. This finding reflects that MRP performs better when there are more granular characteristic information due to the ability to share preference information across levels. However, MRP remains the preferred model based on RMSE compared to weighting and post-stratification even with these coarse income categories.

The omitted variables scenario is where the adjustment models are misspecified due to omitting two individual characteristics (education and city size). The mean percent bias for MRP increases to 18% with omitted variables but this bias is still considerably less than the weighting (44%) and post-stratification (56%) approaches.

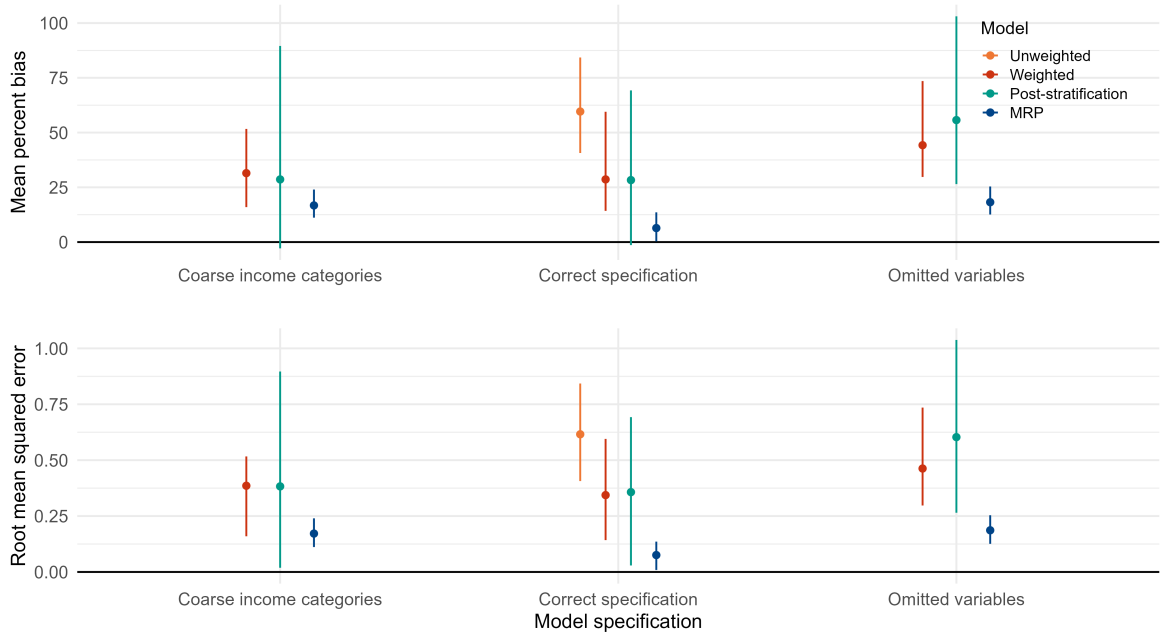

Figure 3: Experimental comparison of model performance under model misspecification scenarios with 1000 survey sample

5.3 Regional WTP estimation

The final set of simulations results uses a "small-area" estimation application to compare the performance of MRP and post-stratification with fixed parameters for estimating mean WTP in each region using a sample size of 1000. The top panels of Figure 3 shows the mean percent bias and RMSE for these models with random sampling (i.e. no bias). For regions with larger populations and hence well represented in the sample, the mean percent bias is minimal for both approaches. The bias increases and becomes more variable for regions with smaller population levels and this trend is especially stronger for the post-stratification models. Across all regions, the unweighted average RMSE is 0.11 for MRP and 0.67 for post-stratification.[17] These results show the potential benefits of MRP for estimating subgroup WTP even in the absence of any survey bias.

The bottom panel of Figure 3 show the same set of results with biased sampling ($\beta_{\text{strength}} = 1$). The patterns uncovered with random sampling are accentuated with increased variability in mean percent bias. With biased sampling, the average RMSE for MRP is 0.14 compared to 0.97 post-stratification models. As shown in Figure A1 of the appendix, increasing the survey sample size to 3000 reduces the relative outperformance of MRP as there is richer information for estimating the fixed effects models.

The experimental evidence in this section demonstrate the performance of MRP relative to competing approaches to addressing non-representative SP survey samples and providing robust estimates of subgroup WTP. The relative performance of MRP is especially pronounced using smaller sample sizes and in regional WTP estimation where partial pooling of preference information becomes more important which is a finding in earlier MRP Monte Carlo experiments in non-SP settings Downes and Carlin (2020b).

MRP empirical application

[17] The average RMSE metrics weighted by region population size for the random sampling scenario are 0.09 for MRP and 0.30 for post-stratification. These weighted average RMSE metrics for the survey bias scenario are 0.12 for MRP and 0.57 for post-stratification.

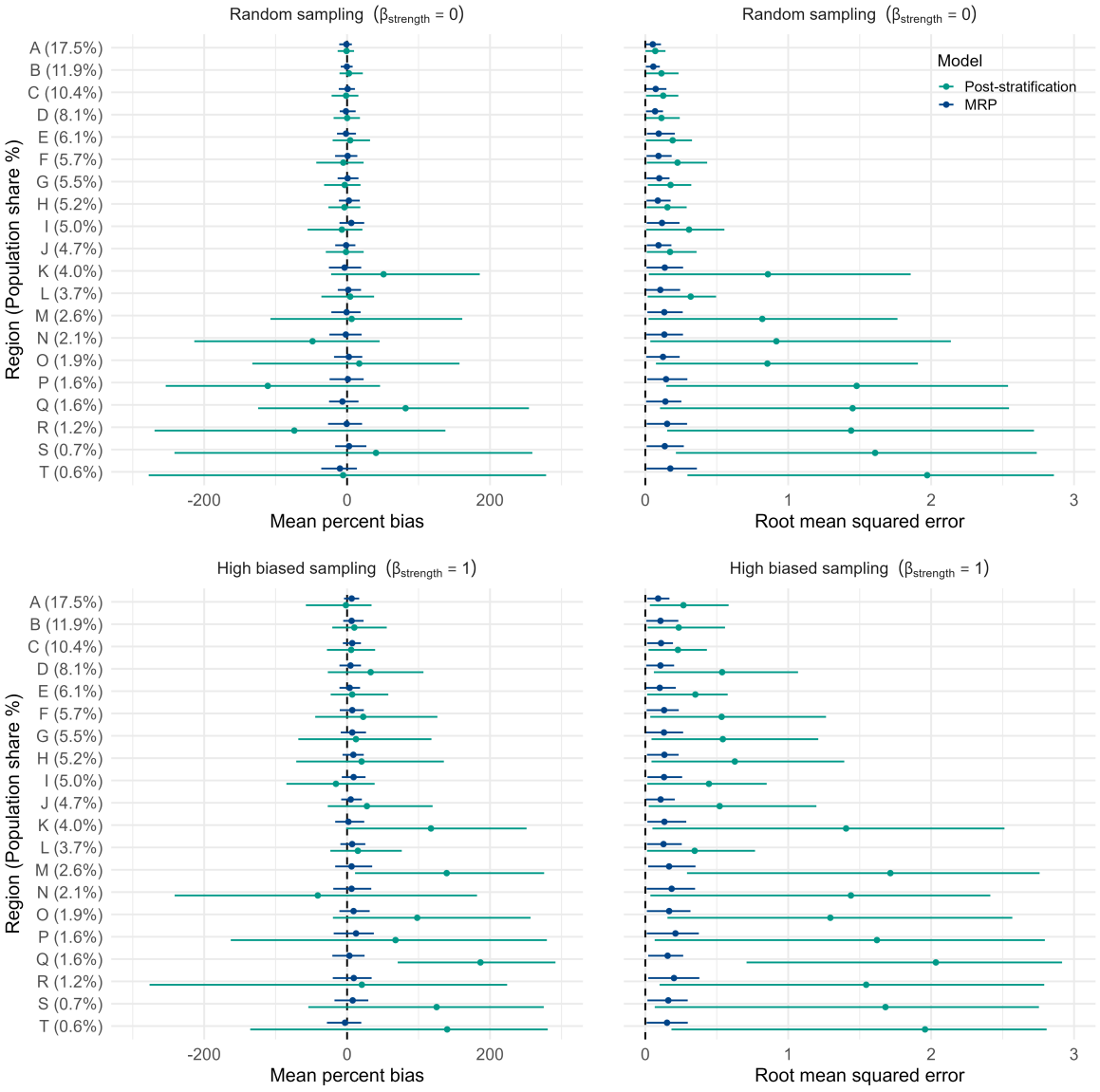

Figure 4: Experimental comparison of MRP and post-stratification for regional WTP estimation with 1000 survey sample

In this section we first describe the two stated preference surveys and the MRP discrete choice model specification employed for each dataset and then describe the post-stratification data sources and approach.[18]

Before describing each survey, we first discuss the similarities between them and highlight one important difference. The purpose of both surveys was to estimate WTP values for changes in environmental conditions in Canada within a public good context. The survey design process for each survey involved three focus groups and a pre-test.[19] The two surveys were administered online in Spring 2021 to respondents recruited from an opt-in internet panel using quota-based sampling. The main difference between the two surveys is their use of different value elicitation formats. The Saskatchewan River Delta (SRD) survey asked multiple valuation questions with each question containing three alternatives with varying attribute levels. The wetland survey used a single binary choice valuation question. Table 1 provides a summary of the two surveys and specific MRP implementations.

6 Wetland study

The purpose of the wetland study was to estimate WTP for a provincial wetland restoration program in the three Canadian prairie provinces. The target population was households living in the provinces of Alberta, Saskatchewan, and Manitoba. The quota-based sampling was done to ensure the respondents represented the population in terms of proportions male and three age categories (18-34 years old, 35-54, and 55+). A total of 1,999 completed surveys were obtained. The valuation question was a single binary choice referendum format using six randomly distributed bid levels ranging from $15 to $865. A text example of the valuation question is presented in Figure A2 and the full wetland survey instrument is included as Appendix B. Additional details on the study is provided in Boldt et al. (2026).

[18] All the data and code for these model implementations are available at https://github.com/plloydsmith/MRPandSP.

[19] The focus groups were conducted online due to the COVID pandemic in the Winter 2021 and included 4-8 participated each. Participants were recruited using random digit dialing from the target populations.

The MRP model specification for the wetland survey data is a binary logit model with the bid amount (p_j) and an alternative-specific constant for the restoration program ($ASC_{program}$) defined as

$$U_{ij} = \lambda(ASC_{program} - p_j) + \varepsilon_{ij} \tag{6}$$

The $ASC_{program}$ is specified using random parameters as

$$ASC_{program} = exp(\beta_0 + \alpha_{a,g[i]}^{age-gender} + \alpha_{p[i]}^{province} + \alpha_{m[i]}^{income} + \alpha_{e[i]}^{education}) \tag{7}$$

$$\begin{aligned}
\alpha_{a,g}^{age-gender} &\sim N(0, \sigma^2_{age-gender}), \quad \text{for } a = 1,2,3,\ g = 1,2 \\
\alpha_{p}^{province} &\sim N(0, \sigma^2_{province}), \quad \text{for } p = 1,2,3 \\
\alpha_{m}^{income} &\sim N(0, \sigma^2_{income}), \quad \text{for } m = 1,...,8 \\
\alpha_{e}^{education} &\sim N(0, \sigma^2_{education}), \quad \text{for } e = 1,...,5
\end{aligned}$$

where 'age-gender' is an interaction term between three age categories and two gender categories (i.e. whether respondent self identified as male or not). We force the ASC term to be positive through the use of an exponential function.[20] There are six parameters to estimate in this model specification including the scale (λ) and mean ASC parameter (β_0) and the four standard deviation parameters. We specify interactions between three age categories and two gender categories (male and non-male) for a total of six joint categories. The income variable includes seven income brackets and a missing category if respondents did not answer the question. The education group includes five categories ranging from "High school or less" to "Advanced degree".[21]

[20] In Figure A6 of Appendix A, we present the main results without exponentiation as a comparison and find similar results.

[21] The full set of income, age, and education categories are provided in Figure A4.

We estimate WTP using three separate provincial models and a combined prairie model.[22] For the combined prairie model, there are 630 subgroups to predict WTP in step 2 of the MRP workflow ($6_{age-gender} \times 3_{province} \times 7_{income} \times 5_{education}$ subgroups). We have respondents in 78% of the 630 subgroups and these represented subgroups make up 86% of the population. The welfare measure is the WTP for a provincial wetland restoration program to restore in the Prairie Pothole Region.[23]

7 Saskatchewan River Delta study

The purpose of the SRD study was to estimate the WTP to restore the ecological condition of a large freshwater inland delta located between the provinces of Saskatchewan and Manitoba. The target population was all Canadian households. Quota-based sampling was done to align the sample to be represented of the population in terms of proportions male, three age categories (18-34, 35-54, and 55+), and place of residence in each province or territory. A total of 949 respondents are used in the analysis after removing 16 identified yea-sayers.[24] The SRD survey used a 'typical' choice experiment set-up presenting respondents with a choice between the status-quo scenario and two alternative scenarios. The alternatives were described using four ecological attributes (lake sturgeon, waterfowl population, muskrat abundance, habitat in good ecological condition) and a cost attribute (an increase in income taxes for 20 years). Each person responded to six choice sets that were experimentally designed using a Bayesian and priors from preliminary models using a pre-test survey of 400 people. An example choice card is presented in Figure A3 and the full survey instrument is included as Appendix C. Additional details on the study is provided in Lika et al. (2025).

[22] The separate provincial models omit the province random parameter.

[23] Respondents were randomized into two different restoration levels: 10% or 20% of wetlands in the Prairie Pothole Region of their province. Preliminary models indicate that there is not a statistical difference in welfare measures across the two restoration levels. Thus reported WTP estimates are a weighted average across these two levels as in Vossler et al. (2023).

[24] Yea sayers were identified as respondents who always choose the most expensive option in the 6 choice tasks and agreed or strongly agreed with at least one of the following two statements: (i) "*It is important to conserve the delta, no matter how much it costs*" and (ii) "*The added cost I am willing to pay is to protect the environment in general and not just to protect the delta.*"

The MRP model specification for the SRD restoration survey data is a multinomial logit model with the bid amount (p_{jt}) and a vector of the 4 ecological attribute levels x_{jt} included as variables along with an ASC for a program alternative $ASC_{program}$ (i.e. non-status-quo alternative).

$$U_{ijt} = ASC_{program} + \lambda(\omega_i x_{jt} - p_{jt}) + \varepsilon_{ijt} \tag{8}$$

The vector ω_i for the 4 ecological attributes is specified using random parameters as

$$\omega_i = \beta_0 + \alpha_{a,g[i]}^{age-gender} + \alpha_{p[i]}^{province} + \alpha_{m[i]}^{income} + \alpha_{e[i]}^{education} \tag{9}$$

$$\begin{aligned}
\alpha_{a,g}^{age-gender} &\sim N(0, \sigma^2_{age-gender}), \quad \text{for } a = 1,..,5,\ g = 1,2 \\
\alpha_{m}^{income} &\sim N(0, \sigma^2_{income}), \quad \text{for } m = 1,...,7 \\
\alpha_{e}^{education} &\sim N(0, \sigma^2_{education}), \quad \text{for } e = 1,...,5 \\
\alpha_{p}^{province} &\sim N(0, \sigma^2_{province}), \quad \text{for } p = 1,...11
\end{aligned}$$

The age-gender interaction, income, and education random parameters are similar to the wetland survey specification, but there are now 5 age categories and 7 income categories (including one category for missing if the respondent did not answer the question). There are 11 province categories in the data but many of these provinces only have a handful of respondents given the quota-sampling strategy employed.[25] In total, there are 6 mean parameters and 16 standard deviation parameters to estimate in this MRP specification.

For the SRD study, there are 3,201 subgroups to impute WTP values ($10_{age-gender} \times 5_{education} \times 6_{income} \times 11_{province}$ minus 99 subgroups not in the post-stratification data).

[25] We have grouped the 1 respondent in each of the three northern Canadian territories into one 'province'.

There are respondents in only 18% of these 3,201 subgroups but these represented subgroups make up 57% of the total population. We calculate the compensation variation for a program that increases all four ecological attributes by 25% relative to their status quo levels and also calculate marginal WTP (MWTP) measures for a 1% increase in each of the ecological attributes.

Table 1: Summary of the two stated preference surveys and MRP implementation

	Wetland study	Saskatchewan River Delta study
Survey data		
Target population	Adults in 3 provinces	Adults in Canada
Sampling strategy	Quota-based sample: age-gender, province	age, gender, province
Respondent recruitment	Opt-in internet panel	Opt-in internet panel
Sample size	1999	965
Elicitation question	Single binary choice	Six choice sets with 3 alternatives and 4 non-price attributes
Model specification		
Model type	Binary logit	Multinomial logit
Number of mean parameters	2	4
Number of standard deviation parameters	4	16
Imputation of WTP		
Number of subgroups to impute WTP	900	3201
Percent of subgroups with respondents	78%	18%

Percent of population in represented subgroups	86%	57%

8 Post-stratification data

In addition to the survey data, implementing MRP also requires information on the target population to calculate subgroup counts (N_s) for use in Equation 4. The auxiliary data for the two studies comes from the 2021 Canadian Census Public Use Microdata File (PUMF). Similar microdata exists for many countries and the analogous data in the United States is the American Community Survey Public Use Microdata Sample. The PUMF contains information from 980,868 individuals representing 2.7% of the Canadian population and is a high quality sample as evidenced by unit population weights ranging from 36.92 to 37.07. Restricting the data to people aged 18 years and older, there is information on 777,421 Canadian adults. The PUMF file includes 126 individual variables which provides insights into the wide range of respondent characteristics that could potentially be included in the MRP modelling. For this analysis, we use a subset of the variable information for which we have corresponding survey data. We focus on the same five variables to calculate subgroup counts for both studies: gender, age, income, education, and province.

Three use cases of multi-level regression and postratification (MRP) for stated preference research

We frame the presentation of the application results around three beneficial MRP use cases of MRP to demonstrate specific way the approach can aid SP research.[26]

[26]Model diagnostics for all models and parameters are presented in Tables A4, A5, A6, and A7 of the Appendix.

9 Generating population-relevant estimates from non-representative samples

The first application demonstrates how MRP can help address the non-representativeness of survey samples to generate more population-relevant estimates. Starting with the SRD study, Figure 5 illustrates the differences between sample respondent characteristics and the target population using census data. The gender, age categories, and province sample shares closely mimic the population shares, which is not surprising as these were built into the quota-based sampling designs. However, when considering respondent characteristics that were not explicitly included in the quota-based sampling design, we observe stark differences between the sample income and education levels and the population. These differences motivate the use of the MRP approach.

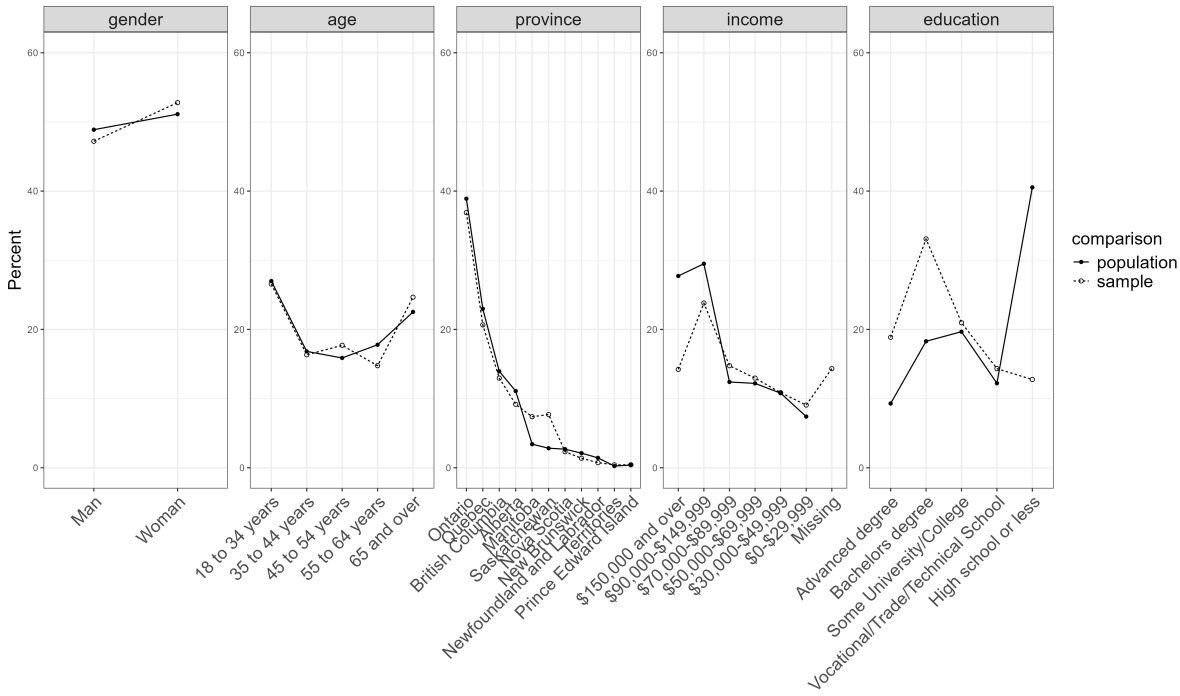

Figure 5: Comparison of Saskatchewan River Delta (SRD) survey sample and target population

Table 2 presents the WTP-space model parameter results for the MRP specification alongside sample-based multinomial logit (MNL) and random parameters logit (RPL).[27] Parameter means on attribute variables are interpreted as MWTP for a 0.1% increase in the levels. Parameters estimates for the attributes are fairly similar as no adjustments have been made for sample representativeness during Step 1 of the MRP approach.

[27]The RPL model specifies a normal distribution for all non-price attributes.

Table 2: WTP-space model estimates for a SRD restoration program using MNL, RPL, and MRP models

	MNL	RPL	MRP
Mean Sturgeon	0.176 (0.013)	0.189 (0.015)	0.153 (0.051)
Mean Habitat	0.225 (0.020)	0.224 (0.022)	0.203 (0.106)
Mean Waterfowl	0.164 (0.016)	0.175 (0.016)	0.143 (0.083)
Mean Muskrat	0.149 (0.016)	0.149 (0.018)	0.133 (0.083)
Scale (λ)	0.403 (0.018)	0.495 (0.023)	0.393 (0.018)
Program constant	0.180 (0.044)	1.497 (0.172)	0.175 (0.045)
SD Sturgeon: individual		0.201 (0.023)	
SD Habitat: individual		0.277 (0.037)	
SD Waterfowl: individual		0.071 (0.043)	
SD Muskrat: individual		0.218 (0.034)	
SD Program: individual		4.012 (0.211)	
SD Sturgeon: age-gender			0.030 (0.021)
SD Sturgeon: education			0.036 (0.038)
SD Sturgeon: income			0.022 (0.019)
SD Sturgeon: province			0.113 (0.047)
SD Habitat: age-gender			0.099 (0.038)
SD Habitat: education			0.165 (0.106)
SD Habitat: income			0.095 (0.048)
SD Habitat: province			0.051 (0.041)
SD Waterfowl: age-gender			0.035 (0.023)
SD Waterfowl: education			0.150 (0.089)
SD Waterfowl: income			0.021 (0.020)
SD Waterfowl: province			0.030 (0.027)
SD Muskrat: age-gender			0.057 (0.028)
SD Muskrat: education			0.145 (0.088)

	MNL	RPL	MRP
SD Muskrat: income			0.041 (0.033)
SD Muskrat: province			0.046 (0.029)
nObs	5694	5694	5694

The standard deviation of the posterior distribution for the parameter estimates are presented in parentheses. For the RPL model, the standard deviation (SD) parameters are specified at the individual level. For the MRP model, the standard deviation (SD) parameters are specified at the respondent characteristic level and there are 10 age-gender levels, 5 education group levels, 7 income levels, and 11 province levels. Parameter means on attribute variables are interpreted as marginal WTP for a 0.1% increase. Parameter diagnostic metrics for R-hat and effective sample size are provided in Table A4.

Figure 6 shows the posterior distribution of estimated WTP for a restoration program that improves all four SRD attribute levels by 25%. The estimated sample average WTP is \$179 using a MNL specification and \$184 using a RPL specification. The mean WTP is estimated to be \$148 using the MRP approach which is 20% lower compared to the MNL or RPL approaches. Results of Poe combinatorial tests comparing these distributions find statistically significant differences (Poe et al., 2005).[28] Besides comparing the averages, the interquartile range of the posterior distribution is higher using the MRP approach as the model accounts for both estimation and post-stratification adjustment uncertainty.

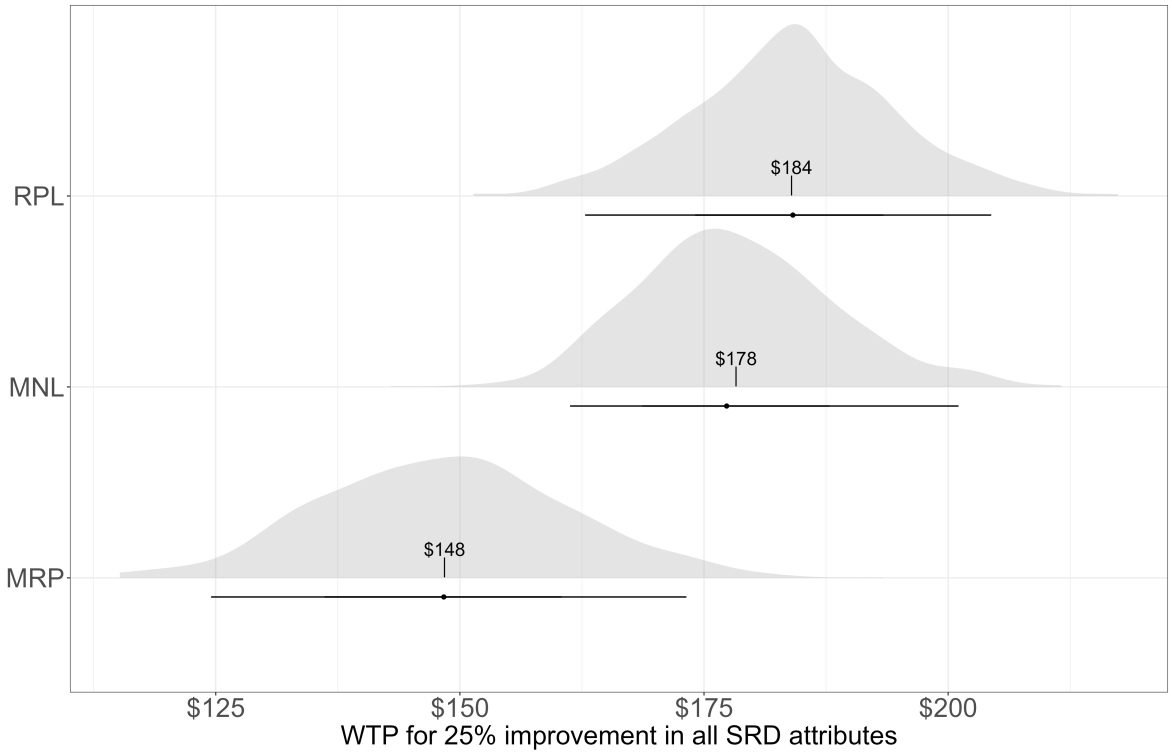

Figure 6: Comparison of WTP estimates for a SRD restoration program using MNL, RPL, and MRP models

To investigate whether the MRP and RPL differences in WTP values vary across specific attributes, Figure 7 presents the mean estimated marginal WTP for a 1% increase in each of

[28] In particular, the posterior probability that the WTP distribution of the MRP model is greater than the WTP of the MNL model is 3% and this posterior probability is 1% comparing the MRP and RPL model's WTP. The full set of Poe test results for all model comparisons is provided in Appendix A.

the four environmental attributes levels. While the RPL model estimates a higher MWTP for each attribute, the differences in MWTP between MRP and RPL vary considerably, ranging from a 5% difference for lake sturgeon to a 57% difference for waterfowl. Using the Poe test to compare distributions, the probability that the MRP model produces a higher WTP is 37% and 32% for lake sturgeon and and good ecological condition but substantially less at 3% and less 1% for muskrat abundance and waterfowl population. This provides evidence that the difference between sample and population estimates may be specific to the environmental change under consideration.

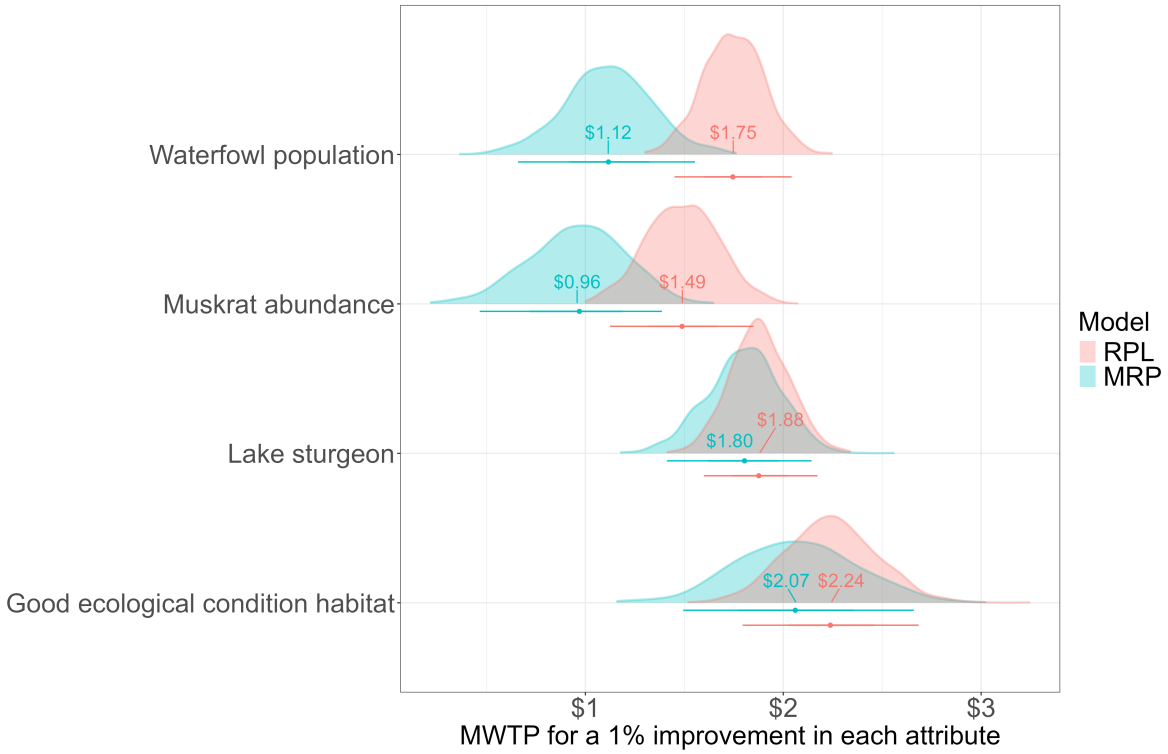

Figure 7: Estimated mean MWTP for a 1% improvement in each SRD environmental attribute

Turning to the wetland study, Figure A4 shows the comparison across all three Prairie provinces for the survey sample and the target population using census data. Similar figures for each province are provided in Figure A5. The marginal distribution of gender and age are quite similar across the sample and population, as in the SRD study. However, important differences

emerge when we consider a broader set of characteristics as the sample is more educated with substantially less people with "High school or less" compared to the population.

Figure 8 shows the estimated posterior distribution of mean WTP for a wetland restoration program in each of the three provinces as well as a combined Prairie model estimate for the MRP and a sample-based logit model approach. The mean WTP using MRP is lower than the sample logit model for Alberta and Manitoba, but higher for Saskatchewan. Furthermore, there is substantial overlap in the estimated posterior distributions across the two models. The combined Prairie model mean WTP estimate is $399 using the logit model and $362 using MRP. Poe tests comparing distributions confirms that there is little evidence to suggest that the models generate different WTP distributions.

Taken together, these sets of comparisons show that MRP can result in lower or higher WTP estimates compared to models using sample information only, MRP differences can vary by specific attributes, and single binary choice data may not provide enough information on how preferences vary across people for MRP to address sample representation issues. We discuss this last point further in the conclusion section.

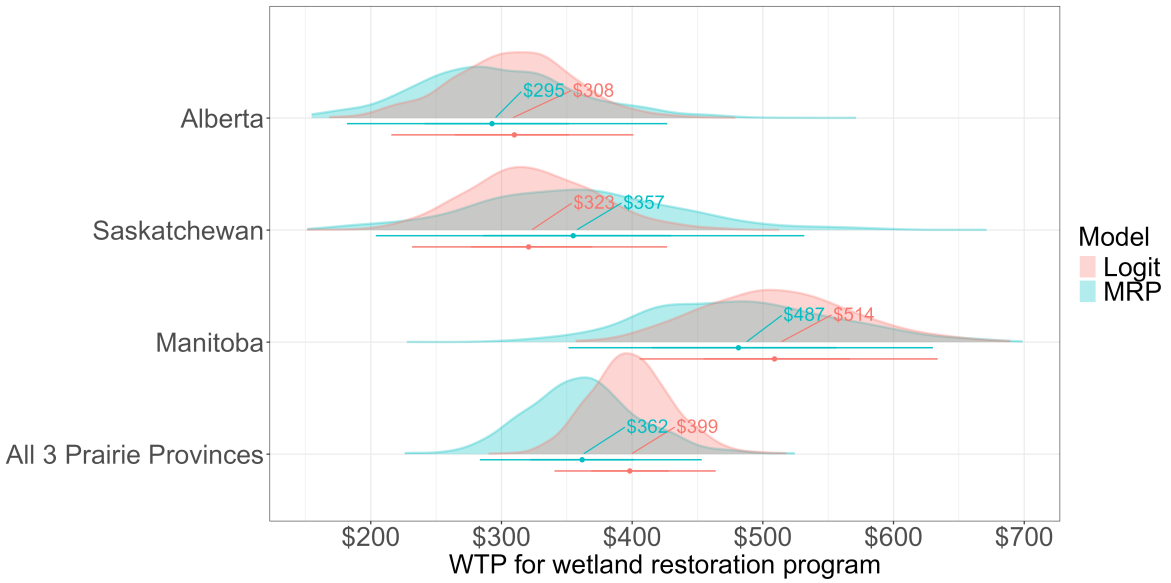

Figure 8: Estimated mean WTP for a provincial wetland restoration program across logit and MRP specifications

10 Estimating preference heterogeneity and distributional impacts across people

The second application of MRP for SP research is to provide an alternative approach for assessing preference heterogeneity and estimating distributional impacts across people. This use extends the notion of "small-area estimation" (i.e. using national surveys to infer local preferences) in the MRP literature to the common SP context of using general population surveys to estimate subgroup preference information and builds on the regional results presented in the Monte Carlo simulation section.

The MRP approach combines the regularization benefits of random parameters, which mitigates overfitting concerns, with the partial pooling of preference information across people to estimate more reasonable WTP estimates. This approach is particularly promising when specific subgroups of interest are poorly, or not at all, represented in the survey data. Furthermore, the MRP approach facilitates investigations of population heterogeneity, as opposed to merely sample heterogeneity, through its post-stratification step.

The top distribution of Figure 9 shows the individual-level mean WTP for the 25% restoration program in the SRD using the RPL model. The bottom distribution of Figure 9 shows the distribution of WTP for the same program using the MRP approach for each of the 3,201 subgroups weighted to reflect to their population shares. Thus, while the RPL model shows the *sample* WTP distribution, the MRP results show the *population* WTP distribution.[29]

Each of these imputed 3,201 subgroup WTP estimates from the MRP approach come with their own uncertainty and Figure 10 shows the posterior distribution for two of these subgroups. The subgroup in the province of British Columbia has a mean WTP of $330 compared to a mean WTP of $5 for someone in the province of Newfoundland and Labrador. Both estimates

[29]The MRP approach also provides more reasonable bounds on the WTP distribution across the population with over 99% of subgroup WTP estimates being between $0 and $325 (i.e. the highest cost level shown in the survey) compared to the RPL model where 88% of individual estimates are between these bounds. This comparison masks the differences in distributions due to combining four uncorrelated random parameters for each attribute. Figure A7 in the appendix shows the same figure for each attribute which highlights the wide bounds for the RPL individual-level estimates compared to the MRP approach.

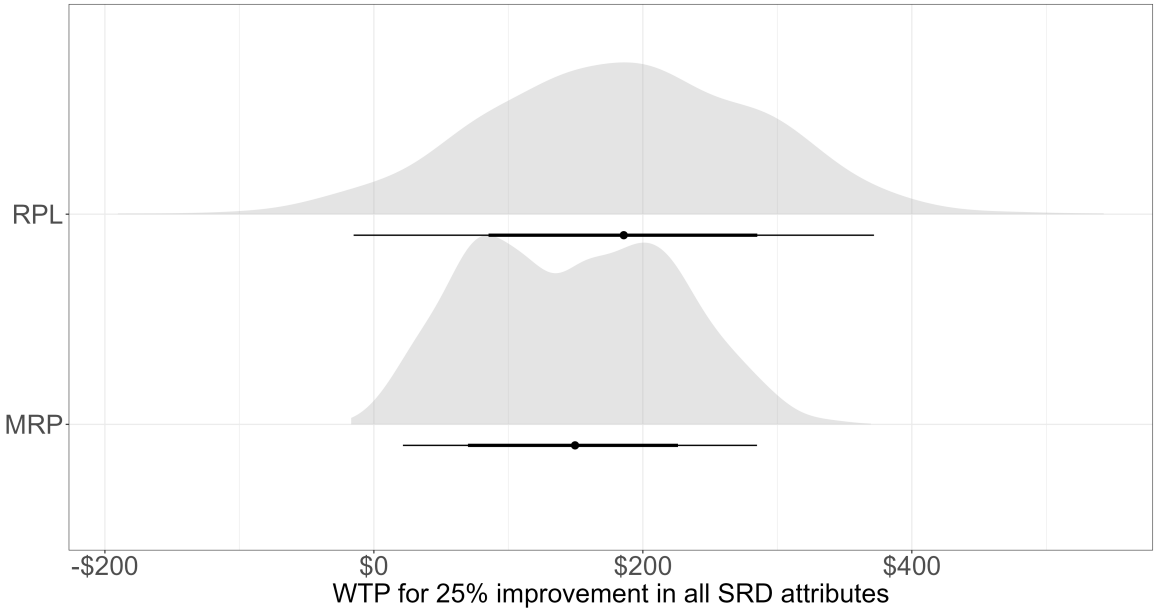

Figure 9: Distribution of WTP for a SRD restoration program using the MRP and RPL approach

reflect the use of partial pooling as while the sample sample contains a person that fits British Columbian subgroup definition, the Newfoundlander subgroup is not represented in the data.

We also demonstrate how MRP provides a natural approach for estimating welfare measures at different levels of aggregation between the overall population and the 3,201 subgroups, focusing on the subnational level and across the educational categories. As we divide our sample into different subgroups, sample representativeness issues are more pronounced. For instance, the representativness of respondents living in British Columbia within the sample is weaker compared to the national-level representativeness due to the small sample sizes and the sampling design. In Figure 11, the mean WTP estimates for the 25% restoration program in the SRD for each province and the territories. These provincial WTP estimates, derived from weighted provincial subgroups, are not 'holding everything else constant' as would be the case in interpreting a set of provincial interaction terms, but rather reflect the unique population characteristics (e.g. income differences, education, age) of each province. Provinces with larger respondent samples, such as Ontario and Quebec, yield tighter posterior distributions, whereas

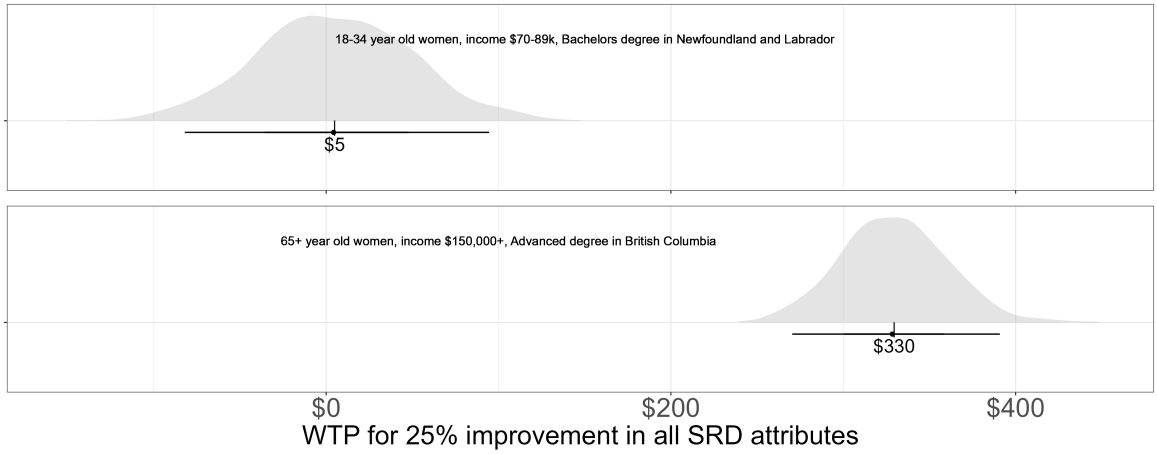

Figure 10: Subgroup-level WTP estimates using the MRP approach

the WTP estimate for the territories, with only a handful of respondents, exhibit greater uncertainty.

Figure 12 presents the estimated mean WTP for the same SRD restoration program by educational categories. We find evidence of a clear education gradient with more educated individuals having a higher WTP (e.g. $257 for people with advanced degrees) compared to lower education individuals (e.g. $81 for people with high school or less). As in the provincial estimates, these are not partial estimates holding all other respondent characteristics "fixed" but rather reflect the average WTP for individuals within each educational category.

11 Assessing impacts of anomalous responses using consistent target population-relevant estimates

The third beneficial use case of MRP for SP research is assessing the impacts removing potentially anomalous responses using consistent target population-relevant estimates. A common consideration in SP research is the exclusion of respondents identified as having provided potentially anomalous responses to the key valuation question(s) (Johnston et al. 2017; Poe

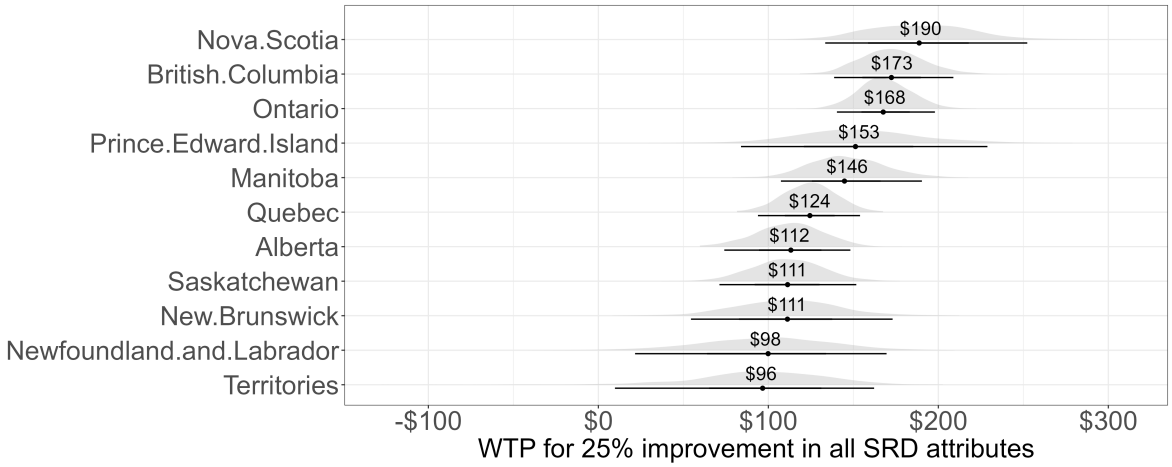

Figure 11: Mean WTP for a SRD restoration program by province

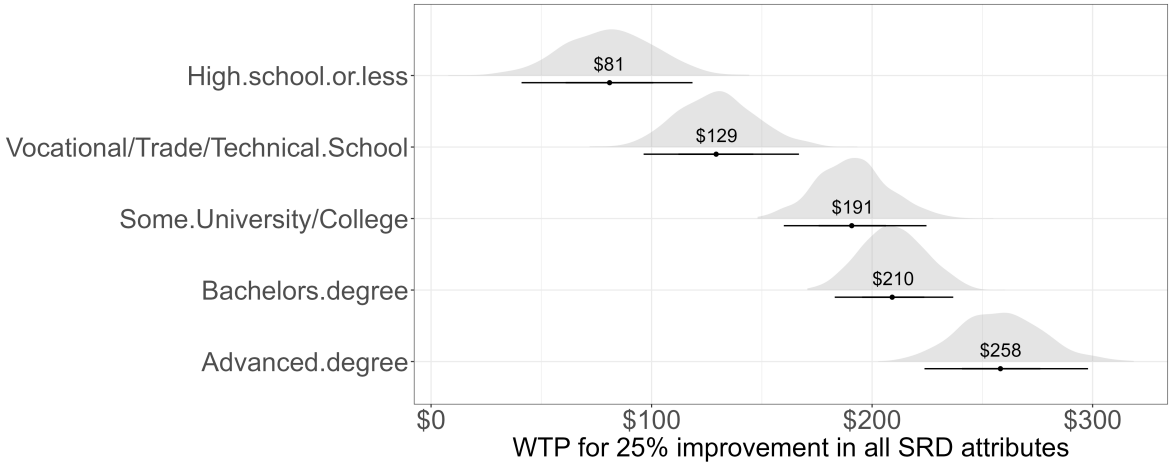

Figure 12: Mean WTP for a SRD restoration program by education level

2016). These respondents with monikers such as "yea-sayers", "scenario-rejectors", "speedsters", "straight-liners", and "protest votes" are typically identified through follow-up questions, trap questions, the speed they do the survey, straight line choices in repeated questions, or some combination of these signals. Excluding such respondents to improve the validity of responses may inadvertently exacerbate representation errors if these respondents are not randomly distributed across the sample. Due to the post-stratification step, the MRP approach maintains a consistent target population for deriving welfare measures, regardless of the degree of non-representativeness in the various subsamples used for analysis.

To demonstrate this MRP use case, we use the wetland survey data and the MRP approach to compare WTP estimates using the full sample to WTP estimates using a restricted 'validity' sample, which excludes potentially anomalous responses. To improve the validity of welfare estimates, we purposely choose quite restrictive definitions and remove identified 'yea-sayers' and respondents who perceived the survey as inconsequential.[30] After these restrictions, the overall sample size drops from 1,999 in the full sample to 1,127 in the restricted sample.[31] This reduction in sample sizes is admittedly quite substantial and the focus here is to illustrate the approach rather than take a definite stand on what is or is not a valid response.

We first assess the representativeness effects of using the validity sample and then present the MRP WTP result comparisons. Figure A8 compares the representativeness of the full sample, the validity sample, and the target population across the three Prairie provinces. Analogous province-specific comparisons are provided in Figure A9. Across the respondent characteristics, the validity sample is generally more unrepresentative of the target population than the full sample. This finding provides some empirical evidence that validity checks that

[30] Yea-sayers were identified as voting yes to a restoration program and strongly disagreeing with this statement "*I voted as if my household would actually face the costs shown*" or strongly agreeing with "*It is important to restore wetlands in the Canadian prairies no matter how high the cost*". Inconsequential respondents were identified if they said that the government is "unlikely" or "very unlikely" to take the survey votes into account. The specific text of the consequentiality beliefs question was: *When the [$province] government decides whether or not to implement the proposed plan you just voted on, how likely do you think it is that the provincial government will take into account your vote and that of the other respondents to this study in its decision-making?*

[31] In each province, the change in sample sizes are 870 to 526 in Alberta, 811 to 459 in Manitoba, and 318 to 192 in Saskatchewan.

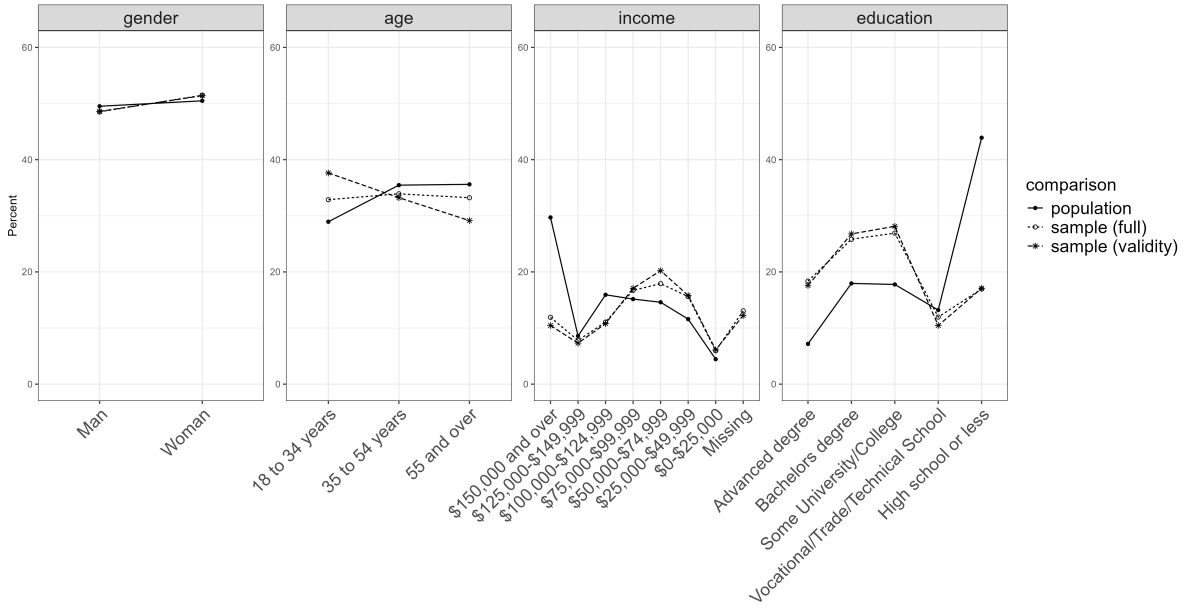

Figure 13: Comparison of wetland study sample, validity sample, and population across all three Priarie provinces

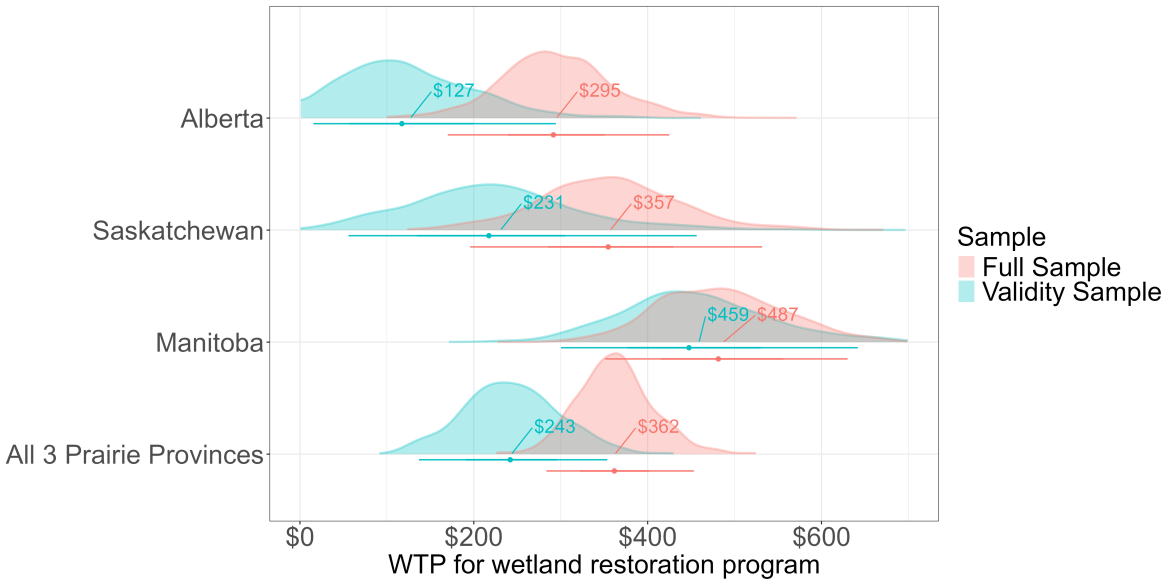

Figure 14: Comparing the impacts of removing invalid responses using consistent population-estimate frames

improve the validity of responses by excluding certain anomalous responses can exacerbate representativeness issues.

Figure 14 presents the estimated MRP WTP distributions for a wetland restoration program across the three provinces using the full sample and the validity sample.[32] The mean WTP estimates are lower using the validity sample compared to the full sample. Excluding 'yeasayers' necessarily decreases WTP as only yes votes are removed but the effect on WTP of excluding respondents with inconsequential beliefs is less clear a priori but there is empirical evidence that people with consequential beliefs also have higher WTP (Herriges et al. 2010). Boldt et al. (2026) uses the same wetland survey data and finds that people with consequential beliefs have a higher WTP. The important point for the comparison is that the post-stratification target population frame is consistent across the two MRP estimates even though the sample used for analysis is different. Thus, we can be sure that these differences are not due to differences in subsample respondent characteristic composition that are identified in Figure 13. The MRP approach prevents the confounding of assessing the impacts of improving sample validity while sample representativeness also changes if the post-stratification step is not conducted.

Conclusion

This paper introduces the multi-level regression and post-stratification (MRP) modelling workflow to improve welfare measurement using SP data. Through simulation experiments and survey data, we demonstrate how MRP can support SP researchers in generating population-relevant estimates from non-representative samples, estimating preference heterogeneity and the impacts across subpopulations, and assessing the implications of removing anomalous responses using consistent population frame. The MRP approach combines the regularization benefits of random parameters, which mitigates overfitting concerns, with the partial pooling of preference information across people by modelling group-level rather than individual-level preferences to estimate more reasonable WTP estimates. Because heterogeneity is specified at the respondent characteristic level, the MRP approach is more widely applicable than RPL and

[32] Appendix A provides an example of how to include region-level (i.e. Level 2) variables.

LC models to preference sparse settings such as data with smaller survey sample sizes, single binary choice data, or when specific subgroups of interest are poorly, or not at all, represented in the survey data. Furthermore, the MRP approach facilitates investigations of population heterogeneity, as opposed to merely sample heterogeneity, through its post-stratification step. In MRP complements existing approaches to assessing preference heterogeneity at the individual-level such as RPL and LC models with a focus on putting the heterogeneity "to work" to improve the population relevance of estimates.

The MRP approach has broader implications for SP survey design and administration. We outline three key considerations here. First, the MRP approach provides an additional rationale for using increased sample sizes in SP research as additional data is useful for robustly estimating subgroup welfare measures that underpin generalizations to population-relevant estimates. Second, the MRP approach provides further justification for collecting more detailed preference information per respondent than a SBC question. While SBC questions enhance validity, they reveal less about preference variability within the sample, which is critical for robust subgroup WTP imputation and the post-stratification steps in MRP. By contrast, standard 'choice experiment' format with multiple questions and three alternative offer richer insights into preference heterogeneity, but raise validity concerns (Johnston et al. 2017; Meginnis et al. 2021). Repeated binary choice elicitation formats may strike an optimal balance between the two for addressing both measurement and representation errors (Petrolia and Interis 2013). Third, incorporating questions about respondent characteristics that are predictive of WTP and can be readily matched to auxiliary population information broadens the range of variables that can be integrated into MRP models.

While MRP has promise as a useful tool for SP research, it also has several limitations. Most notably, MRP addresses observed differences between sample and population but does not address systematic unobserved differences. While MRP can mitigate some biases in the sample, it relies on the strong assumption that there is ignorable nonresponse within subgroup, that is, each subgroup's estimated WTP represents its true WTP. Further research is needed to address unobservable non-response bias and selection effects (Bailey 2024; Johnston and Abdulrahman

2017). Another challenge is in choosing the number of subgroups to impute as there is a tradeoff between having enough imputed subgroups to capture preference heterogeneity in the population while also ensuring robust predictions. MRP works well when the choice model estimated in the first step is able to robustly predict subgroup welfare but this becomes more challenging as the subgroup count increases as less represented subgroups necessarily rely on relatively more preference sharing from other subgroups. While that is a feature of the partial pooling embedded in the MRP approach and will be reflected in larger uncertainty over the prediction, the sharing of information also carries the assumption that observed characteristics influence preferences similar across subgroups.[33] Finally, MRP requires post-stratification population data for the respondent characteristics used in adjustment. This auxiliary data may not always be available and especially for some of the key determinants of preference heterogeneity.

This paper is an initial exploration of MRP's potential in non-market valuation research, leaving several avenues for future investigations. Beyond SP applications, MRP shows promise in revealed preference contexts, such as recreation demand models, where sample representativeness concerns arise - e.g. in citizen science (Cameron and Kolstoe 2021; Jayalath, Lloyd-Smith, and Becker 2023) or mobile device tracking datasets. The empirical applications used standard census variables, yet there are opportunities to expand the set of included respondent characteristics to include non-census variables such as attitude, belief, and voting behavior variables, provided these are available at the target population level.[34] Future research could also advance more complex MRP model specifications such as integrating LC logit specifications or individual-level random parameters, including region-level variables to better impute regional WTP values[35], introducing correlation among random parameters, and using machine learning techniques to enhance variable selection in the model stage (Broniecki, Leemann, and Wüest 2022). One particular promising avenue is using structured priors on random parameters for

[33] For unrepresented subgroups (i.e. subgroups with no survey respondents), their WTP is being completely determined by other people.

[34] Relaxing the data demands of MRP, the multilevel regression with synthetic post-stratification (MrsP) approach allows the one to combine several population datasets using marginal distributions (Leemann and Wasserfallen 2017)

[35] Appendix A provides an example of how to include region-level (i.e. Level 2) variables.

ordered predictors such as education level to better reflect the relationship between respondent characteristics and preferences (Gao et al. 2021). The MRP approach also blurs the lines between a primary valuation and benefit transfer study by imputing WTP for subgroups absent (or severely underrepresented) in the data. In this way, MRP resembles function-based transfer exercise but focusing on extrapolating from sample to population rather than from study to policy site. Future research could evaluate the accuracy of MRP out-of-sample extrapolations.

References

Arrow, Kenneth, Robert Solow, Paul Portney, Edward Leamer, Roy Radner, and Howard Shuman. 1993. "Report of the NOAA Panel on Contingent Valuation."

Bailey, Michael A. 2024. *Polling at a Crossroads: Rethinking Modern Survey Research*. Methodological Tools in the Social Sciences. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108697798.

Baker, Reg, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista J. Gile, and Roger Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1 (2): 90–143. https://doi.org/10.1093/jssam/smt008.

Banzhaf, H. Spencer. 2017. "Constructing Markets: Environmental Economics and the Contingent Valuation Controversy." *History of Political Economy* 49 (Supplement): 213–39. https://doi.org/10.1215/00182702-4166335.

Bishop, Richard C., Kevin J. Boyle, Richard T. Carson, David Chapman, W. Michael Hanemann, Barbara Kanninen, Raymond J. Kopp, et al. 2017. "Putting a Value on Injuries to Natural Assets: The BP Oil Spill." *Science* 356 (6335): 253–54. https://doi.org/10.1126/science.aam8124.

Boldt, Liam, Patrick Lloyd-Smith, Ken Belcher, and Peter Boxall. 2026. "Public Willingness to Pay for Wetland Restoration in the Canadian Prairie Pothole Region." *Canadian Journal of Agricultural Economics/Revue Canadienne d'agroeconomie*. https://doi.org/10.1111/cjag.70010.

Bonnichsen, Ole, and Søren Bøye Olsen. 2016. "Correcting for Non-Response Bias in Contingent Valuation Surveys Concerning Environmental Non-Market Goods: An Empirical Investigation Using an Online Panel." *Journal of Environmental Planning and Management* 59 (2): 245–62. https://doi.org/10.1080/09640568.2015.1008626.

Boyle, Kevin J., Mark Morrison, Darla Hatton MacDonald, Roderick Duncan, and John Rose. 2016. "Investigating Internet and Mail Implementation of Stated-Preference Surveys While Controlling for Differences in Sample Frames." *Environmental and Resource Economics* 64 (3): 401–19. https://doi.org/10.1007/s10640-015-9876-2.

Broniecki, Philipp, Lucas Leemann, and Reto Wüest. 2022. "Improved Multilevel Regression with Poststratification Through Machine Learning (autoMrP)." *The Journal of Politics* 84 (1): 597–601. https://doi.org/10.1086/714777.

Bürkner, Paul-Christian. 2017. "Brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (August): 1–28. https://doi.org/10.18637/jss.v080.i01.

Cameron, Trudy Ann, and Sonja H. Kolstoe. 2021. "Using Auxiliary Population Samples for Sample-Selection Correction in Models Based on Crowd-Sourced Volunteered Geographic Information." *Land Economics*, October. https://doi.org/10.3368/le.98.1.040720-0050R1.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles* 76 (1): 132. https://doi.org/10.18637/jss.v076.i01.

Carson, Richard T., and Theodore Groves. 2007. "Incentive and Informational Properties of Preference Questions." *Environmental and Resource Economics* 37 (1): 181–210. https://doi.org/10.1007/s10640-007-9124-5.

Carson, Richard T., Theodore Groves, and John List. 2014. "Consequentiality: A Theoretical and Experimental Exploration of a Single Binary Choice." *Journal of the Association of Environmental and Resource Economists* 1 (1): 171–207. http://econpapers.repec.org/article/ucpjaerec/doi_3a10.1086_2f676450.htm.

Cohn, Nate. 2024. "Trump Leads in 5 Key States, as Young and Nonwhite Voters Express

Discontent with Biden." *The New York Times*, May. https://www.nytimes.com/2024/05/13/us/politics/biden-trump-battleground-poll.html.

Downes, Marnie, and John B. Carlin. 2020a. "Multilevel Regression and Poststratification as a Modeling Approach for Estimating Population Quantities in Large Population Health Studies: A Simulation Study." *Biometrical Journal* 62 (2): 479–91. https://doi.org/10.1002/bimj.201900023.

———. 2020b. "Multilevel Regression and Poststratification Versus Survey Sample Weighting for Estimating Population Quantities in Large Population Health Studies." *American Journal of Epidemiology* 189 (7): 717–25. https://doi.org/10.1093/aje/kwaa053.

Gao, Yuxiang, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. 2021. "Improving Multilevel Regression and Poststratification with Structured Priors." *Bayesian Analysis* 16 (3): 719–44. https://doi.org/10.1214/20-BA1223.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)."

Gelman, Andrew, and Thomas Little. 1997. "Postratification into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology*, no. 2: 127–35. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X19970023616.

Glenk, Klaus, Robert J. Johnston, Jürgen Meyerhoff, and Julian Sagebiel. 2020. "Spatial Dimensions of Stated Preference Valuation in Environmental and Resource Economics: Methods, Trends and Challenges." *Environmental and Resource Economics* 75 (2): 215–42. https://doi.org/10.1007/s10640-018-00311-w.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. John Wiley & Sons.

Hausman, Jerry. 2012. "Contingent Valuation: From Dubious to Hopeless." *Journal of Economic Perspectives* 26 (4): 43–56. https://doi.org/10.1257/jep.26.4.43.

Herriges, Joseph, Catherine Kling, Chih-Chen Liu, and Justin Tobias. 2010. "What Are the Consequences of Consequentiality?" *Journal of Environmental Economics and Management* 59 (1): 67–81. https://doi.org/https://doi.org/10.1016/j.jeem.2009.03.004.

Hoffman, Matthew D., and Andrew Gelman. 2014. "The No-u-Turn Sampler: Adaptively

Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15: 1593–623.

Jayalath, Tharaka A., Patrick Lloyd-Smith, and Marcus Becker. 2023. "Biodiversity Benefits of Birdwatching Using Citizen Science Data and Individualized Recreational Demand Models." *Environmental and Resource Economics* 86 (1): 83–107. https://doi.org/10.1007/s10640-023-00788-0.

Johnston, Robert J. 2006. "Is Hypothetical Bias Universal? Validating Contingent Valuation Responses Using a Binding Public Referendum." *Journal of Environmental Economics and Management* 52 (1): 469–81. https://doi.org/10.1016/j.jeem.2005.12.003.

Johnston, Robert J., and Abdulallah S. Abdulrahman. 2017. "Systematic Non-Response in Discrete Choice Experiments: Implications for the Valuation of Climate Risk Reductions." *Journal of Environmental Economics and Policy* 6 (3): 246–67. https://doi.org/10.1080/21606544.2017.1284695.

Johnston, Robert J., Kevin J. Boyle, Wiktor (Vic) Adamowicz, Jeff Bennett, Roy Brouwer, Trudy Ann Cameron, W. Michael Hanemann, et al. 2017. "Contemporary Guidance for Stated Preference Studies." *Journal of the Association of Environmental and Resource Economists* 4 (2): 319–405. https://doi.org/10.1086/691697.

Keeter, Scott, and Leah Christian. 2012. "A Comparison of Results from Surveys by the Pew Research Center and Google Consumer Surveys." *Washington, DC: Pew Research Center*.

Keeter, Scott, N Hatley, C Kennedy, and Arnold Lau. 2017. "What Low Response Rates Mean for Telephone Surveys." https://www.pewresearch.org/methods/2017/05/15/what-low- response-rates-mean-for-telephone-surveys/ accessed July 8, 2024.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, et al. 2018. "An Evaluation of the 2016 Election Polls in the United States." *Public Opinion Quarterly* 82 (1): 1–33. https://doi.org/10.1093/poq/nfx047.

Kling, Catherine L., Daniel J. Phaneuf, and Jinhua Zhao. 2012. "From Exxon to BP: Has Some Number Become Better Than No Number?" *Journal of Economic Perspectives* 26 (4): 3–26. https://doi.org/10.1257/jep.26.4.3.

Kolstoe, Sonja, Brian Vander Naald, and Alison Cohan. 2022. "A Tale of Two Samples: Understanding WTP Differences in the Age of Social Media." *Ecosystem Services* 55 (June): 101420. https://doi.org/10.1016/j.ecoser.2022.101420.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science* 53 (1): 107–21. https://doi.org/10.1111/j.1540-5907.2008.00360.x.

Leemann, Lucas, and Fabio Wasserfallen. 2017. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation." *American Journal of Political Science* 61 (4): 1003–22. https://doi.org/10.1111/ajps.12319.

———. 2020. "Measuring Attitudes – Multilevel Modeling with Post-Stratification (MrP)." In, 371–84. SAGE Publications Ltd. https://doi.org/10.4135/9781526486387.

Lika, Elisabeta, Ken Belcher, Tim Jardine, Sabine Liebenehm, Patrick Lloyd-Smith, and Graham Strickert. 2025. "The Economic Value of Improving the Ecological Condition of the Saskatchewan River Delta, Canada." *Ecosystem Services* 75: 101763. https://doi.org/https://doi.org/10.1016/j.ecoser.2025.101763.

Lindhjem, Henrik, and Ståle Navrud. 2011. "(PDF) Using Internet in Stated Preference Surveys: A Review and Comparison of Survey Modes." *International Review of Environmental and Resource Economics* 54 (4): 30951. https://doi.org/10.1561/101.00000045.

Maas, Harro, and Andrej Svorenčík. 2017. ""Fraught with Controversy": Organizing Expertise Against Contingent Valuation." *History of Political Economy* 49 (2): 315–45. https://doi.org/10.1215/00182702-3876493.

Macaskill, James, and Patrick Lloyd-Smith. 2022. "Six Decades of Environmental Resource Valuation in Canada: A Synthesis of the Literature." *Canadian Journal of Agricultural Economics/Revue Canadienne d'agroeconomie* 70 (1): 73–89. https://doi.org/10.1111/cjag.12304.

McFadden, Daniel, and Kenneth Train. 2017. *Contingent Valuation of Environmental Goods: A Comprehensive Critique*. Cheltenham, UK: Edward Elgar Publishing. http://www.e-elgar.com/shop/contingent-valuation-of-environmental-goods.

Meginnis, Keila, Michael Burton, Ron Chan, and Dan Rigby. 2021. "Strategic Bias in Dis-

crete Choice Experiments." *Journal of Environmental Economics and Management* 109 (September): 102163. https://doi.org/10.1016/j.jeem.2018.08.010.

"OMB". 2023. "Circular A-4," September. https://obamawhitehouse.archives.gov/node/15644.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2017. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12 (4): 375–85. https://doi.org/10.1093/pan/mph024.

Pasek, Josh. 2018. *Anesrake: ANES Raking Implementation*. https://cran.r-project.org/package=anesrake.

Penn, Jerrod M., Daniel R. Petrolia, and J. Matthew Fannin. 2023. "Hypothetical Bias Mitigation in Representative and Convenience Samples." *Applied Economic Perspectives and Policy* 45 (2): 721–43. https://doi.org/10.1002/aepp.13374.

Petrolia, Daniel R., and Matthew G. Interis. 2013. "Should We Be Using Repeated-Choice Surveys to Value Public Goods?" November. https://papers.ssrn.com/abstract=2354495.

Poe, Gregory L. 2016. "Behavioral Anomalies in Contingent Values and Actual Choices." *Agricultural and Resource Economics Review* 45 (2): 246–69. https://doi.org/10.1017/age.2016.25.

Prosser, Christopher, and Jonathan Mellon. 2018. "The Twilight of the Polls? A Review of Trends in Polling Accuracy and the Causes of Polling Misses." *Government and Opposition* 53 (4): 757–90. https://doi.org/10.1017/gov.2018.7.

Sandstrom-Mistry, Kaitlynn, Frank Lupi, Hyunjung Kim, and Joseph A. Herriges. 2023. "Comparing Water Quality Valuation Across Probability and Non-Probability Samples." *Applied Economic Perspectives and Policy* 45 (2): 744–61. https://doi.org/10.1002/aepp.13375.

Selb, Peter, and Simon Munzert. 2011. "Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information." *Political Analysis* 19 (4): 455–70. https://doi.org/10.1093/pan/mpr034.

Skibba, Ramin. 2016. "The Polling Crisis: How to Tell What People Really Think." *Nature* 538 (7625): 304–6. https://doi.org/10.1038/538304a.

Stan Development Team. 2024. *Stan Modeling Language Users Guide and Reference Manual*. https://mc-stan.org/.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. 2 edition. Cambridge ; New York: Cambridge University Press.

Vossler, Christian A., Stéphane Bergeron, Maurice Doyon, and Daniel Rondeau. 2023. "Revisiting the Gap Between the Willingness to Pay and Willingness to Accept for Public Goods." *Journal of the Association of Environmental and Resource Economists* 10 (2): 413–45. https://doi.org/10.1086/721995.

Vossler, Christian A., Maurice Doyon, and Daniel Rondeau. 2012. "Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments." *American Economic Journal: Microeconomics* 4 (4): 145–71. https://doi.org/http://www.aeaweb.org/aej-micro/.

Vossler, Christian A., David A. Keiser, Catherine L. Kling, and Daniel J. Phaneuf. 2025. "Information Scripts and the Incentive Compatibility of Discrete Choice Experiments." *Journal of the Association of Environmental and Resource Economists* 12 (2): 459–92. https://doi.org/10.1086/731527.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91. https://doi.org/10.1016/j.ijforecast.2014.06.001.

Warshaw, Christopher, and Jonathan Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *The Journal of Politics* 74 (1): 203–19. https://doi.org/10.1017/S0022381611001204.

Whitehead, John C., Andrew Ropicki, John Loomis, Sherry Larkin, Tim Haab, and Sergio Alvarez. 2023. "Estimating the Benefits to Florida Households from Avoiding Another Gulf Oil Spill Using the Contingent Valuation Method: Internal Validity Tests with Probability-Based and Opt-in Samples." *Applied Economic Perspectives and Policy* 45 (2): 705–20. https://doi.org/10.1002/aepp.13352.