# COMPARING WATER QUALITY VALUATION ACROSS PROBABILITY AND NON-PROBABILITY SAMPLES

By

Kaitlynn Sandstrom

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Agricultural, Food, and Resource Economics – Master of Science

2022

## ABSTRACT

This thesis compares the results of a stated preference survey administered to three samples: one non-probability sample and two non-probability samples. The probability sample is an address-based sample from the USPS postal delivery file, while the two non-probability samples are from the opt-in panels, MTurk and Qualtrics. The survey used a single binary referendum contingent valuation question with respondents voting on a water quality change at a cost to their household. To understand differences in economic values across samples, we compared results of logit models that relate the referendum vote to cost and each water quality index. Several tests reveal differences across samples. First, almost all parameters were significantly different across samples except for water clarity. Second, we compared marginal willingness to pay (MWTP). However, many of the MWTP estimates for individual water quality indices were not significantly different across the three sources. Third, we calculated total WTP (TWTP) for a range of non-marginal changes. The MTurk values were always significantly greater than the address sample at the 1% level, and the Qualtrics values were significantly greater than the address sample for changes up to about a 20% improvement. In summary, we find that the non-probability methods generate different valuation results than the probability-based sample, especially in terms of TWTP.

## ACKOWLEDGEMENTS

This thesis would not have been possible without the support and collaboration from numerous mentors and colleagues, which I would like to mention here. First, I would like to thank the U.S EPA for their funding and support in facilitating this project. This includes other EPA water quality grantees and EPA personnel who have provided beneficial discussions, and especially want to thank Chris Moore, Rob Johnston, Cathy Kling and Roger von Haefen. Our research has benefited from funding by Michigan AgBioResearch and from EPA Star Grant R836168 awarded to Michigan State University.

I sincerely want to acknowledge and give my thanks to my major professor, Dr. Frank Lupi, who's guidance and support has made this project possible. His advice these past two years has been integral in every step of the process. In addition, I want to thank and Joseph A. Herriges, Jan R. Stevenson and Hyunjung Kim for their contributions to this project. I would also like to thank my committee members, Dr. Vincenzina Caputo and Dr. David Ortega, for being a part of my master's journey and welcome their comments/suggestions. Lastly, I would like to thank my wonderful husband Smeet Mistry for supporting me during this academic journey and all that it has entailed.

## TABLE OF CONTENTS

# CHAPTER 1: Comparing Water Quality Valuation Across Probability and Non-Probability Samples

## 1.1: Introduction

Michigan is endowed with an abundance of waterbodies with over 26,000 inland lakes, 3,000 miles of Great Lakes shorelines, and 36,000 river miles (Curell, 2012). Unsurprisingly, water resource management is a key issue in Michigan. Federal, state, and local agencies tasked with water resources management must consider both the costs and benefits of potential programs when decision-making. Often, calculating the benefits of such programs can be quite difficult given that environmental improvements are not reflected in market prices, i.e., environmental benefits have non-market values. One way to quantify environmental benefits is through non-market valuation methods, which use observed or stated behavior to infer a value for changes in quality. Such methods contextualize the value of non-market goods by assigning a monetary value, often known as willingness to pay (WTP).

One way to estimate WTP is through stated preference (SP) surveys, which rely on asking respondents how they would behave in a hypothetical scenario. SP surveys are useful in estimating WTP values since they are the only method available to estimate non-use values (values for quality that are unrelated to observable behaviors). Often, stated preference data is more suited for policy evaluation since policies represent a hypothetical version of the future and can experimentally vary attributes such as environmental quality to create variation needed to identify parameters. Our research estimates values for freshwater ecosystem services in Michigan using a non-market valuation survey that asks respondents whether they would vote for a proposed change in water quality at a stated cost.

An important step in SP survey design is choosing an appropriate collection method (Johnston et al. 2017; Champ 2017). The traditionally preferred methods rely on probability sampling, such as addressed-based mailing or random-digital dialing. Probability sampling ensures that every member of the population of interest has a known probability of being selected to participate. Probability samples are preferred since they reduce the risk of systematic bias related to representation (Baker et al. 2010). While probability sampling methods are the gold-standard for survey sampling, they are often cost-prohibitive. Thus, in the internet era, non-probability online samples have seen rising predominance due to their speed and cost-effectiveness. Non-probability samples are different from probability samples because not everyone in the population has a known and equal chance of being selected. Most non-probability online samples use opt-in panels, where members choose to participate in the panel and are recruited online. While non-probability online samples offer advantages such as rapid collection times and lower costs, they are criticized for their potential biases that may not be mitigated by balancing samples to "represent" population demographics (Baker et al. 2010).

Understanding trade-offs related to non-probability samples is particularly pertinent to SP surveys, which are often complex and costly. Similarly, understanding differences in SP surveys and results due to sampling differences is an important part of best practices for SP (Johnston et al. 2017) and contributes to our understanding of valuation validity (Bishop and Boyle 2019). Some survey literature has found that non-probability samples are less accurate and more variable in their accuracy than probability samples (Yeager et al. 2011). Although a recent review of best practices for SP recommends probability sampling, it also notes that data collection mode may not considerably affect results citing several SP studies with mixed findings (Johnston et al. 2017). As methods change and access and familiarity with the internet increase,

continuing research is important since bias between non-probability online samples and traditional methods may change.

Given these tradeoffs, the aim of this research is to investigate how sociodemographic, attitudinal, and WTP values from a SP survey compare across probability and non-probability samples. We build on previous literature by providing evidence of the extent to which responses gathered via non-probability sampling differ from responses gathered via a representative sample of the population. Furthermore, this research specifically focuses on environmental SP research which is underrepresented in the literature comparing probability and non-probability internet samples. The SP survey was implemented to three sample types: one probability sample and two non-probability samples. The probability sample is an address-based sample from the USPS postal delivery sequence file and used a push-to-web design with mailed invitations to visit a website (Dillman 2017). The two non-probability online samples are from the opt-in panels MTurk and Qualtrics[1]. MTurk is an Amazon web-service where workers complete tasks, such as surveys, for small payments. Qualtrics is an opt-in panel using a range of proprietary methods and incentives. The SP survey used a single binary referendum contingent valuation question with respondents voting on a water quality change at a cost to their household. To understand differences in economic values across samples, we compared results of logit models that relate the referendum vote to cost and water quality indices. To test for differences across survey sources, we compared parameters, marginal willingness to pay (MWTP), and total WTP (TWTP) for a range of non-marginal changes. Overall, we have mixed findings, but generally conclude that the non-probability methods generate different valuation results than the probability-based sample. Based on these findings, we suggest that non-probability samples can be used in low-

[1] In 2020, Qualtrics began offering access to a probability-based sample in partnership with NORC at the University of Chicago, but almost all studies in the literature to date have used the more cost-effective opt-in panel.

stake applications, while probability samples are preferred for population studies and policy applications.

### 1.2: Literature Review

The gold-standard for valuation survey research has consistently been random samples drawn from frames consistent with the population (Johnston et al. 2017). Traditional survey methods include face-to-face interviewing, telephone interviewing, and self-administered mail surveys, which make use of representative sample frames (address and phone lists). These typically rely on addressed-based sampling (ABS) or random digit dialing (RDD). However, these methods come with challenges related to increasing non-response, increasing costs, and longer timeframes.

As a response to these challenges, the use of internet surveys has been increasing. There are clear advantages to using the internet as a medium: it is cheaper, quicker, and allows for interesting ways to present information. Furthermore, some evidence suggests that mode effects may be insignificant showing that surveys administered via the internet do not produce estimates significantly different than surveys administered via other methods (Baker et al. 2010; Braunsberger et al. 2007; Fleming & Bowden 2009; Lindhjem & Navrud 2011a; Lindhjem & Navrud 2011b; and Windle & Rolfe 2011). One exception to this is Boyle et al. (2016) who found that survey mode affects welfare estimates. The evidence also suggests that survey mode may not affect data quality, and in some instances, surveys administered via the internet can improve data quality when compared to other modes through higher concurrent validity, less survey satisficing, fewer protest responses, and greater reporting of socially undesirable attitudes/behaviors (Baker et al. 2010; Braunsberger et al. 2007; & Navrud 2011a; Lindhjem & Navrud 2011b; and Windle & Rolfe 2011). Additionally, some evidence has shown that choice

consistency does not differ between internet devices and that mobile devices can produce higher data quality when compared to desktops/laptops (Liebe et al. 2015).

Many internet surveys can be done using probability sampling. These surveys often use ABS and a push-to-web design, where invitations asking residents to complete an online survey are sent by mail to a random sample of addresses. However, many internet surveys use opt-in panels, where the respondent chooses to participate in the panel and is often recruited online. These samples are not typically probability samples; even though the probability of selection from the panel is known in some cases, the probability of selection from the general population is unknown. Because of their non-traditional and often opaque selection mechanisms, a main issue with non-probability online samples surrounds representation of the full range of people, knowledge, attitudes, behaviors and values as in the population, which may not be well reflected by merely having balanced representation in a few demographic categories.

To date, numerous studies have compared probability and non-probability samples to determine whether there are significant differences between the two in terms of demographics and substantive estimates. Generally, studies find demographic differences (Atkeson et al. 2011; Chang & Krosnick 2009; Pennay et al. 2018; Zack et al. 2019). The evidence is mixed about estimates, but most find that estimates from non-probability and probability samples differ, with probability samples being more reliable even after weighting (Baker et al. 2010; Chang & Krosnick 2009; Pennay et al. 2018; Yeager et al. 2011; Zack et al. 2019). However, some found that the results were similar, especially after weighting (Atkeson et al. 2011; Weinberg et al. 2014). These findings are also true for studies specifically using SP surveys. Many found that the non-probability and probability samples differed in terms of demographics (Boyle et al. 2016; Bonnichsen & Olsen 2016; Grandjean et al. 2009; Olsen 2009). Again, the evidence is mixed

regarding the reliability of estimates. Some find that preference estimates differ between samples (Boyle et al. 2016; Bonnichesn & Olsen 2016), while others found that they produce similar preference estimates (Grandjean et al. 2009; Olsen 2009). One recent study (Whitehead et al. 2021) investigates the internal validity of a probability and a non-probability sample for a CV survey by testing sensitivity to cost, sensitivity to scope, and by comparing income elasticities of WTP. They find that the probability sample generally exhibits greater internal validity.

Even though non-probability samples are often subject to coverage bias, statistical adjustments through weighting or modeling could potentially minimize or eliminate bias. The most common method to correct bias is standard demographic weighting, but many other techniques exist. At the sample design stage, standard quota sampling is one option, which uses panel member information to create demographically balanced samples. Post-survey methods include model-based and sample matching. At the post-survey stage, post-stratification adjustments are often used, which uses standard demographic weighting with attitudinal/behavior measures that are potentially predictors of bias. The evidence surrounding the usefulness and reliability of such adjustment methods are mixed. For studies that specifically examined SP surveys, some found that adjustments did not eliminate bias (Bonnichsen & Olsen 2016), while others found that estimates were improved through generalized regression weighting (GREG) (Dever et al. 2018) or raking and propensity weighting (Roshwalb et al. 2016). Evidence from non-SP studies generally found that adjustments were useful in reducing bias but offer only a partial remedy (Baker et al. 2010; Pennay et al. 2018; Yeager et al. 2011; Zack et al. 2019) or had no effect (Chang & Kosnick 2009).

In addition to the above, there are concerns about the data quality of opt-in online panels. Many of these concerns relate to satisficing, fraudulent behavior, and professional respondents.

Satisficing is when respondents put in less cognitive effort, resulting in inaccurate and inconsistent responses. Often, it is difficult to measure satisficing since it may be confounded with other effects, resulting in very few "true experiments." The limited body of evidence shows that there tends to be less satisficing in self-administered online surveys (Baker et al. 2010). For stated-preference surveys specifically, evidence shows that respondents who do poorly on attention checks yield less efficient estimates (Gao et al. 2016). Fraudulent behavior relates to self-misrepresentation or the use of bots to maximize rewards. Generally, surveys of narrow populations are more prone to fraud, which reduces data quality (Brazhkin 2020). Methods to reduce or detect satisficing and fraudulent behavior include attention checks, bogus questions, open-ended questions, self-reports of effort, response times, analyzing choice response patterns and selection of non-substantive responses, digital fingerprinting, as well as many other methods (Aguinis et al. 2021; Baker et al. 2010; Brazhkin 2020; Chmielewski & Kucker 2020; Teitcher et al. 2015). Professional respondents are those who participate in surveys frequently. The issue with professional respondents is non-naivete or panel conditioning, which refers to a change in behavior or attitudes due to repeated survey completion. Evidence shows that "hyperactive" respondents who complete surveys frequently significantly impact estimates, whereas "experienced" respondents who have long panel tenure may be less of an issue (Sandorf et al. 2020). Unfortunately, correcting for fast/slow or low/high effort respondents does not correct the impact of professional respondents (Sandorf et al. 2020).

One specific opt-in online panel is Amazon's Mechanical Turk (MTurk), which pays respondents to perform tasks, such as completing surveys. MTurk is an Amazon web-service where workers complete tasks, such as surveys for payments. These tasks are only visible to workers who meet pre-defined criteria and workers can sort through tasks based on reward size,

time for completion, and description before they decide to accept a task. Researchers can also screen workers via approval ratings and can choose to deny payments based on performance. Several studies have examined the quality of MTurk data. Like other opt-in online panels, MTurk faces representation challenges, such as self-selection bias, and data quality challenges such as inattention, self-misrepresentation, non-naivete, and vulnerability to bots (Aquinis et al. 2021). In terms of representation, most studies find that MTurk respondents are less representative than probability samples, but more representative than other convenience samples (Berinsky et al. 2012; Paolacci et al 2010; Weinberg et al. 2014; Zack et al 2019). In terms of data quality, the evidence is mixed. Some studies find that MTurkers pay more or equal amounts of attention when compared to other online or convenience samples (Berinsky et al. 2012; Paoacci et al. 2010) or to a probability sample (Weinberg et al. 2014). In terms of professional respondents, one study found that habitual survey takers and repeat survey taking are not a problem (Berinsky et al. 2012). Additionally, a study found that MTurk did better than a probability sample in terms of comprehension check items, time to complete, nonresponse, variation in responses (Weinberg et al. 2014). Conversely, other studies have found that MTurk suffers significantly from data quality problems. One study found that 36% of MTurkers failed a proxy for careful participation (Downs et al. 2010), while another found that MTurk suffers from response inconsistency, statistically improbable comments, and unusual comments, indicating low quality responses (Chmielewski & Kucker 2020).

Another opt-in panel used in this study is Qualtrics. Qualtrics is slightly different from MTurk because Qualtrics uses a variety of opt-in methods and incentive types to populate panels. Specifically, Qualtrics aggregates panel respondents initially recruited by other firms. In this case, researchers contract with Qualtrics, not the individual workers. The main differences

between the two services are that: 1) Qualtrics allows for quota sampling, which ensures that the sample matches the demographics of the target population, 2) Qualtrics uses a range of sampling methods to create their panels, and 3) Qualtrics does not have a screening mechanism like MTurk does, which allows researchers to set rating requirements so that only workers with certain ratings can complete jobs. Researchers that have used MTurk for previous surveys can also exclude the previous respondents from a survey, if desired. Several studies have compared Qualtrics to MTurk (or to other similar online opt-in panels) in a non-SP setting. In terms of demographics, some studies have found that Qualtrics differs from the probability sample/the general population/or the target population (Beymer et al. 2018; Roulin 2015), but less so than MTurk (Zack et al. 2019). However, this may be country dependent, since Boas et al. (2020) find that Qualtrics is more demographically and politically representative in the US, but MTurk is more representative in India. In terms of outcome variables, studies find that the different panels produce different results from each other (Armstrong et al. 2020), from the targeted population (Beymer et al. 2017) and from the general population (Zack et al. 2019). However, some find that Qualtrics outperforms MTurk (Zack et al. 2019), while others find that MTurk performs better (Owens & Hawkins 2019). Others find that the panels produce similar results when compared to each other and to a probability sample (Roulin 2015). Some studies have compared data quality and the findings are similarly mixed. Some find that both MTurk and Qualtrics have good reliability (Roulin 2015) and high attention rates (Beymer et al. 2017). However, other studies find that MTurk participants pay more attention and better acquire and recall information than Qualtrics participants (Boas et al. 2020; Owens & Hawkins 2019). Additionally, Owens & Hawkins (2019) find that the quality and generalizability of Qualtrics data is not improved by eliminating participants who fail attention checks. Thus, in the U.S., Qualtrics seems closer to

the general population than MTurk, but the evidence is mixed, and it is unclear whether the data quality is the same.

As internet surveys and, subsequently, non-probability sampling become more popular it is important to ask whether non-probability samples can serve as an appropriate replacement for probability samples, especially in terms of representation. Our research seeks to contribute to the growing literature by investigating whether non-probability samples are representative when compared to a probability sample in terms of demographic, attitudinal, and economic values. Currently, research on non-probability versus probability sampling is sparse when it comes to environmental SP surveys. Given the high costs of SP surveys, continued research on non-probability sampling is particularly important. Our research contributes to the literature in an additional way by comparing two popular and sometimes controversial opt-in panels, MTurk and Qualtrics.

### 1.3: Background and Data

An online SP survey was implemented to estimate WTP values for water quality improvements in the state of Michigan. Water quality was described by four indices: a water clarity score based on secchi depth; a water contact score based on bacterial counts; an aquatic wildlife score based on biological condition, and a gamefish biomass score. These scores were presented in the survey via text, graphics, and maps depicting the average scores for each watershed across the Lower Peninsula of Michigan. The water quality scores each range from 0 (worst possible quality) to 100 (best possible quality). Scores were described to respondents in writing and depicted using a color bar labeled with levels in a manner analogous to previous water quality indices and ladders used by EPA. Each index description included questions to encourage the respondents to engage with the information. Following presentation of the

baselines, the water quality improvements were also shown to respondents in the same ways as the baseline scores.

The SP survey was developed according to suggestions by Johnston et al. (2017) for survey development. The survey included information about the respondent's usage of Michigan water bodies, a description of the water quality indices, a description of current water quality, a policy scenario, and socio-demographic and attitudinal questions (see appendix). Other design features include questions on consequentiality, survey bias and an attention question. The policy scenario included a single binary referendum contingent valuation question with respondents voting on a water quality change at a cost to their household. Across respondents, there were 30 experimentally varied scenarios for water quality changes and costs, with four cost alternatives and five water quality change levels for each index. The cost levels ranged from $65 to $965, while the water quality changes ranged from 0-20 points. A D-efficient design was used to vary attributes (attribute table located in Appendix G.1). Each scenario was assigned at random, and the order that indices were presented and summarized was randomized across respondents.

The design of the survey followed the steps and logic of a CV survey conducted on the Deepwater Horizon Oil Spill (Bishop et al. 2017). Specifically, the survey used a referendum format asking people to vote on a water quality plan, which would impose a one-time household income tax. Additionally, steps were taken to ensure that the proposed plan and tax were seen as consequential, i.e., the respondents believed their answers would affect the passing of the proposed plan and that they would have to pay the tax amount shown in the survey (Carson & Groves 2007; Herriges et al, 2010). To reinforce policy consequentiality, respondents were told that the survey results would be shared with policy makers and that there was only one plan

being considered. In addition, people had to provide income information before they were shown a tax amount, which reinforced the consequentiality of the tax instrument.

The survey was developed over the course of a few years and in four stages, as outlined in Figure 1. The first development stage consisted of focus groups conducted by the EPA that tested survey language, water quality metrics, and the general survey design, which we leveraged and built upon (Moore et al., personal correspondence). Next, using an iterative survey testing and updating approach following Kaplowitz et al. (2004), survey questions were tested using a series of one-on-one cognitive interviews with MSU students and Michigan residents. The third development stage was the first survey pilot, which was conducted on about 600 MTurk respondents in 3 waves. After each wave, the survey was analyzed and updated. The final development stage was the second survey pilot, which was conducted on 3000 Qualtrics respondents. During this pilot, the focus was on testing alternative water quality indices, as well as a full data analysis to inform the final survey (Lupi et al., 2022). Prior to the pilot and the final survey implementation, additional cognitive interviews were used to test and refine the survey.

The final survey was implemented to the three sample types: an addressed-based sample, Qualtrics, and MTurk. Except for a few items like how user IDs were handled, the surveys were identical with the same experimental designs and protocols. The probability sample is an address-based sample from the USPS postal delivery file and uses a push-to-web mail design. Cash incentives included $1 pre-survey to all addressees with the invitation letter, and on the third invite letter a $20 post-survey completion incentive was offered of which about one-third did survey but instructed the money be put into the project. The ABS survey received a 23% response rate yielding 2,531 observations. The two non-probability online samples include 1,238 respondents from MTurk and 3,094 from Qualtrics. For Qualtrics, we paid five dollars per

response, but there is no way to know the amount or type of incentive respondents received. We paid MTurk workers five dollars for completed surveys, and they had to be workers that reside in Michigan that did not previously complete a survey for us. In order to assure that the surveys were being completed by the target audience, the surveys included certain questions such as affirming the respondent was over 18, affirming the respondent was involved in household decision-making, and two types of checks in each survey source related to Michigan residency. Before the data was analyzed, the data was cleaned, and a final sample was created. This step included dropping respondents who did not respond to the vote, those who did not live in the study region, and those with potentially fraudulent responses.

One of the risks of opt-in online panels is bots and other types of fraudulent answers (Johnston et al. 2021). Fraud detection occurred at multiple stages in the survey process. First, fraudulent answers were detected by Qualtrics and removed by Qualtrics staff using a variety of detection tools as well as reading through our open-ended response fields for gibberish. Qualtrics also flagged speeding, straight lining, and non-sensical responses. Second, fraudulent answers were detected during the data cleaning stage. This mostly included responses from MTurk, which does not have the fraud-detection feature like Qualtrics. Those who gave who responses using straight lining tactics (giving the same response to a series of grouped questions) or those who gave non-sensical/gibberish answers were dropped from the final sample. This resulted in 74 observations dropped due to straight lining (1% of all responses), and 38 observations dropped due to non-sensical answers (0.6% of all responses).

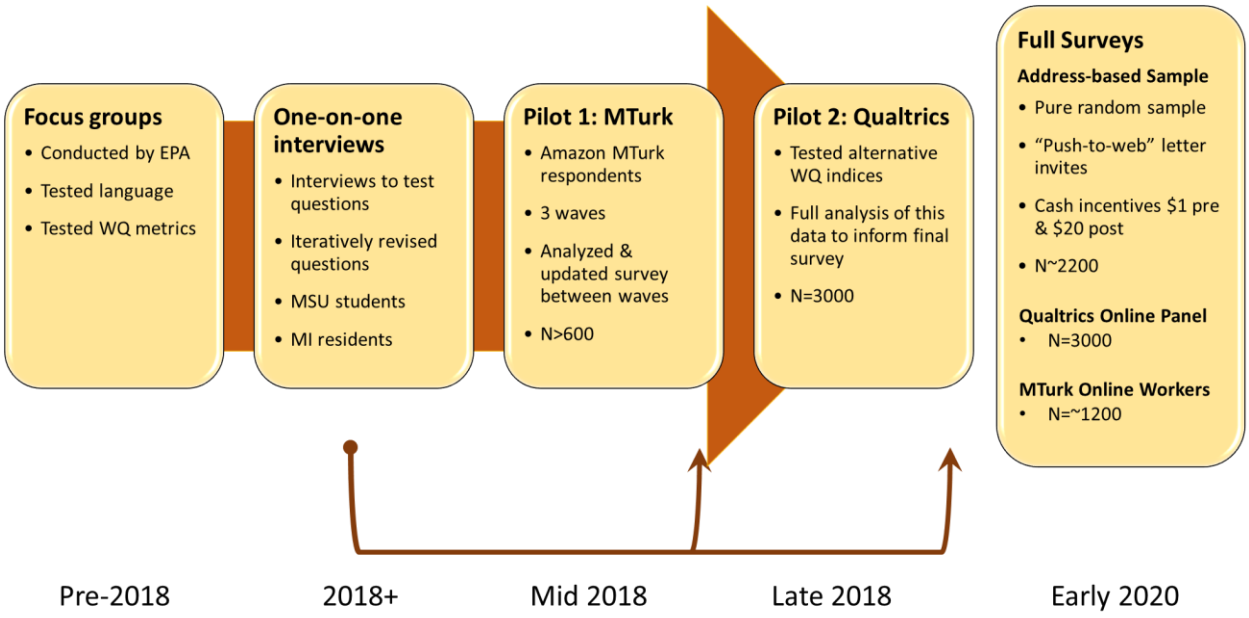


**Figure 1.** Major steps and timing for the extensive survey development, testing and pilot surveys that preceded the final surveys with the three sample types

### 1.4: Econometric Methods and Specification

Analysis of the SP data is based on the Random Utility Maximization (RUM) Model. The RUM model is based on the hypothesis that individuals, when presented with a set of alternatives comprised of different attributes and levels, maximize their utility by choosing the alternative that gives them the greatest satisfaction (McFadden 1974). This study presented respondents with a dichotomous choice between two alternatives: a baseline water quality (current state) at no cost and an improved state for water quality at a cost. Each respondent, *i*, was asked whether they would vote for or against the change. Individual utility is thus a function of proposed policy price, water quality scores, and some degree of randomness for factors influencing choice that cannot be observed by the researcher. Consider respondent, *i*, facing *j* alternatives. The utility function ($U_{ij}$) for the respondent is given by:

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \qquad (1)$$

where $V_{ij}$ is an observed component and $\varepsilon_{ij}$ is a random component. Here we assume that $V_{ij}$ is a linear utility function, so $V_{ij} = \beta x_{ij}$. In which case, $\beta$ is a vector of parameters, $x_{ij}$ is a vector of explanatory variables, and $\varepsilon_{ij}$ is assumed to be distributed i.i.d extreme value. Since there are two alternatives, each individual considers their utility without and with the program:

$$U_{i0} = V_{i0} + \varepsilon_{i0} = \beta_{i0} Q_{i0} + \theta_{i0} Y_i + \varepsilon_{i0,} \quad (2)$$

$$U_{i1} = V_{i1} + \varepsilon_{i1} = \alpha_i + \beta_{i1} Q_{i1} + \theta_{i1} (Y_i - P_i) + \varepsilon_{i1,} \quad (3)$$

Where $U_{i0}$ is baseline utility, $U_{i1}$ is the utility of the improved water quality and price with the program, $\alpha$ is a constant, $Q_{i0}$ is the baseline water quality level, and $Q_{i1}$ is the improved water quality level under the proposed policy, $Y_i$ is the individual's income, and $P_i$ is the cost of the policy to the individual. In this study, Q consists of water clarity, water contact, fish biomass, and wildlife scores. We assume that if a respondent voted yes for the referendum, then $U_{i1} > U_{i0}$. This also means that, if person *i* voted yes, then their WTP for water quality improvements is greater than the program cost.

To estimate the model, we use a binary logit model. Thus, we model the probability of individual *i* voting yes as $Pr_1$:

$$Pr_{i1} = \frac{e^{\alpha + \beta dQ_i - \theta P_i}}{1 + e^{\alpha + \beta dQ_i - \theta P_i}}$$

where *Pr* is the probability that respondent *i* will respond with a yes vote, $\beta$ is a vector of marginal utility parameters to be estimated, $\alpha$ is a constant, and Q is a vector of explanatory variables for the four water quality indices.

### 1.5: Results

This section contains results of our preliminary data analysis and comparisons of the three sample types. Key descriptive statistics are described in Figure 2. In our comparison samples, there were 6,071 observations with complete data: 2,016 from the address-based sample, 1,069

from the MTurk sample, and 2,986 from the Qualtrics sample. For all samples, the median income was $62,500 (note that income was reported in ranges in the surveys), which is similar to Michigan's state 2019 Census median income of $59,584. Across 6 demographics, we found that the three samples differed in three key areas: age, employment status and children in household. The MTurk sample skewed younger while the Qualtrics sample skewed older when compared to the address sample (the percentage under age 55 was 50 % for the address sample, 91% for MTurk, and 38% for Qualtrics). Employment status also differed. For MTurk, only 3% were retired, while the percentage retired for the address sample was 28% and 36% for Qualtrics. The share of respondents with children in their household was higher for MTurk than the ABS whereas Qualtrics had a lower share. The MTurk sample was also more educated than the others.

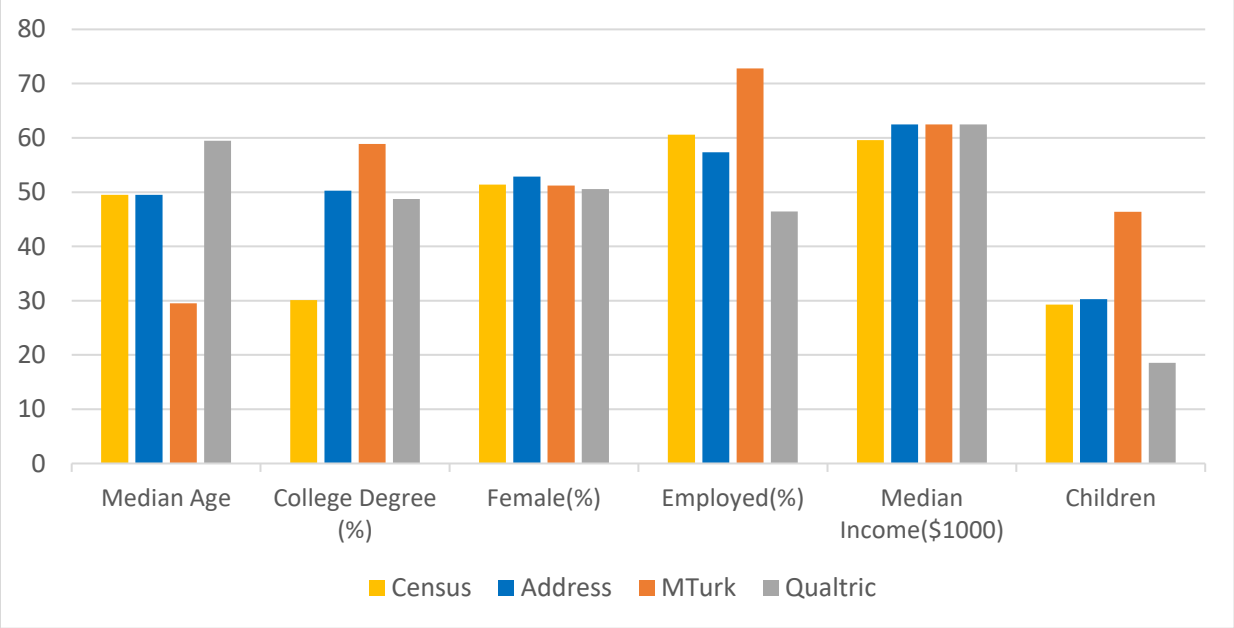


**Figure 2.** Demographic comparisons of each sample and Census

Several attitudinal scale variables were also examined across the samples. Two types of "importance" attitudinal questions were included in the survey. The first was at the beginning of the survey, where respondents were primed to think about a variety of public policy issues with a

set of questions on the importance of various policy issues. The second set of importance questions were about specific qualities of the proposed plan and appeared prior to the vote to encourage respondents to think about the plan that was just described to them. In addition, after the vote there were a series of Likert agreement questions regarding various aspects of their vote. In general, these attitude measures were very similar across samples. There were 29 possible attitude questions to compare, and the average scores differed from the ABS by more than 10-percentage points for only three items. The MTurk sample scored 10% lower for "it is important to cut taxes" than the ABS but were more likely to explain their vote by opposing taxes. Another question explaining votes asked whether the respondent agreed with the statement "If the plan is implemented, I personally would not benefit." In this case, the MTurk sample was more likely to agree with the statement scoring 19% more than the ABS while the Qualtrics sample scored 12% more than the ABS. Thus, overall, the three samples produced similar attitudes about policy stances and qualities of the proposed plan, with a few minor differences.

The survey also examined how households value "use" and "non-use" components of water quality changes. "Use values" include the value of visiting and using the waterbodies impacted by the plan, such as improved fishing, better swimming conditions, and more wildlife to view when visiting and using waterbodies. "Non-use values" include the value of water quality changes for their own sake, even if the respondent doesn't ever visit and use the waterbodies. The respondents were asked to consider how much of the value of the proposed water quality plan comes from "use" and "non-use" values. Again, this question was represented as a 5-point scale, ranging from "almost all value from use" to "almost all value from non-use." All three samples got the majority of their value from a mix of use and non-use values and, on average, got slightly more value from non-use values. Of the three samples, Qualtrics got the

most value from non-use values, but neither the average scores from Qualtrics or MTurk differed from ABS by more than 10-percentage points. Using chi-square tests, we find that the responses to the use/non-use question are significantly different between the samples. However, the general patterns are very similar across samples, as evident by the graph below. Overall, most people get equal value from use and non-use, while few people get value from either all non-use or use.

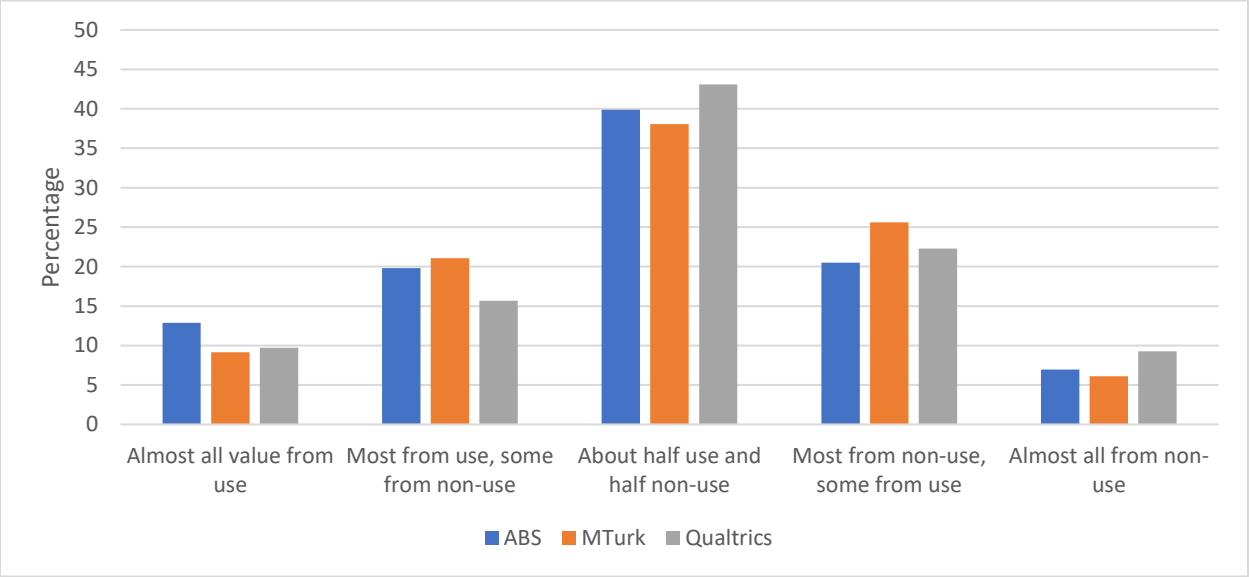


**Figure 3.** Use and non-use value comparisons of each sample

The dependent variable in our study is a dummy variable taking a value of 1 if the individual voted for the proposed policy and 0 if the individual voted against for the proposed policy. The main independent variable of interest is cost, which is the cost of the proposed policy for each individual. Figure 4 depicts how the percentage of those who voted "yes" varies over proposed cost and across samples. Based on economic theory, we expect cost to have a negative effect on the probability of voting "yes". The other independent variables are changes in water clarity, water contact, fish biomass, and wildlife scores, although the experimental design does not hold these constant across the price levels for the sake of estimation efficiency. Based on

economic theory, we expect higher costs to have a negative effect on the probability of voting "yes". We see in the figure that the expected relationship holds for the ABS sample with the percent yes declining monotonically with costs. In general, both the other samples also decline with costs, but the relationship is not strictly monotonic. We also see that the Qualtrics respondents are more likely to vote yes at all costs than the ABS, with the MTurk respondents always being the most likely to vote yes in each case.

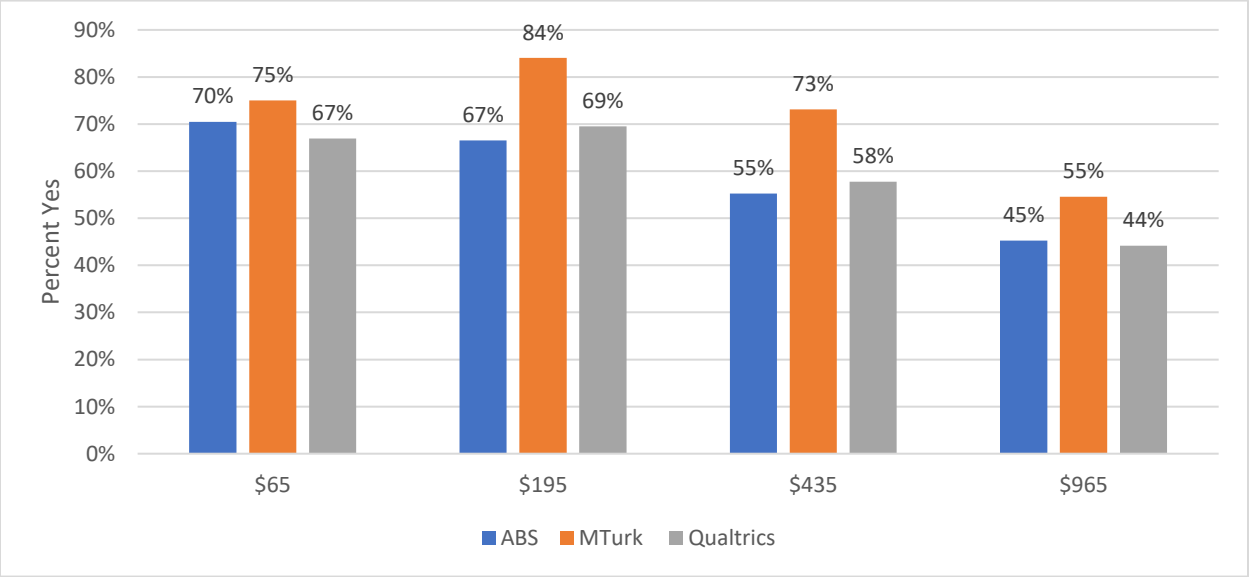


**Figure 4**. Percent of Yes Vote by Cost*

* Figure 4 mutes the steepness of vote response to cost since, for efficiency, the experimental design tends to have small changes in quality at low costs more often than at high costs; specifically, average changes in quality at the highest cost are about double those at the lowest cost. Thus, all else is not equal across the price levels.

Using information from the surveys, a logit model was estimated that relates the referendum vote to cost and each water quality index. [2] The regression uses data from all three

[2] The core model does not include socio-demographic information. Inclusion of socio-demographic information did not change the general pattern of the valuation results. The coefficients on cost are negative and the coefficients on water quality indices are positive in all samples. Additionally, the inclusion of socio-demographic information did not change the significance level for any cost or water quality parameter. Importantly, income did not have a significant effect on voting for MTurk, contrary to expectations for some forms of validity (Whitehead et al., 2021). See appendix for estimation results.

sample types, although each sample type has its own interaction terms which effectively results in separate parameters for each sample. Coefficients, standard errors, and test statistics are reported in Table 1. MWTP is reported since the individual parameter estimates cannot be used to indicate a variable's marginal effect. As expected, the coefficient on cost is negative and the coefficients on water quality indices are positive in all samples. Additionally, nearly every variable is significant at the 1% level demonstrating that there is significant sensitivity to the scope of the quality change in each sample type. Several tests yield mixed evidence on differences across samples. Comparing the parameters reveals that all parameters were significantly different across samples except for water clarity. We also examined the marginal willingness to pay (MWTP) for individual water quality indices. Although the values differ sometimes by more than a factor of two, most were not significantly different across the three sources; one exception was the low and insignificant MWTP for fish biomass in the Qualtrics sample and water contact in the MTurk sample.

Next, we computed TWTP for a range of non-marginal changes by simultaneously moving each index by the same increment, which is shown in Figure 5 along with the 95% confidence intervals. We found that for the relevant range of water quality changes from baseline to 30 additional percentage points on the indices, the MTurk values were always significantly greater than the address sample at the 1% level, and the difference grows with the size of the changes. Although the Qualtrics values were significantly greater than the address sample for changes up to about a 22% improvement (13 points on the indices), the Qualtrics values showed less scope sensitivity (were less responsive to quality changes) and became lower than the address sample for changes that were 33% or larger (21 points on the indices).

We also computed TWTP for a range of non-marginal changes for each individual index, which is shown in the appendix, along with the 95% confidence intervals. The fish biomass score is the only index where a sample had almost constant values across the range of improvements, which was for the Qualtrics sample. This meant that ABS had lower values than Qualtrics for up to a 20-point increase, and higher values past this point. For each index, MTurk always had the highest TWTP values.

**Table 1.** Logit Model: Estimates and Statistics

| Variable | ABS | MTurk | Qualtrics | Differences Test |
|---|---|---|---|---|
| Cost | -0.0012*** | -0.0018*** | -0.0013*** | 4.60* |
| SE | 0.0001 | 0.0002 | 0.0001 | |
| Δ in Water Clarity Score | 0.0140* | 0.0190* | 0.0138** | 0.32 |
| SE | 0.0054 | 0.0083 | 0.0045 | |
| MWTP | 11.2479 | 10.8388 | 10.6941 | |
| Δ in Water Contact Score | 0.0117* | 0.0369*** | 0.0179*** | 6.67** |
| SE | 0.0054 | 0.0082 | 0.0045 | |
| MWTP | 9.4302 | 21.0467 | 13.8535 | |
| Δ in Fish Biomass Score | 0.0166*** | 0.0237*** | 0.0021 | 11.73*** |
| SE | 0.0043 | 0.0065 | 0.0036 | |
| MWTP | 13.3664 | 13.5512 | 1.6263 | |
| Δ in Wildlife Score | 0.0135* | 0.0344*** | 0.0112* | 6.51** |
| SE | 0.0053 | 0.0081 | 0.0045 | |
| MWTP | 10.8689 | 19.6694 | 8.6855 | |
| Constant | 0.4136*** | 0.7622*** | 0.5215*** | 2.81 |
| SE | 0.1172 | 0.1718 | 0.0963 | |
| N | 2016 | 1069 | 2986 | 6539 |
| LogL | | | | -4274.91 |
| Wald $\chi^2$ | | | | 451.58 |

*Note: ***, **, * represents significance at the 1%, 5%, and 10% level*

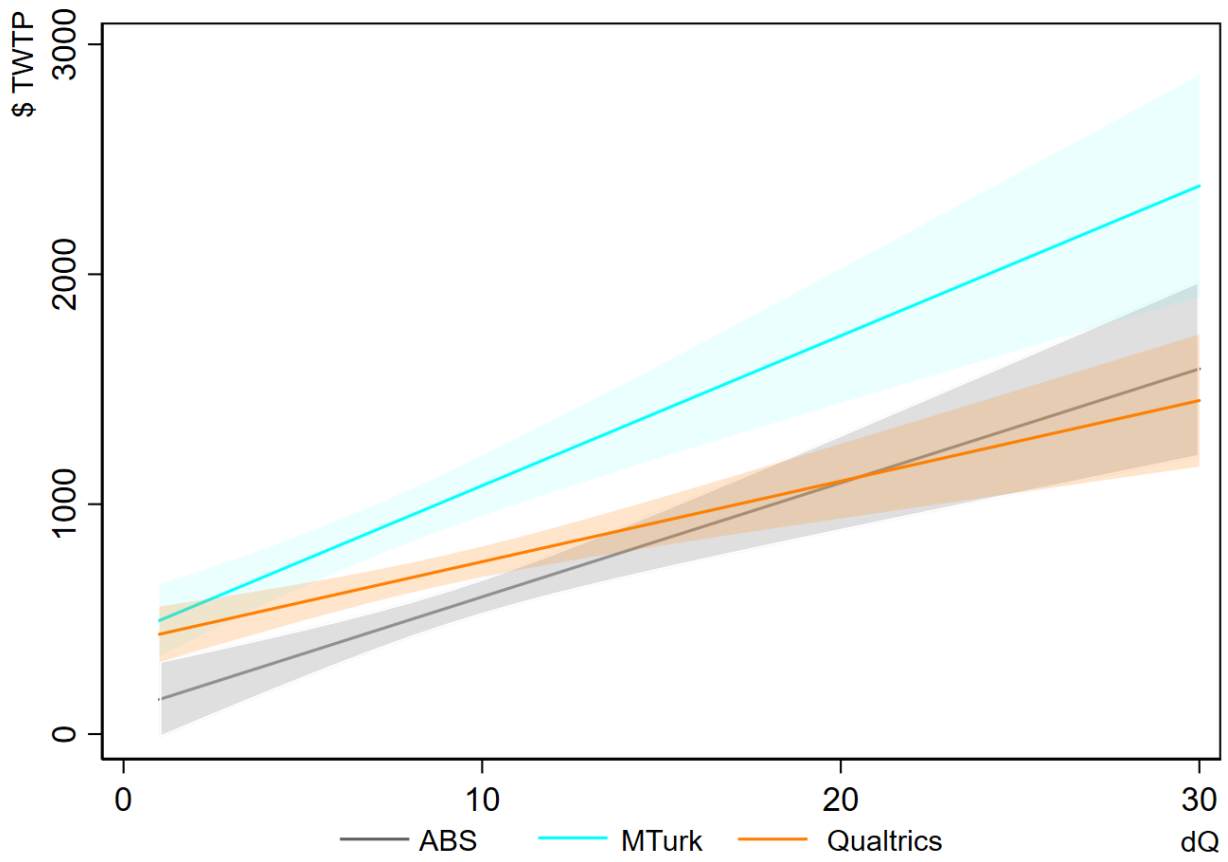


**Figure 5**. TWTP and 95% confidence intervals for a range of non-marginal changes to all four water quality indices

### 1.6: Discussion

Previous research comparing probability and non-probability samples has been mixed but generally finds that demographics and estimates from non-probability and probability samples differ. Even though research has been conducted on the topic, continuing research is necessary as internet access improves and recruiting methods for non-probability online panels change. Additionally, there is not an abundant amount of literature that specifically compares probability and non-probability samples for environmental SP studies. There is also the question of whether the research trend is true across geographic regions. Thus, this study contributes to the growing body of literature by evaluating the appropriateness of non-probability online samples as an alternative to probability samples for environmental SP surveys by examining differences

between demographic, attitudinal, and economic values. This is particularly relevant considering the time and monetary costs of SP surveys. The world is increasingly interconnected, and the internet plays an outsized role in communication making non-probability online samples are a cost-effective alternative for reaching most people. While these non-probability samples may be subject to greater bias, it is important to determine if they offer a viable alternative by achieving similar results to traditional probability sampling for SP.

Overall, the results of our research lead to mixed conclusions about the probability versus non-probability samples. We find that while the samples differ in some demographics, such as the relative youth of the MTurk respondents, the samples were very similar on average across 29 attitudinal variables and on the extent that they voted based on use versus non-use. In valuation models that included income, the MTurk sample income was not significant in explaining votes suggesting a potential validity problem. In terms of MWTP, most were not significantly different except for the fish biomass in the Qualtrics sample and water contact in the MTurk sample. Despite these similarities, there were notable differences between the probability and non-probability samples in terms of WTP for non-marginal changes. The non-probability samples were more likely to vote yes and had higher values than the probability sample[3], especially MTurk. Based on these results, we can generally conclude that the non-probability methods generate different valuation results than the probability-based sample, especially in terms of TWTP.

Previous literature supports using probability-based samples for population-based applications. Our findings do not contradict this evidence and we recommend that probability samples be used for population studies. Considering that most policies are for non-marginal

[3] Including socio-demographic information in the model did not change the general pattern of results. In addition, performing a weighting procedure did not change the valuation results, as it did in some previous studies.

changes, and that we find differences between the samples in WTP for non-marginal changes, we also recommend that probability samples are used for policy applications and any high-stakes decisions. However, there are applications where non-probability sampling could be appropriate. For example, non-probability sampling could be useful for informing low-stakes decisions, for simple A/B tests, testing methods, or for preliminary/investigative testing.

One limitation of this study is that there are unknowns due to the fact that Qualtrics aggregates panel respondents initially recruited by other firms. This means we lack information about opt in methods, the incentives Qualtrics respondents received, and the response rate for this sample. This is one avenue that future research could explore. Another limitation is that the MTurk sample size is about half the size of ABS and Qualtrics, which can affect internal validity comparisons and is something that should be considered by future studies. Despite these limitations, these results are pertinent given the growing use of non-probability samples. Due to this growth, there is a wide range of issues for future research to explore. Future studies could employ different types of surveys, explore various policy questions, or utilize different non-probability platforms. They could also explore various methodological or data quality devices. For example, this study used raking to weight the results to determine if differences persisted (see the Appendix), but a variety of other weighting methods could be used. As non-probability sampling grows in prevalence, continuing research will be necessary.

## BIBLIOGRAPHY


Aguinis, Herman, Isabel Villamor, and Ravi S. Ramani. "MTurk Research: Review and Recommendations." *Journal of Management* 47.4 (2021): 823-837.

Armstrong, Beth, et al. "How does Citizen Science compare to online survey panels? A comparison of food knowledge and perceptions between the Zooniverse, Prolific and Qualtrics UK Panels." *Frontiers in Sustainable Food Systems* 4 (2020): 306.

Atkeson, Lonna Rae, et al. "Considering mixed mode surveys for questions in political behavior: Using the Internet and mail to get quality data at reasonable costs." *Political Behavior* 33.1 (2011): 161-178.

Baker et al. Prepared for the AAPOR Executive Council by a Task Force operating under the auspices of the AAPOR Standards Committee. "Research synthesis: AAPOR report on online panels." *Public Opinion Quarterly* 74.4 (2010): 711-781.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political analysis* 20.3 (2012): 351-368.

Beymer, Matthew R., Ian W. Holloway, and Christian Grov. "Comparing self-reported demographic and sexual behavioral factors among men who have sex with men recruited through Mechanical Turk, Qualtrics, and a HIV/STI clinic-based sample: Implications for researchers and providers." *Archives of sexual behavior* 47.1 (2018): 133-142.

Bishop, R., and K. Boyle. 2019. Reliability and validity in nonmarket valuation. *Environmental and Resource Economics* 72:559–582.

Bishop, Richard C., et al. "Putting a value on injuries to natural assets: The BP oil spill." *Science* 356.6335 (2017): 253-254.

Boas, Taylor C., Dino P. Christenson, and David M. Glick. "Recruiting large online samples in the United States and India: Facebook, mechanical turk, and qualtrics." *Political Science Research and Methods* 8.2 (2020): 232-250

Bonnichsen, Ole, and Søren Bøye Olsen. "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 (2016): 245-262.

Boyle, Kevin J., et al. "Investigating Internet and mail implementation of stated-preference surveys while controlling for differences in sample frames." *Environmental and Resource Economics* 64.3 (2016): 401-419.

Braunsberger, Karin, Hans Wybenga, and Roger Gates. "A comparison of reliability between telephone and web-based surveys." *Journal of Business Research* 60.7 (2007): 758-764.

Brazhkin, Vitaly. "“I have just returned from the moon:” online survey fraud." *Supply Chain Management: An International Journal* (2020).

Carson, Richard T., and Theodore Groves. “Incentive and informational properties of preference questions.” *Environmental and Resource Economics* 37 (2007):181–210.

Champ P.A. 2017. Collecting Survey Data for Nonmarket Valuation. In: Champ P.A., Boyle K.J., Brown T.C. (eds) A Primer on Nonmarket Valuation. The Economics of Non-Market Goods and Resources, 59-98, Springer, Dordrecht.

Chang, Linchiat, and Jon A. Krosnick. "National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality." *Public Opinion Quarterly* 73.4 (2009): 641-678.

Chmielewski, Michael, and Sarah C. Kucker. "An MTurk crisis? Shifts in data quality and the impact on study results." *Social Psychological and Personality Science* 11.4 (2020): 464-473.

Curell, C. (2021, July 29). *Michigan Water Facts*. MSU Extension. Retrieved December 14, 2021, from https://www.canr.msu.edu/news/michigan_water_facts.

Dever, Jill A., Ann Rafferty, and Richard Valliant. "Internet surveys: Can statistical adjustments eliminate coverage bias?" *Survey Research Methods.* 2.2 (2008).

Dillman, D.A. 2017. The promise and challenge of pushing respondents to the Web in mixed-mode surveys. *Survey Methodology* 43(1):3–30.

Downs, Julie S., et al. "Are your participants gaming the system? Screening Mechanical Turk workers." *Proceedings of the SIGCHI conference on human factors in computing systems* (2010).

Fleming, Christopher M., and Mark Bowden. "Web-based surveys as an alternative to traditional mail methods." *Journal of environmental management* 90.1 (2009): 284-292.

Freeman, A.M., J.A. Herriges, and C.L. Kling. 2014. *The Measurement of Environmental and Resource Values: Theory and Methods*, 3rd edition, RFF Press.

Gao, Zhifeng, Lisa A. House, and Jing Xie. "Online survey data quality and its implication for willingness-to-pay: A cross-country comparison." *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* 64.2 (2016): 199-221.

Grandjean, Burke D., Nanette M. Nelson, and Patricia A. Taylor. "Comparing an internet panel survey to mail and phone surveys on willingness to pay for environmental quality: a national mode test." *64th annual conference of the American association for public opinion research* (2009).

Herriges, Joseph, Catherine Kling, Chi-Chen Liu, and Justin Tobias. “What are the consequences of consequentiality?” *Journal of Environmental Economics and Management* 59 (2010):67–81.

Johnston, R.J., F. Lupi, K. Moeltner, E. Besedin, Z. Yao, T. Ndebele, S. Crema, S. Peery, H. Kim and J.A. Herriges. “Do You Know Who’s Answering Your Survey? Expanding Threats to the Integrity of Online Panel Data in Environmental and Resource Economics.” *Association of Environmental and Resource Economists (AERE) Summer Conference* (2021) June 3-5, online.

Johnston, R.J., K.J. Boyle, W. Adamowicz, J. Bennett, R. Brouwer, T.A. Cameron, W. M. Hanemann, N. Hanley, M. Ryan, R. Scarpa, R. Tourangeau, and C.A. Vossler. "Contemporary guidance for stated preference studies." *Journal of the Association of Environmental and Resource Economists* 4.2 (2017): 319-405.

Kaplowitz, M., F. Lupi, and J. Hoehn. 2004. "Multiple-methods for developing and evaluating a stated preference survey for valuing wetland ecosystems. " In *Questionnaire Development, Evaluation, and Testing Methods*, (S. Presser, et al., eds). 503-524. Wiley:New Jersey.

McFadden, D. “Conditional logit analysis of qualitative choice behavior.” In P. Zarembka (Ed.), *Frontiers in econometrics.* New York: Academic Press (1974): 105–142.

Liebe, Ulf, et al. "Does the use of mobile devices (tablets and smartphones) affect survey quality and choice behaviour in web surveys?" *Journal of choice modelling* 14 (2015): 17-31

Lindhjem H., & Navrud S. “Are internet surveys an alternative to face-to-face interviews in contingent valuation?” *Ecol Econ* 70.9 (2011a):1628–1637

Lindhjem H., & Navrud S. “Using internet in stated preference surveys: a review and comparison of survey modes.” *Int Rev Environ Resour Econ* 5.4 (2011b): 309–351

Lupi, F., J.A. Herriges, H. Kim, and R.J. Stevenson. Getting off the Ladder: Disentangling Water Quality Indices to Enhance the Valuation of Divergent Ecosystem Services, working paper, July 2022.

Olsen, Søren Bøye. "Choosing between internet and mail survey modes for choice experiment surveys considering non-market goods." *Environmental and Resource Economics* 44.4 (2009): 591-610.

Owens, Joel, and Erin M. Hawkins. "Using online labor market participants for nonprofessional investor research: A comparison of MTurk and Qualtrics samples." *Journal of Information Systems* 33.1 (2019): 113-128.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. "Running experiments on amazon mechanical turk." *Judgment and Decision making* 5.5 (2010): 411-419.

Pennay, Darren, et al. "The Online Panels Benchmarking Study: a total survey error comparison of findings from probability-based surveys and non-probability online panel surveys in Australia." *Australian National University: Center for Social Research & Methods*. 2 (2018).

Roshwalb, Alan, Zachary Lewis, and Robert Petrin. "The Efficacy of Nonprobability Online Samples." *JSM Proceedings, Survey Research Methods Section* (2016): 3657-3666.

Roulin, Nicolas. "Don't throw the baby out with the bathwater: Comparing data quality of crowdsourcing, online panels, and student samples." *Industrial and Organizational Psychology* 8.2 (2015): 190.

Sandorf, Erlend Dancke, Lars Persson, and Thomas Broberg. "Using an Integrated Choice and Latent Variable Model to Understand the Impact of "Professional" Respondents in a Stated Preference Survey." *Resource and Energy Economics* (2020): 101178.

Teitcher, Jennifer EF, et al. "Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs." *The Journal of Law, Medicine & Ethics* 43.1 (2015): 116-133.

Valliant, R., & Dever, J. A. "Survey weights: a step-by-step guide to calculation." College Station, TX: Stata Press (2018).

Weinberg, Jill D., Jeremy Freese, and David McElhattan. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample." *Sociological Science* 1 (2014).

Whitehead, John C., et al. "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." Appalachian State University: Department of Economics Working Paper. (2021): 21-13.

Windle, Jill, and John Rolfe. "Comparing Responses from Internet and Paper-Based Collection Methods in More Complex Stated Preference Environmental Valuation Surveys." *Economic Analysis & Policy* 41.1 (2011).

Yeager, David S., et al. "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples." *Public opinion quarterly* 75.4 (2011): 709-747.

Zack, Elizabeth S., John Kennedy, and J. Scott Long. "Can nonprobability samples be used for social science research? A cautionary tale." *Survey Research Methods* 13.2 (2019).

## APPENDIX A: SURVEY SECTIONS OVERVIEW

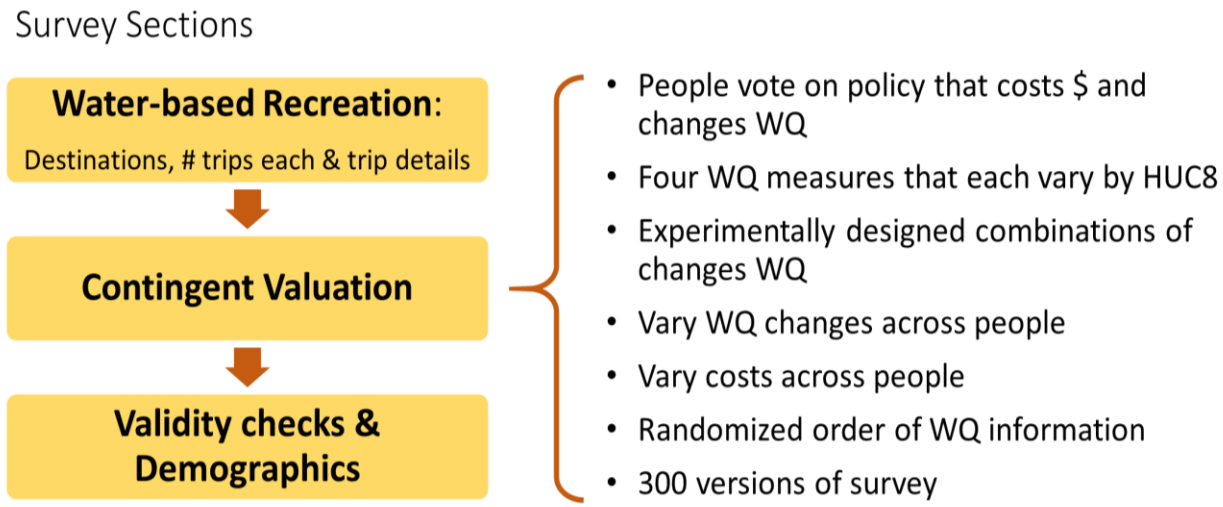


**Figure 6.** An overview of the survey sections

## APPENDIX B: SELECTED SURVEY SECTIONS

**Water Quality Indices Introduction:**

**II. This part of the survey is about the current conditions of Michigan's water resources.**

Water is essential to human life and to the health of the environment. Good water quality in lakes, rivers, and streams allows for a wide range of human uses and supports a rich and varied community of plants and animals.

However, measuring water quality can be tricky. Conditions that are good for one use of a lake or stream may not be ideal for another use. For example, a perfectly clear and sterile water body may be ideal for swimming, but would not support a diverse fish population.

In this survey, we will describe measures of water quality that focus on four categories:

- **Water Clarity** – The clarity of waterbodies (how far one can see through the water).
- **Water Contact** – The suitability of waterbodies for contact such as wading and swimming.
- **Fishing Water Quality** – The suitability of waterbodies for recreational fishing
- **Wildlife Water Quality** – The ability of waterbodies to support healthy and diverse populations of *naturally* occurring aquatic plants and animals.

**Figure 7.** Introduction to the water quality portion of the survey

**Water Clarity Introduction:**

*Water Quality Scores*

Water quality is often measured with separate water quality scores. A water quality score is a way to express the effects of important water quality characteristics (such as water clarity, bacteria levels, fishing quality and aquatic wildlife quality). All the water quality scores used in this survey range from 0 (worst possible quality) to 100 (best possible quality).

*A Water Clarity Score*

Scientists can create a water clarity score by measuring how far into the water a person can see. Water clarity is naturally higher in a typical lake than in a typical river. To make a water clarity score, scientists measure:

- How far into a lake or river someone can see an object

As the score increases, the water clarity of the waterbody increases.

Levels of the Water Clarity Score:

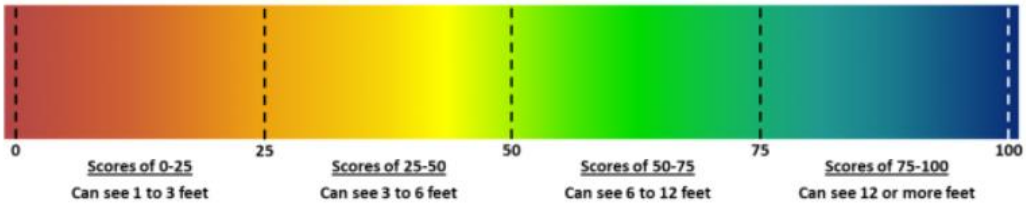


**Figure 8.** Introduction to the water clarity portion of the survey

**Water Contact Introduction:**

*A Water Quality Score for Water Contact*

People making contact with waterbodies may face a risk of illnesses such as diarrhea. The risk of water contact increases when levels of harmful bacteria increase. Scientists can measure bacteria levels in the water to determine a water quality score for recreational water contact such as boating, wading and swimming. A key bacterial indicator is:

- Fecal coliform - bacteria from sewage and animal waste

As the water contact score increases, the waterbody is more suitable for different water contact activities.

Levels of the Water Quality Score for Water Contact:

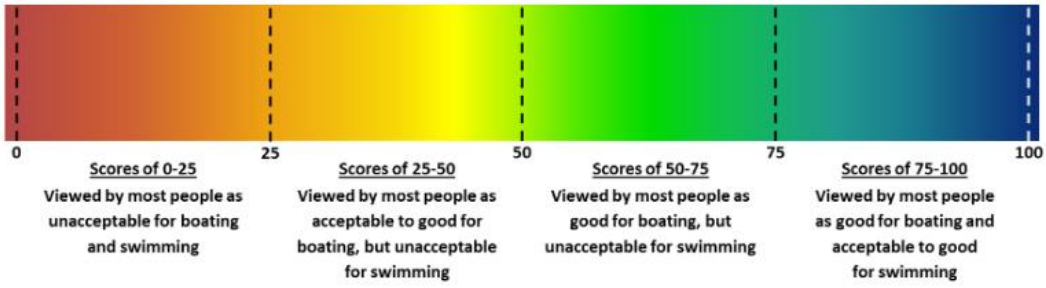


**Figure 9.** Introduction to the water contact portion of the survey

**Fishing Water Quality Introduction:**

*A Recreational Fishing Water Quality Score*

Scientists combine several measures of water quality to determine the water quality score for recreational fishing for game fish. Some of those measures are:

- Catch rates for warm water species such as panfish and bass
- Catch rates for cool water species such as walleye
- Catch rates for cold water species such as trout and salmon

As the score increases, fishing at the waterbody improves.

Levels of the Recreational Fishing Water Quality Score:

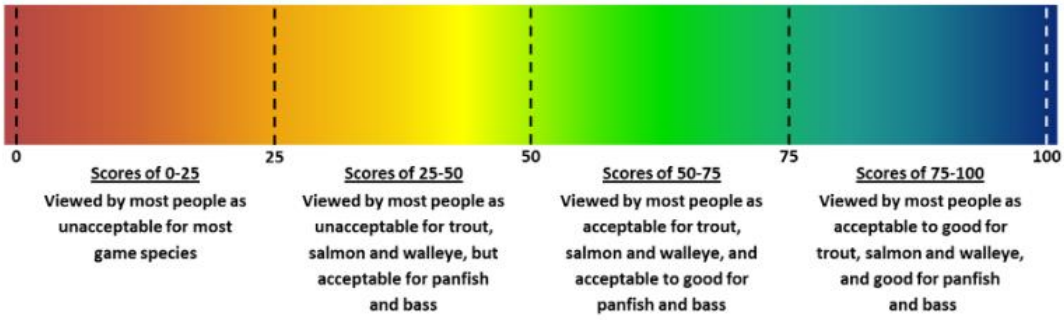


**Figure 10.** Introduction to the fishing water quality portion of the survey

**Wildlife Water Quality Introduction:**

*A Wildlife Water Quality Score*

Scientists rate the health of an **aquatic** ecosystem by combining measures related to the variety, abundance, and condition of naturally occurring animal and plant species that live in the water. Scientists also consider ecosystem functions that include processes such as photosynthesis and the cycling of nutrients, energy and organic matter.
Some of those measures include:

- Species richness - the number of naturally occurring aquatic species
- Sensitive species - number of natural species that are sensitive to alterations of the ecosystem
- Ecological function - the current condition of the ecosystem relative to its natural condition

The wildlife score of a lake, river, or stream is measured relative to its **natural state** before any changes due to human activity (e.g., pollution, development, etc). The more similar the aquatic wildlife measures are to their natural levels, the higher the wildlife score the water body receives.

Levels of the Wildlife Water Quality Score:

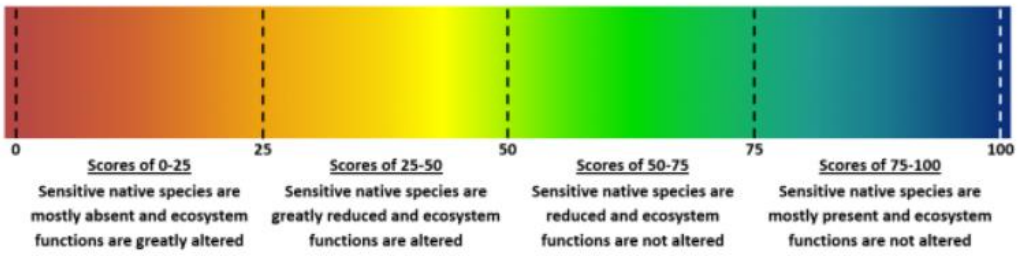


**Figure 11.** Introduction to the wildlife water quality portion of the survey

**Introduction to policy scenario and vote:**

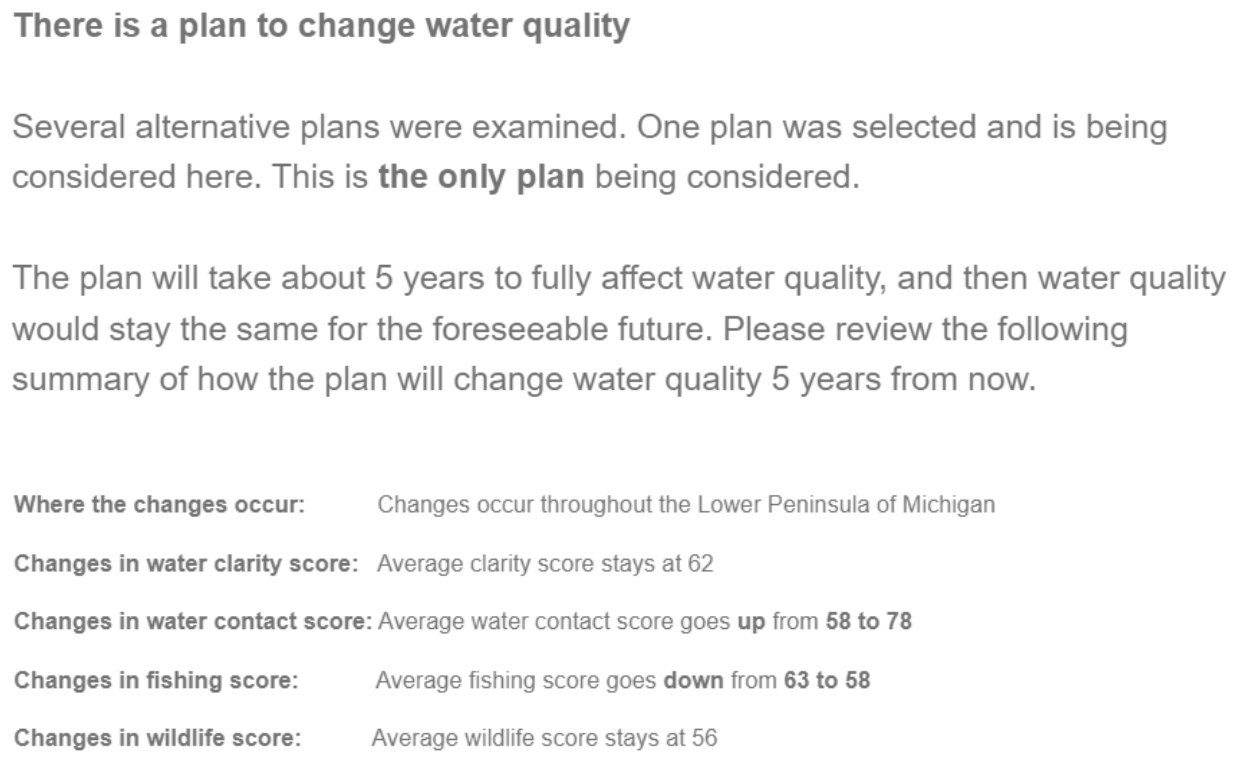


**Figure 12.** Introduction to the policy scenario portion of the survey

**Example of one-time payment:**

Based on your answers, the one time cost to your household would be **$195.**

**Your vote**: There are valid reasons you might vote **for** or **against** the plan for your one-time cost of $195. Some people may vote for the plan because they feel the water quality changes are worth their cost.

Some people will vote **against** the plan because

- they feel the water quality changes are not worth their cost, or
- they prefer to spend the money on something else instead, or
- they have some other reason to vote against the plan.

Some people may like the plan, but still vote **against** it because they feel

- the changes in water quality are too small, or
- there are not enough changes in areas they care about, or
- there are not enough changes in the water quality score they care about

Whatever your reasons, a vote for or against the plan is legitimate. We need you to consider the water quality changes and your cost, and then decide what is best for you.

I understand I can vote **for** or **against** the program, and I should pick what is best for my household.

**Figure 13.** Introduction to the payment vehicle portion of the survey

**Voting Page:**

Before voting, please keep the following in mind:

- This is the only plan under consideration.
- Survey results **will be shared** with policy-makers and agencies that may implement this plan.
- If the plan is implemented, the one-time cost to your household is $195.

| How the plan changes water quality in Michigan's Lower Peninsula | |
|---|---|
| **Where the changes occur** | Changes occur **throughout the Lower Peninsula of Michigan**. |
| **Changes in water clarity score** | Average water clarity score goes stays at 62 |
| **Changes in water contact score** | Average water contact score goes goes **up** from **58 to 78** |
| **Changes in fishing score** | Average fishing score goes goes **down** from **63 to 58** |
| **Changes in wildlife score** | Average wildlife score goes stays at 56 |
| **One-time cost to your household** | **$ 195** |

17. Considering that the policy would change water quality as described above, do you vote for or against the plan, which will cost your household the onetime tax of $195?

I vote **for** the plan

I vote **against** the plan

18. Please share some reasons for your answer to question 17.

**Figure 14.** The voting portion of the survey

## APPENDIX C: ROBUSTNESS CHECKS

To test the sensitivity of these logit results, we conducted various robustness tests. These included three procedures: imputation, raking, and an attention check. The goal of imputation was to replace any socio-demographic data that was missing due to non-response. First, education, children, employment, and gender[4] were recoded as binary variables (i.e., no bachelor's/bachelor's and above, no children/children, unemployed/employed, and male/female). Respondents who had item nonresponse for these variables were assigned a random number between 0 and 1 for each piece of missing data. If the random number was lower than the mean for each variable, then the respondent was assigned a 1 and if the random number was higher than the mean they were assigned a 0. For those missing income data, the respondent was assigned a value equal to the mean income of completed responses. All missing data replacement was done individually by source. Nonresponse represented a very small portion of the sample: 1.2%, 1.8%, 1.1%, 1.6% and 1.4% for education, children, employment, gender, and income, respectively. After all missing data was replaced, we applied the new data to our logit model that included socio-demographic information. The results can be seen in Appendix D, labeled as Model 3 and 4. Overall, imputation did not change the general pattern of our results. The signs and significance of all cost and water quality parameters remained similar.

Next, we applied a raking procedure to this imputed sample, which involved weighting the data and trimming large weights. Raking is a process where observations in selected groups are weighted so that the relative size of each group in the sample matched the relative size of each group in the American Community Survey. The selected groups involved classifications of

[4] Less than 0.3% of the sample identified as "other" for gender. For these people, data was imputed using the same randomized procedure as for the item nonresponse to ensure the entire sample was represented in a binary fashion.

gender, education, children, income, and age. Extreme weights were trimmed by group and were considered any weight that was five times the mean. After the extreme weights were timed, the data was reweighted using the same raking procedure. This process of raking, trimming, and reweighting was repeated until no extreme weights were produced (Valliant & Dever. 2018). For ABS and Qualtrics, the data was re-weighted only once. For Mturk, the data was re-weighted 50 times. While raking did not change the sign and significance of the cost parameters, it did decrease the significance of several water quality parameters and it did so across samples. So the overall pattern remained the same, but the results became weaker. Results can be seen in Appendix D as Models 5 and Model 6 (where Model 5 has no socio-demographic information and Model 5 includes socio-demographic parameters).

## APPENDIX D: MODELS 2-5†

**Table 2.** Additional Logit Models

| | Model 2<br>b/se | Model 3<br>b/se | Model 4<br>b/se | Model 5<br>b/se | Model 6<br>b/se |
|---|---|---|---|---|---|
| Vote_yes | | | | | |
| Address_Cons | -0.1705 | 0.4136*** | 0.0246 | 0.3608** | 0.3743 |
| | -0.2752 | -0.1172 | -0.2614 | -0.1302 | -0.2874 |
| MTurk_Cons | 0.8562** | 0.7622*** | 0.9047** | 0.7237** | 0.9621* |
| | -0.308 | -0.1718 | -0.3051 | -0.2757 | -0.427 |
| Qualtrics_Cons | 0.8886*** | 0.5215*** | 0.9158*** | 0.5557*** | 0.7781** |
| | -0.2179 | -0.0963 | -0.2175 | -0.1172 | -0.2562 |
| Cost*Address | -0.0014*** | -0.0012*** | -0.0013*** | -0.0011*** | -0.0012*** |
| | -0.0001 | -0.0001 | -0.0001 | -0.0001 | -0.0002 |
| Cost*MTurk | -0.0018*** | -0.0018*** | -0.0018*** | -0.0016*** | -0.0018*** |
| | -0.0002 | -0.0002 | -0.0002 | -0.0003 | -0.0003 |
| Cost*Qualtrics | -0.0013*** | -0.0013*** | -0.0013*** | -0.0012*** | -0.0012*** |
| | -0.0001 | -0.0001 | -0.0001 | -0.0001 | -0.0001 |
| DelCLS_ABS | 0.0140* | 0.0140* | 0.0135* | 0.0136* | 0.0131* |
| | -0.0058 | -0.0054 | -0.0055 | -0.0061 | -0.0061 |
| DelCLS_Mturk | 0.0192* | 0.0190* | 0.0194* | 0.0054 | 0.0071 |
| | -0.0085 | -0.0083 | -0.0085 | -0.0126 | -0.0129 |
| DelCLS_Qualtrics | 0.0135** | 0.0138** | 0.0133** | 0.0069 | 0.0066 |
| | -0.0046 | -0.0045 | -0.0046 | -0.0055 | -0.0055 |
| DelWCS_ABS | 0.0177** | 0.0117* | 0.0132* | 0.0058 | 0.0063 |
| | -0.0058 | -0.0054 | -0.0055 | -0.006 | -0.0061 |
| DelWCS_MTurk | 0.0385*** | 0.0369*** | 0.0387*** | 0.0219 | 0.0254* |
| | -0.0083 | -0.0082 | -0.0083 | -0.0129 | -0.0128 |
| DelWCS_Qualtrics | 0.0180*** | 0.0179*** | 0.0180*** | 0.0146** | 0.0146** |
| | -0.0046 | -0.0045 | -0.0046 | -0.0055 | -0.0056 |
| DelFBS_ABS | 0.0150** | 0.0166*** | 0.0157*** | 0.0170*** | 0.0160*** |
| | -0.0046 | -0.0043 | -0.0043 | -0.0047 | -0.0048 |
| DelFBS_MTurk | 0.0261*** | 0.0237*** | 0.0253*** | 0.0311** | 0.0339*** |
| | -0.0066 | -0.0065 | -0.0066 | -0.0102 | -0.0101 |
| DelFBS_Qualtrics | 0.0027 | 0.0021 | 0.0028 | 0.0008 | 0.0016 |
| | -0.0036 | -0.0036 | -0.0036 | -0.0044 | -0.0044 |
| DelWLS_ABS | 0.0132* | 0.0135* | 0.0135* | 0.0113 | 0.0117 |
| | -0.0057 | -0.0053 | -0.0054 | -0.0059 | -0.006 |
| DelWLS_MTurk | 0.0376*** | 0.0344*** | 0.0368*** | 0.0410** | 0.0504*** |
| | -0.0083 | -0.0081 | -0.0083 | -0.0127 | -0.0125 |
| DelWLS_Qualtrics | 0.0115* | 0.0112* | 0.0111* | 0.0168** | 0.0172** |
| | -0.0045 | -0.0045 | -0.0045 | -0.0054 | -0.0055 |
| Gender_ABS | 0.2347* | | 0.2137* | | 0.0977 |
| | -0.1007 | | -0.0951 | | -0.1066 |
| Children_ABS | -0.0266 | | -0.1862 | | -0.1567 |
| | -0.0556 | | -0.1159 | | -0.1306 |

**Table 2** (cont'd)

| | | | | | |
|---|---|---|---|---|---|
| Eduction_ABS | 0.2552* | | 0.2207* | | 0.1737 |
| | -0.106 | | -0.0994 | | -0.1047 |
| Employment_ABS | 0.0178 | | 0.0236 | | -0.0985 |
| | -0.115 | | -0.1079 | | -0.1213 |
| Age_ABS | 0.0006 | | -0.0042 | | -0.0083* |
| | -0.0036 | | -0.0034 | | -0.0038 |
| Income_ABS | 0.0000*** | | 0.0000*** | | |
| | 0 | | 0 | | 0 |
| Gender_MTurk | -0.3161* | | -0.3467* | | -0.3006 |
| | -0.1426 | | -0.1419 | | -0.2178 |
| Children_MTurk | 0.4546** | | 0.4345** | | 0.5855** |
| | -0.1445 | | -0.1437 | | -0.2106 |
| Education_MTurk | 0.3081* | | 0.2852 | | 0.6434* |
| | -0.1507 | | -0.1497 | | -0.2624 |
| Employment_MTurk | -0.0175 | | -0.0226 | | 0.046 |
| | -0.1634 | | -0.1622 | | -0.2275 |
| Age_MTurk | -0.0082 | | -0.0074 | | -0.0119 |
| | -0.0058 | | -0.0057 | | -0.0072 |
| Income_MTurk | 0 | | 0 | | 0 |
| | 0 | | 0 | | 0 |
| Gender_Qualtrics | -0.0551 | | -0.0557 | | -0.1306 |
| | -0.0773 | | -0.0772 | | -0.0935 |
| Children_ Qualtrics | -0.1122 | | -0.1203 | | -0.1994 |
| | -0.109 | | -0.1089 | | -0.1227 |
| Education_ Qualtrics | 0.1824* | | 0.1861* | | 0.1211 |
| | -0.0819 | | -0.0818 | | -0.0955 |
| Employent_ Qualtrics | 0.0207 | | 0.0181 | | 0.3003** |
| | -0.0874 | | -0.0873 | | -0.1097 |
| Age_ Qualtrics | -0.0129*** | | -0.0133*** | | -0.0125*** |
| | -0.0029 | | -0.0029 | | -0.0034 |
| Income_ Qualtrics | <0.0000*** | | <0.0000*** | | <0.0000*** |
| | 0 | | 0 | | 0 |
| Chi-sqr | 590.859 | 473.264 | 590.693 | 262.797 | 428.607 |
| AIC | 7492 | 7907 | 7783 | 7968 | 7795 |
| BIC | 7732.6 | 8027.8 | 8024.2 | 8088.6 | 8037 |
| N | 5863 | 6071 | 6071 | 6071 | 6071 |

* p<0.05, ** p<0.01, *** p<0.001

† Model 2: Core model and sociodemographic information
Model 3: Imputed data
Model 4: Imputed data and sociodemographic information
Model 5: Imputed then raked data
Model 6: Imputed then raked data and sociodemographic information

## APPENDIX E: TWTP FOR A RANGE OF NON-MARGINAL CHANGES

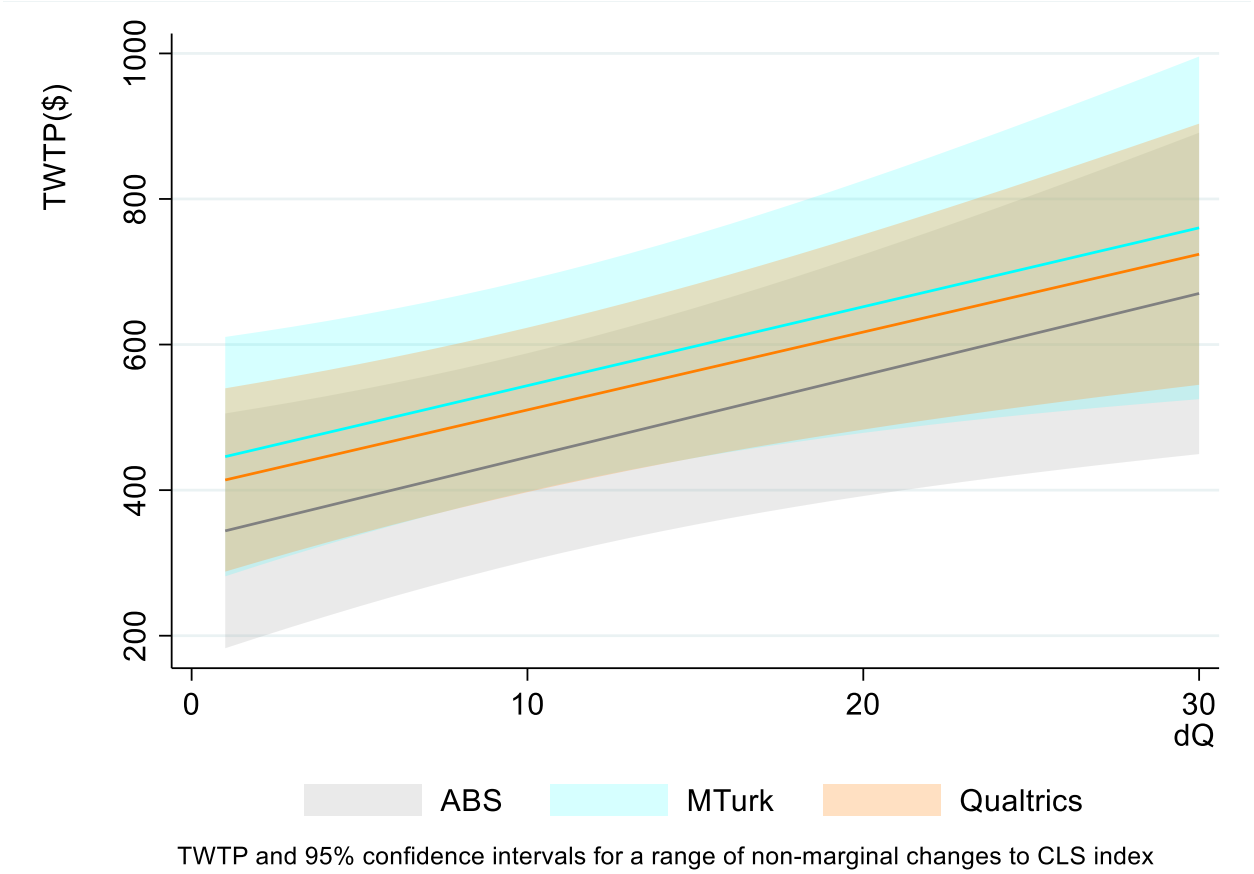


**Figure 15.** TWTP for a range of non-marginal changes to CLS index

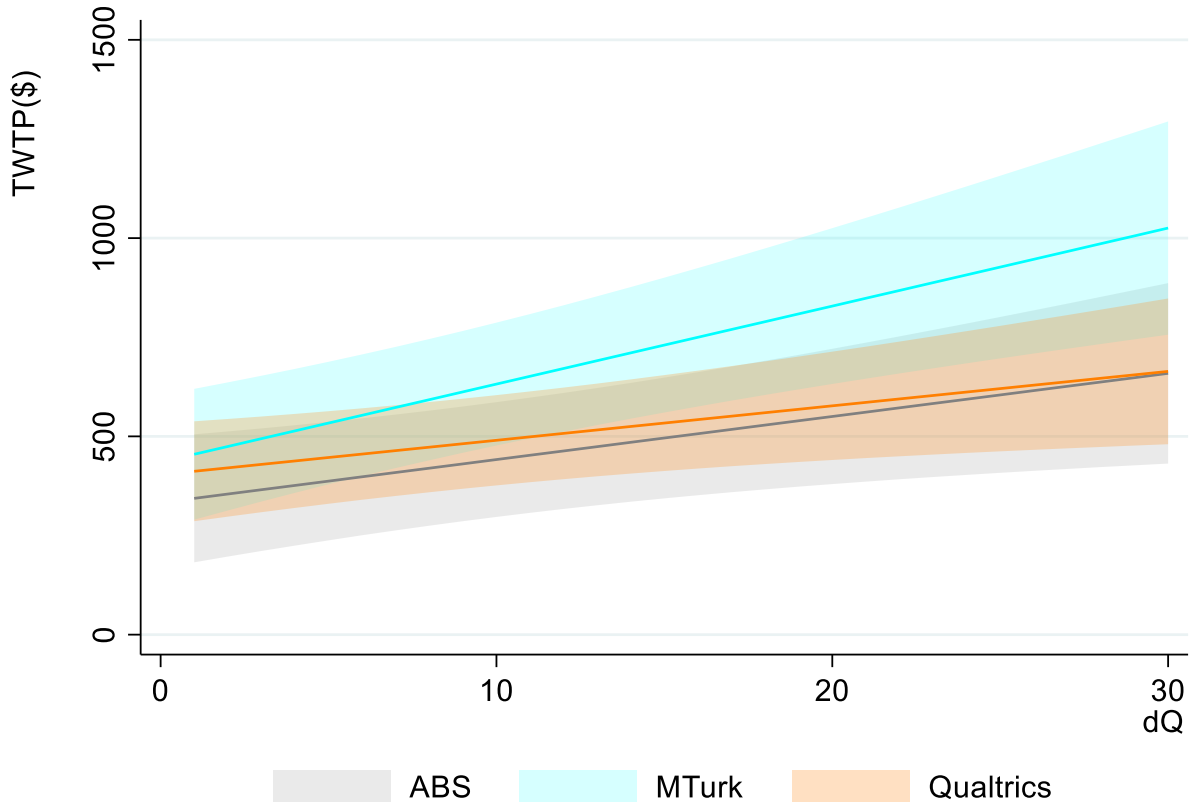


**Figure 16.** TWTP for a range of non-marginal changes to WLS index

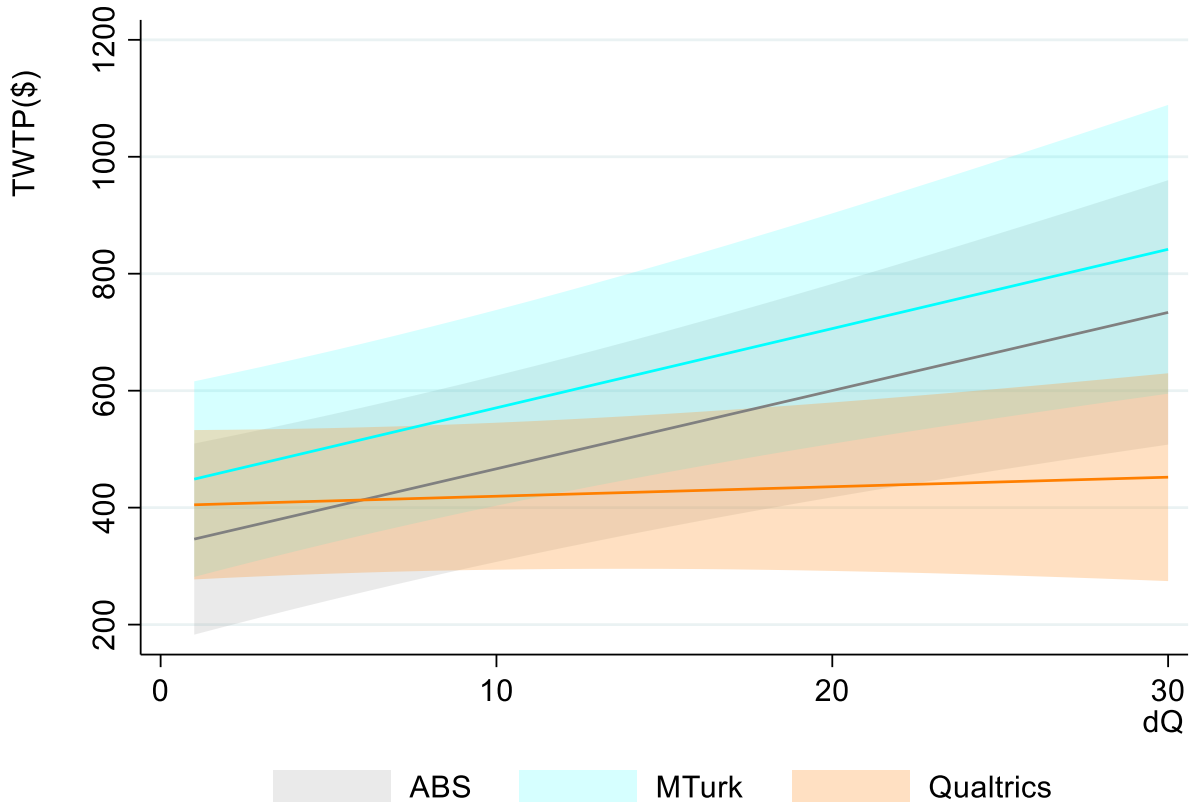


**Figure 17.** TWTP for a range of non-marginal changes to FBS index

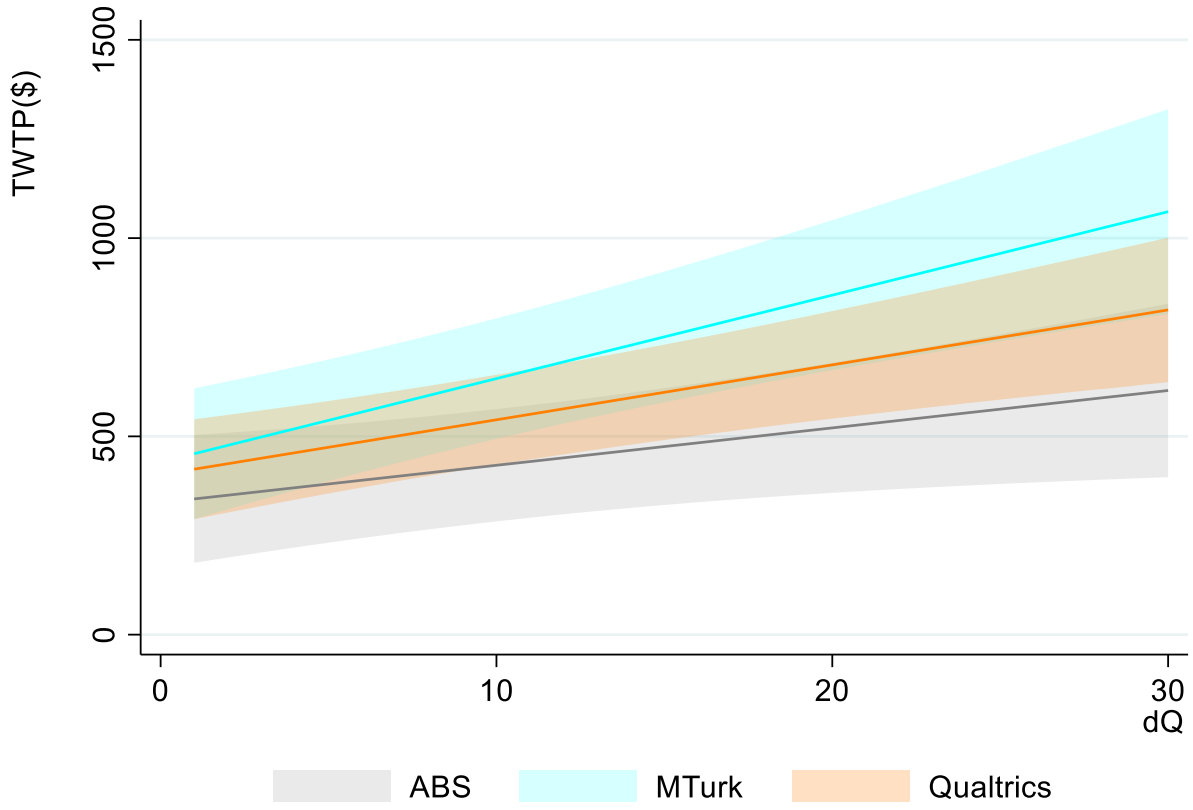


**Figure 18.** TWTP for a range of non-marginal changes to WCS index

## APPENDIX F: MWTP FOR CHANGES IN INDICES

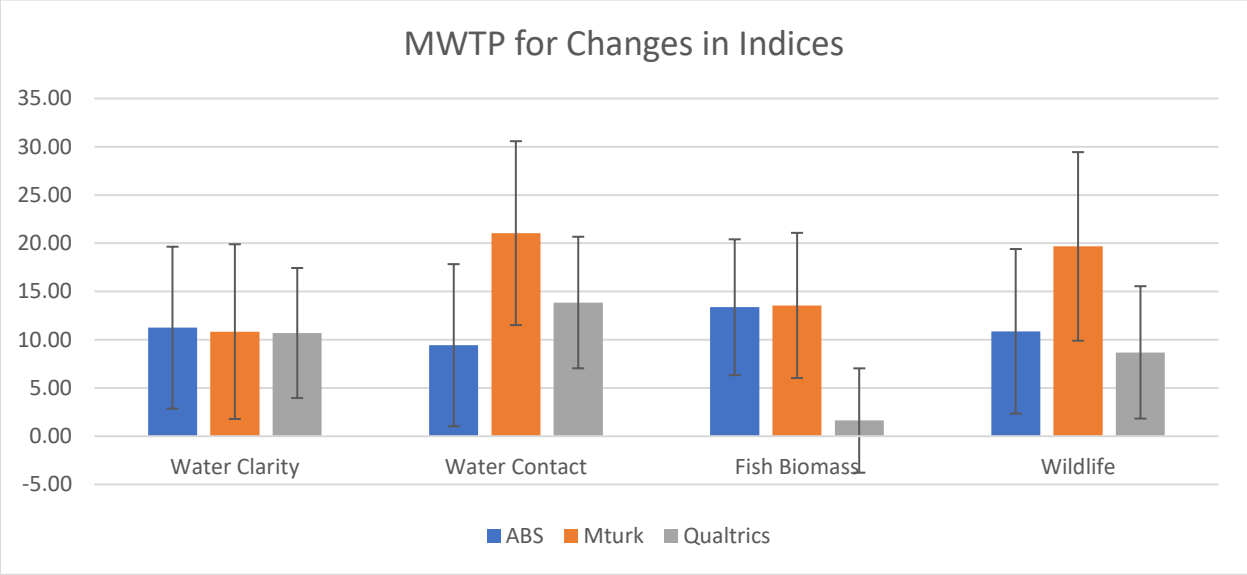


**Figure 19.** MWTP for changes in indices with error bars

## APPENDIX G: ATTRIBUTE TABLE

**Table 3.** Table of attributes and levels

| Attribute | Attribute Level |
|---|---|
| Cost | $65, $195, $435, or $965 |
| Delta Water Clarity Score | 0, 5, 10, 15, or 20 points |
| Delta Water Contact Score | 0, 5, 10, 15, or 20 points |
| Delta Fish Biomass Score | -5, 0, 5, 15, or 20 points |
| Delta Wildlife Score | 0, 5, 10, 15, or 20 points |
| Note: The order the indices were presented and summarized was randomized across respondents | |

# CHAPTER 2: Effect of Controlling for Heterogeneity in Preferences on Valuation Results

## 2.1: Introduction

As discussed in the previous chapter, one of the major concerns associated with non-probability sampling are potential biases that may not be mitigated by balancing samples to "represent" population demographics (Baker et al. 2010). These concerns are particularly relevant given that the non-probability methods in our study generate several key valuation results than that differ in the probability-based samples, even after raking. One possible way to account for biases in coefficient estimates is to control for preference heterogeneity. Preference heterogeneity has long been recognized in SP studies, and it is recommended that analyses of SP data should include observed and unobserved preference heterogeneity (Johnston et al. 2017). Identifying preference heterogeneity is important for policy implications since it may affect the generalizability of results. Unmodeled heterogeneity can lead to a scenario where regression coefficients are correct for the sample, but incorrect for the population. If heterogeneity isn't included in the analysis, then external validity is weakened, and the results aren't generalizable for the population. Preference heterogeneity is particularly relevant to nonprobability sampling because these samples tend to overrepresent certain populations. Depending on how sociodemographic characteristics are related to the outcome variable and the various attribute variables, this overrepresentation may negatively impact the validity of the findings (Pasek 2016). In order to remedy this issue, it is necessary to identify whether preference heterogeneity exists between the samples, and if so, control for them.

To identify preference heterogeneity, we compare the results of the logit model to a logit model with interaction terms, a mixed logit model, and a latent class model. We find that the

logit model is the best fit for the data, given that the other heterogeneity models often do not produce significant results or do not converge.

## 2.2: Literature review

One reason that preference estimates may differ between samples is due to preference heterogeneity. There are three main types of heterogeneity, described by Johnston et al., 2017. First, heterogeneity can occur when there are discrete groups of people, and each group exhibits a different set of preferences (which often calls for a latent class approach). Second, preferences can vary continuously across all individuals (which often calls for a mixed logit model with randomly varying parameters). Third, there may be one shared set of preferences across all respondents, but the error term has differing dispersions (which often class for an error component model). We focus on the first two types in this paper.

Preference heterogeneity is well explored in environmental SP literature, often using a logit model with interaction terms, a mixed logit model, or a latent class model. Additionally, there are many studies that compare preference heterogeneity across these models (Bujosa et al., 2010; Colombo et al., 2009; Hynes et al., 2008; Yoo & Ready, 2014), with some specifically doing so in the context of water quality improvements (Andrepoulos et al., 2015; Biro et al., 2006; Chen et al., 2018; Khan et al., 2019; Kosenius, 2010). Many find that preference heterogeneity exists (Birol et al., 2016, Hynes et al., 2008) and is explained by sociodemographic characteristics (Bujosa et al.,2010; Chen et al., 2018; Khan et al., 2019; Kosenius, 2010; Yoo & Ready 2014). Most studies find that the models produce similar welfare estimates for the population (Andrepoulos et al. 2015; Colombo et al., 2009; Hynes et al., 2008), while one found that the models produce different welfare estimates (with the mixed logit model producing higher estimates) (Bujosa et al., 2010). Many find that the latent class model is the best fit and

better captures preference heterogeneity (Andrepoulos et al., 2015; Birol et al., 2016; Bujosa et al., 2010; Colombo et al., 2009; Kosenius), while some find that the mixed logit model was the better fit (Yoo & Ready, 2014). So, generally speaking, studies that compare preference heterogeneity models find that the models produce similar estimates for the population, but that the latent model fits the data better and better explains preference heterogeneity.

Although preference heterogeneity is well explored in SP literature, not many studies focus on preference heterogeneity in non-probability sampling. One such study is by Thompson & Pickett (2019), which compares results about criminal justice attitudes from five online non-probability samples that were drawn from either a crowdsourcing platform, Amazon Mechanical Turk (MTurk), or an opt-in panel platform, SurveyMonkey Audience to those from the General Social Survey (GSS). They found that their results were likely biased due to effect heterogeneity, since the non-probability samples generated results that were not of the same magnitude as the GSS coefficients, diminishing external validity.

This chapter investigates whether preference heterogeneity exists between the samples and, if so, whether any differences are minimized after controlling for preference heterogeneity. Currently, there is a need for research on preference heterogeneity in non-probability environmental SP surveys and our research seeks to contribute to this growing area. This area is particularly relevant because it is not only important to identify whether there are differences between non-probability and probability sample, but also to determine whether we can control for these differences. My findings to date are mixed. Although all conditional logits with heterogeneity captured with interaction converge, the RP & LC models generally do not.

## 2.3: Econometric Methods and Specification

*Binary Logit Model*

The conditional logit model identifies the population average preference, which is akin to assuming preference homogeneity across respondents since a single utility parameter estimate is estimated for each attribute. When moving away from the population average results, the conditional logit implies that all respondents have the same tastes for each attribute. To incorporate preference heterogeneity within the conditional logit approach, socioeconomic variables must be included as interactions with attributes or as interactions with alternative-specific constants.

To estimate the model, we use a binary logit model. Thus, we model the probability of individual *i* voting yes as $Pr_1$:

$$Pr_{i1} = \frac{1}{1+e^{-(\beta Q_{i1})}} = \frac{e^{\alpha\prime_i+\beta dQ_i-\theta P_i}}{1+e^{\alpha\prime_i+\beta dQ_i-\theta P_i}}$$

where *Pr* is the probability that respondent *i* will respond with a yes vote, β is a vector of marginal utility parameters to be estimated, $\alpha_{i1}$ are alternative specific regressors, and Q is a vector of explanatory variables for the four water quality indices.

*Mixed Logit Model*

The mixed logit model incorporates preference heterogeneity by allowing the utility parameters to vary across respondents. For the mixed logit model, we assume that there is a continuous distribution of $f(\beta_i)$ in the population. To get the unconditional choice probabilities, The unconditional choice probabilities are the integrals of the standard logit model over a density of parameters $\beta_i$ which depends on parameters $\theta$ (Train 2009).:

$$Pr_{i1} = \int \left(\frac{e^{\alpha\prime_i+\beta dQ_i-\theta P_i}}{1+e^{\alpha\prime_i+\beta dQ_i-\theta P_i}}\right) f(\beta_i|\theta)d(\beta_i)$$

where *Pr* is the probability that respondent *i* will respond with a yes vote, $\beta_i$ is a vector of marginal utility parameters to be estimated, $\alpha_{i1}$ are alternative specific regressors, and $f(\beta_i)$ is a density function.

Integral 5 can be approximated via simulation. The simulated probability is given by:

$$\tilde{P}_{i1} = \sum_{r} \left(\frac{e^{\alpha\prime_i + \beta dQ_i - \theta P_i}}{1 + e^{\alpha\prime_i + \beta dQ_i - \theta P_i}}\right)/R$$

where *R* is a set of draws from *f*(β).

*Latent Class Model*

The latent class model incorporates preference heterogeneity via discrete distributions, whereas the mixed logit uses continuous distributions. The latent class model assumes that a population is comprised of classes, or different "types" of people. Within each type or "class", preferences are assumed to be homogenous, but preferences are assumed to be heterogenous across classes. The latent class model is constructed by linking a series of conditional logit models, where each class (C) has its own conditional logit model. So if there are four classes, there will be four conditional logit models. Each class conditional logit model will then have its own set of utility functions. The link between the series of conditional logit models is also a conditional logit model, called a class assignment model. In the class assignment model, each individual has a probability of belonging to each class. In addition, each individual has a probability of choosing one of the alternatives (conditional on belonging to the class). In sum, the latent class logit model explains preference heterogeneity across individuals conditional on the probability of membership in a latent class.

For the latent class model, the unconditional probability of the observed panel of choices is a weighted average over the c classes with weight $\pi_c$. This is given by:

$$P_{i1|c} = \sum_{c} \pi_{ic} \prod_{t=1}^{T} [\frac{e^{\alpha\prime_i + \beta dQ_i - \theta P_i}}{1 + e^{\alpha\prime_i + \beta dQ_i - \theta P_i}}]$$

where class assignment is given by:

$$\pi_{ic} = \frac{e^{V_{ic}}}{\sum_{c \in C} e^{V_c}}$$

## 2.4: Results

*Logit Model + Socio-demographic Interaction Terms*

We incorporated preference heterogeneity using a conditional logit model with interaction terms, a mixed logit model, and a latent class model. The logit model with interaction terms is the same as the original logit model, except it includes interaction terms between sociodemographic variables (employment, education, children, income, and gender) and each water quality variable and sample dummy variables. Thus, for each sample the interactions together show the degree to which the different samples vary by sample's demographics with interactions in the model and the utilities and water quality utilities. Coefficients, standard errors, test statistics, and marginal willingness to pay (MWTP) are reported in Table 2.1. As expected, the coefficient on cost is significantly negative and the nearly all the coefficients on water quality indices are significantly positive in all samples. Of all the interaction terms, only the interaction between employment and Qualtrics, education and Qualtrics, education and MTurk, children and Mturk, income and ABS, income and Qualtrics, and gender and Qualtrics are significant. For these significant interaction terms, all have a positive relationship with the outcome variable, except the employment and gender interaction terms. These significant interaction terms indicate that on average preference heterogeneity does exist, especially in the non-probability samples. However, the preference heterogeneity is not the same across samples.

Several tests yield mixed evidence on differences across samples. Unlike in the original logit model, half of the parameters were significantly different across samples: fish biomass, wildlife score, children, income, gender, and the constants. Interestingly, the cost, water clarity, and water contacts scores are not significant in the interaction model but were significant in the original model. This suggests that incorporating preference heterogeneity controls for differences in cost and some water quality parameters. We also examined the marginal willingness to pay (MWTP) for individual water quality indices. Most were not significantly different across the three sources; as before in the conditional logit without heterogeneity, the one exception was the low and insignificant MWTP for fish biomass in the Qualtrics sample.

*Mixed Logit Model*

After estimating the logit models, we used a mixed logit model to incorporate preference heterogeneity. For this model, the interaction terms between the water quality indices and the sample sources were specified as random parameters. However, this model did not converge. To simplify, we ran a mixed logit model for each sample source. In addition, we ran the models using different estimation commands and different specified numbers of Halton draws (nrep). Most models did not converge, and it was not consistent across commands. In addition, these models yielded almost no significant heterogeneity (the standard deviations for the distribution of the parameters).

Using Arne Hole's mixlogit routine in Stata with 10 nreps, the model converges for every sample. However, this is a very low number of draws to be reliable. Using mixlogit with 50 nreps and no constant, MTurk did not converge but ABS and Qualtrics did. Using mixlogit with 50 nreps and a constant, ABS did not converge but MTurk and Qualtrics did. Using the cmmixlogit command in stata, none of the models converged. Next, we simplified the model

further and included only one water quality index in the model. We compared the results of mixlogit with 50 nreps and ccmixlogit with 50 draws. Of these, only CLS for MTurk and Qualtrics did not converge. However, these models produced almost no significant standard deviations of parameters. We also ran the model in NLogit, but none of the models converged.

*Latent Class Model*

5 latent class models were estimated, where model membership was specified by each of the sociodemographic variables. For each model, two classes were specified. Similar to the mixlogit models, the latent class models did not converge or did not produce significant heterogeneity. This is true when the models were estimated using the expectation-maximization (EM) algorithm, as well as when the models were estimated using maximum likelihood estimation.

## 2.5: Discussion

Identifying preference heterogeneity is an important step in the analysis of SP data since it may affect the generalizability of results. If heterogeneity is unaccounted for, it may be possible that the results are correct for the sample but unrepresentative for the population. This is especially true for non-probability data, since these samples tend to overrepresent certain segments of the population. Previous research has shown that preference heterogeneity often exits due to sociodemographic differences. While this has not been extensively studied in non-probability samples, one study has shown that such differences likely biased the results, with the non-probability sample generating different values than the probability sample (Thompson & Pickett 2019).

We attempt to investigate preference heterogeneity by estimating a conditional logit model with interaction terms, a mixed logit model, and latent class models. Overall, the results of

our preference heterogeneity research are inconclusive. The conditional logit model with interaction terms to capture heterogeneity converged and generally indicated that preference heterogeneity exists. This is especially true in the non-probability samples. However, the mixed logit models and the latent class models typically did not converge, and if they did there was almost no significant heterogeneity. There are several possibilities for this result. First, it could be that heterogeneity was simply not important in the samples and, thus, the models did not produce significant standard deviations for the distribution of the parameters. Second, it could be related to the data itself, whether it be a feature of the data quirks or limited sample sizes. Overall, the results did not give a clear indication of whether preference heterogeneity exists and whether it biases our results. Given these inconclusive results, the fact that preference heterogeneity in non-probability samples has not been widely studied, and that non-probability samples tend to overrepresent some populations, there is a call for continued research.

**Table 4.** Conditional Logit Model with Heterogeneity Interactions: Estimates and Statistics

| Variable | ABS | MTurk | Qualtrics | Differences Test |
|---|---|---|---|---|
| Cost | -0.0014*** | -0.0018*** | -0.0013*** | 3.55 |
| SE | (0.0001) | (0.0002) | (0.0001) | |
| Δ in Water Clarity Score | 0.0140* | 0.0192* | 0.0135** | 0.36 |
| SE | (0.0058) | (0.0085) | (0.0046) | |
| MWTP | 10.2897 | 10.8399 | 10.2180 | |
| Δ in Water Contact Score | 0.0117* | 0.0385*** | 0.0180*** | 5.15* |
| SE | (0.0058) | (0.0083) | (0.0046) | |
| MWTP | 13.0488 | 21.7751 | 13.6432 | |
| Δ in Fish Biomass Score | 0.0150** | 0.0261*** | 0.0027 | 11.14*** |
| SE | (0.0046) | (0.0066) | (0.0036) | |
| MWTP | 11.0161 | 14.7982 | 2.0548 | |
| Δ in Wildlife Score | 0.0132* | 0.0376*** | 0.0115* | 7.97** |
| SE | (0.0057) | (0.0083) | (0.0045) | |
| MWTP | 9.7360 | 21.2602 | 8.6863 | |
| Gender | 0.2347* | -0.3161* | -0.0551 | 10.85*** |
| SE | (0.1007) | (0.1426) | (0.0773) | |
| MWTP | 172.5685 | -178.8716 | -41.7372 | |
| Children | -0.0266 | 0.4546** | -0.1122 | 11.16*** |
| SE | (0.0556) | (0.1445) | (0.1090) | |
| MWTP | -19.5702 | 257.2743 | -85.1007 | |
| Education | 0.2552* | 0.3081* | 0.1824* | 0.66 |
| SE | (0.1060) | (0.1507) | (0.0819) | |
| MWTP | 187.6492 | 174.3655 | 138.2945 | |
| Employment | 0.0178 | -0.0175 | 0.0207 | 0.04 |
| SE | (0.1150) | (0.1634) | (0.0874) | |
| MWTP | 178.8961 | -9.9162 | 15.6993 | |
| Age | 0.0006 | -0.0082 | -0.0129*** | 8.62** |
| SE | (0.0036) | (0.0058) | (0.0029) | |
| MWTP | .4491 | -4.6622 | -9.7829 | |
| Income | 0.0000*** | -0.0000 | 0.0000*** | 12.23*** |
| SE | (0.0000) | (0.0000) | (0.0000) | |
| MWTP | .0037 | -.0005 | .0031 | |
| Constant | -0.1705 | 0.8562** | 0.8886*** | 10.24*** |
| SE | (0.2752) | (0.3080) | (0.2179) | |
| N | | | | 5863 |
| LogL | | | | -3710.1098 |
| Wald $\chi^2$ | | | | 590.86 |

*Note: ***, **, * represents significance at the 1%, 5%, and 10% level*

## BIBLIOGRAPHY


Andreopoulos, Dimitrios, et al. "Handling preference heterogeneity for river services' adaptation to climate change." *Journal of environmental management* 160 (2015): 201-211.

Baker et al. Prepared for the AAPOR Executive Council by a Task Force operating under the auspices of the AAPOR Standards Committee. "Research synthesis: AAPOR report on online panels." *Public Opinion Quarterly* 74.4 (2010): 711-781.

Birol, Ekin, Katia Karousakis, and Phoebe Koundouri. "Using a choice experiment to account for preference heterogeneity in wetland attributes: The case of Cheimaditida wetland in Greece." *Ecological economics* 60.1 (2006): 145-156.

Bujosa, Angel, Antoni Riera, and Robert L. Hicks. "Combining discrete and continuous representations of preference heterogeneity: a latent class approach." *Environmental and Resource Economics* 47.4 (2010): 477-493.

Chen, W. Y., Hua, J., Liekens, I., & Broekx, S. "Preference heterogeneity and scale heterogeneity in urban river restoration: A comparative study between Brussels and Guangzhou using discrete choice experiments". *Landscape and Urban Planning, 173* (2018): 9-22.

Colombo, Sergio, Nick Hanley, and Jordan Louviere. "Modeling preference heterogeneity in stated choice data: an analysis for public goods generated by agriculture." *Agricultural Economics* 40.3 (2009): 307-322.

Hynes, Stephen, Nick Hanley, and Riccardo Scarpa. "Effects on welfare measures of alternative means of accounting for preference heterogeneity in recreational demand models." *American journal of agricultural economics* 90.4 (2008): 1011-1027.

Johnston, R.J., K.J. Boyle, W. Adamowicz, J. Bennett, R. Brouwer, T.A. Cameron, W. M. Hanemann, N. Hanley, M. Ryan, R. Scarpa, R. Tourangeau, and C.A. Vossler. "Contemporary guidance for stated preference studies." *Journal of the Association of Environmental and Resource Economists* 4.2 (2017): 319-405.

Khan, Sufyan Ullah, et al. "Valuation of ecosystem services using choice experiment with preference heterogeneity: a benefit transfer analysis across inland river basin." *Science of the Total Environment* 679 (2019): 126-135.

Kosenius, Anna-Kaisa. "Heterogeneous preferences for water quality attributes: The case of eutrophication in the Gulf of Finland, the Baltic Sea." *Ecological Economics* 69.3 (2010): 528-538.

Thompson, A. J., & Pickett, J. T. "Are relational inferences from crowdsourced and opt-in samples generalizable? Comparing criminal justice attitudes in the GSS and five online samples:. *Journal of Quantitative Criminology* (2019): 1-26.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Yoo, James, and Richard C. Ready. "Preference heterogeneity for renewable energy technology." *Energy Economics* 42 (2014): 101-114.