April, 2009

Mode Effects and Other Potential Biases in Panel-based Internet Surveys: Final Report

By

Patricia A. Taylor, Ph.D., Professor of Sociology
Nanette M. Nelson, M.S., Assistant Research Scientist
Burke D. Grandjean, Ph.D., Professor of Statistics and Sociology
Bistra Anatchkova, Ph.D., Manager of Survey Research
David Aadland, Ph.D., Associate Professor of Economics

With the assistance of

Mark McNulty, Ph.D., former Senior Research Scientist
Brian Harnisch, Assistant Research Scientist
Michael Dorssom, M.A., Assistant Research Scientist

Wyoming Survey & Analysis Center
University of Wyoming
1000 E. University Avenue, Dept. 3925
Laramie, WY 82071
(307) 766-2189 • wysac@uwyo.edu • www.uwyo.edu/wysac

WYSAC Technical Report #SRC-905

Supported under Cooperative Agreement # USEPA 83359101

Project Officer

Alan Carlin, Ph.D.
National Center for Environmental Economics
Office of Policy, Economics, and Innovation
U.S. Environmental Protection Agency, Washington, DC 20460

Prepared for

National Center for Environmental Economics
Office of Policy, Economics, and Innovation
U.S. Environmental Protection Agency, Washington, DC 20460

Citation for this document: WYSAC (2009) *Mode Effects and Other Potential Biases in Panel-based Internet Surveys: Final Report,* by P.A. Taylor, N.M. Nelson, B.D. Grandjean, B. Anatchkova, and D. Aadland. (WYSAC Technical Report No. SRC-905). Laramie, WY: Wyoming Survey & Analysis Center, University of Wyoming.

Short reference: WYSAC (2009), *Mode Effects Final Report*.

Table of Contents

List of Tables

List of Figures

Mode Effects and Other Potential Biases in Panel-based Internet Surveys: Final Report

1. Executive Summary

Surveys by phone, mail, or personal interview have been commonly used to estimate the public's willingness to pay (WTP) for environmental quality and other non-market goods. However, traditional survey methods are becoming increasingly problematic because of changes in technology (e.g., cell phones, caller-id) and the proliferation of bulk-mailings, marketing research, and telemarketing activities. These societal trends are making Americans difficult to contact for surveys and, when contacted, reluctant to participate.

A relatively new method for eliciting WTP is the use of standing panels of respondents for surveys administered via the Internet. With non-probability (volunteer) panels, the methods of inferential statistics are unfounded, and therefore any attempt to generalize results to a definable population is suspect. However, Knowledge Networks (KN) recruits its panelists through probability-based sampling. KN web surveys therefore hold promise as a cost-effective alternative to other survey modes.

During the spring and summer of 2008 the Wyoming Survey & Analysis Center (WYSAC) at the University of Wyoming directed national surveys by telephone (1273 respondents) and mail (904), and also a web survey of KN panelists (1162). Each sample was administered a questionnaire about air quality in national parks. By design, the questionnaire was nearly identical for all three modes, as was the sampling frame.

The response rate was much lower for the web survey than by phone or by mail. Response was best in the mail survey, which also showed the greatest yield from additional survey efforts aimed at encouraging response. Exploratory analyses gave indications of differential non-response bias by mode, apparently due to mode-related variation in the mechanisms of self-selection as a survey participant. Phone and mail surveys may involve more self-selection of respondents interested in the topic of a particular survey, whereas a web panel may self-select for those with sedentary lifestyles.

Weighting and matching the respondents did not eliminate significant demographic, behavioral, and attitudinal differences across modes. For example, web respondents were less likely than either phone or mail respondents to have children in the home, to be members of any environmental organizations, to participate in various kinds of outdoor recreation, and to express high satisfaction with the National Park Service. Our econometric models therefore incorporated statistical controls for variables likely to be correlated with both mode of survey administration and WTP.

Results showed that using either a panel-based Internet survey or a mail survey produces a more conservative dollar value for WTP than using a phone survey. Communication with a live interviewer over the phone seems to yield over-statement of true WTP. Though face-to-face interviewing was not part of our research design, the apparent upward bias on WTP due to the effects of social desirability in a phone survey would also be expected in a face-to-face survey.

We found, further, that the variance in WTP left unexplained by our model was higher for the web panel than for either of the other two survey modes. There was a slight negative effect on WTP

from panel conditioning (as measured by duration of panel membership or number of web surveys completed), whereas survey fatigue had no consistent effect in the other two modes. Statistical interactions between mode of survey administration and other explanatory variables were of little substantive importance. In all three modes, the factors affecting WTP were similar and the signs of their effects were consistent with plausible theoretical expectations.

Hence, with appropriate controls, a WTP estimate derived from a KN web survey should be no less accurate than that obtained from a well-designed and well-executed mail or phone survey.

The cost of data collection proved to be highest by mail and lowest by phone. The web survey was in the middle on cost, but closer to the high end than the low. Any of these three modes would be much cheaper than face-to-face interviewing if the goal is to obtain a large representative sample.

Strictly speaking, a KN web panel survey does not solve the methodological problems associated with changes in telephone use, because KN's recruitment process is itself reliant on a fairly traditional telephone survey approach. For example, the low cumulative response rate in our web survey, with the attendant risk of non-response bias, results mainly from KN's low response rate during telephone recruitment of panelists. Hence, to keep pace with cultural and technological changes in telephony, KN will need to adopt the same kinds of tactics for recruitment that all telephone surveyors are now using to improve the representativeness of their samples. These tactics include dual-frame sampling (cell and landline), multi-mode initial contacts (mail, phone, and email), non-contingent incentives, and multi-mode follow-up with non-respondents. Such tactics can be expensive, and their use could well close the fairly narrow gap between the cost of a KN web panel survey and that of a thoroughly designed and implemented mail survey.

Meanwhile, advances in address-based sampling are improving the potential for representativeness of mail surveys. The techniques for maximizing response and minimizing non-response bias in a mail survey are also being refined through systematic research (e.g., Dillman 2007). Like a web survey, a mail survey is well-suited to the use of photos or other visual aids that may be especially helpful when asking about environmental quality. And unlike phone-recruited web panels, mail surveys are not much affected by cultural and technological changes in telephone use.

In sum, our findings demonstrate that mail surveys and probability-based web surveys both merit consideration as alternatives to phone or face-to-face interviewing in studies of willingness to pay for environmental quality.

2. Introduction

During 2008 the Wyoming Survey & Analysis Center (WYSAC) at the University of Wyoming directed a national survey by telephone, on the Internet, and by mail, with a total of more than 3300 respondents nation-wide. The goal of the study is to assist the U.S. Environmental Protection Agency (EPA) in comparing traditional modes of administering a questionnaire with new survey methods that use recruited panels of Internet respondents. Specifically, EPA wants to assess the relative merits of these different survey approaches for estimating the general public's willingness to pay for environmental quality.

The present study uses phone and mail samples drawn by Random Digit Dialing (RDD) for comparison with a sample from an Internet panel recruited by RDD. Each sample was administered a questionnaire developed by WYSAC that focused on the public's valuation of improved air quality in national parks. By design, the questionnaire is nearly identical for all three modes, and also allows comparisons on some items with a separate nation-wide telephone survey of over 4000 respondents that WYSAC conducted for the National Park Service. Controlling, insofar as possible, for mode differences in the demographic characteristics of respondents, we performed statistical tests for mode effects in reported behaviors and attitudes. We used econometric modeling to examine mode differences in willingness to pay to reduce ozone pollution in the parks. This report summarizes some relevant prior research, our methodology, and the main findings of the study. Several appendices present the methods and results in further detail.

2.1. Background

In its current Strategic Plan, EPA (2006) affirms its commitment to improving the nation's air quality (Goal 1: Clean Air), and establishes targets for the reduction of ground-level ozone as well as particulate matter. Furthermore, the Plan specifically calls for partnership with the NPS, in the U.S. Department of the Interior, to confront the problem of air pollution in national parks. As the Plan acknowledges, meeting clean-air targets will require that EPA rely not only on federal but also on state, tribal, and local programs. This presents challenges, given budget constraints at all levels. Accurate estimation of the public's willingness to pay for clean air in the national parks can contribute to mobilizing the necessary cooperation and the political will.

More generally, proper management of any environmental resource (air quality, water quality, etc.) requires an understanding of the value that society places on such resources. Because of the absence of formal markets for most environmental resources, their value must be estimated using elicited measures of willingness-to-pay (WTP). Surveys by phone, mail, or personal interview have been commonly used to estimate WTP. However, traditional survey methods are becoming increasingly problematic because of changes in technology (e.g., cell phones, call screening through caller-id) and the proliferation of bulk-mailings, marketing research, and telemarketing activities. These societal trends are making Americans difficult to contact for surveys and, when contacted, reluctant to participate.

A relatively new method for eliciting WTP is the use of standing panels of respondents for surveys administered via the Internet. However, key properties of the information gathered through panel-based Internet surveys are not yet fully understood. WYSAC's aim here is to examine differences between Internet panel surveys and studies conducted by two other modes of survey administration, telephone and mail. This is consistent with another objective of EPA's Strategic Plan: to enhance science and research. Our main focus in this study is on mode effects *per se* (differences attributable to the medium through which the respondent is questioned), but we also consider differences that may be due to dissimilarity in the populations sampled, differential non-response, or other factors.

2.2. Organization of this Report

The next section of the report reviews some relevant prior research on web surveys, mode differences, and the contingent valuation method for estimating willingness to pay. This is followed by a discussion of the methods used for the WYSAC study. Then the basic results are compared across modes, and an econometric analysis is presented. Finally, the relative costs of data collection by the different modes are considered. An appendix provides frequency distributions and matched cross-tabulations by mode for every survey variable, along with the full script used in the telephone interviewing. Three other appendices present the verbal responses, separately by mode, on open-ended questions. Two more appendices provide details on the web survey, including screenshots of the questionnaire and the field report prepared by the web panel vendor. A final appendix includes all correspondence sent to households in the mail sample, along with a copy of the mail questionnaire.

3. Previous Research

3.1. Web Surveys and Web Panels: Seeking Representative Samples

Surveys using the World-Wide Web (hereafter termed "web" or "Internet" without distinction) provide an inexpensive way to reach a large number of potential respondents (Dillman, 2007). A web survey is vastly cheaper than in-person interviewing, and potentially more cost-effective than a phone or mail survey.

Like a computer-aided telephone interview or a personal interview using a hand-held computer (and in contrast to a mail survey), a web questionnaire can be programmed for complex question structures involving randomization or logical skips. Like a mail questionnaire or a personal interview (and in contrast to a phone survey), a web survey can use photographs or other visual displays to help respondents think about what they are being asked to consider (see Dillman, 2007). This capability can be especially useful in studies of environmental issues, such as pollution.

However, web surveys also face a number of challenges in providing data of sufficient quality to be applicable to the population of interest. Except for small, special-purpose populations (e.g., the employees of a single company, all on the same email system), there is typically no comprehensive *sampling frame* from which to select a representative pool of Internet users. Therefore, web surveys are generally based to some degree on self-selection, even when researchers attempt to attract participants using different Internet service providers (Couper, 2000).

When the objective is to generalize results to the entire U.S. population, *coverage error* is a major concern; some households are simply unreachable in a web survey. Though a majority of U.S. adults now have access to the Internet, such access is still far from universal (NTIA, 2004). Access differs, not surprisingly, by age, education, income, and other demographic characteristics. Web surveyors often weight the responses, using estimates of access and/or propensity-to-respond along with demographic variables, in an effort to make them more representative of the population (see Kehoe & Pitkow, 1996). However, correct weighting is difficult because population parameters on some key demographics (to say nothing of attitudes and behaviors) are typically unknown (Couper, 2000). While incomplete coverage of the population might be partially addressed through weighting, no amount of weighting after the fact can adjust for lack of information about who is or is not online (see Andrews *et al.*, 2003).

Web-panel surveys offer an alternative to seeking a new pool of potential respondents every time a web survey is conducted. Two approaches can be distinguished: non-probability and probability web panels (see Berrens *et al.*, 2003; Couper, 2000). A non-probability panel consists of members who do not have a known probability of being selected. In other words, the initial target population is a purely self-selected "sample." Volunteers are recruited via appeals on popular websites and Internet portals, or by other means. At the time of registration for the panel, basic demographic data are collected to create a large database of potential respondents for future surveys. The panelists invited to participate in any given survey might then be determined at random, perhaps specified so as to include pre-determined proportions in various demographic subgroups (e.g., half male and half female). In essence, the individuals surveyed in a non-probability Internet panel comprise at best a quota sample from the register of initially self-selected panelists. They may give the appearance of representativeness, especially on the demographic variables used to form the subgroups, but without an initial random selection into the panel of potential invitees, formal methods of statistical inference (significance tests, confidence intervals) on such a "sample" are wholly unfounded.

By contrast, a probabilistic approach to panel design recruits panel members from a Random Digit Dialing (RDD) sample of household telephone numbers. Knowledge Networks (KN) is a leading vendor of such panels; see Appendix F for KN's description of the approach. Households in the RDD sample for which a reverse-lookup address can be obtained are initially contacted by U.S. mail through an introductory letter, and later by telephone. Phone numbers that do not yield a valid mailing address through reverse lookup are telephoned directly. In its early years, KN was credited with a "cooperation rate" of about 56% for this stage of the recruitment process (Huggins & Eyerman, 2001), although cooperation has almost certainly declined as it has for telephone surveys in general. Currently, KN reports an average "household recruitment rate" of about 22% and even lower overall "response rates" (see Appendix F, pages 76-77).

A major advantage of using RDD to recruit a web panel is that households without Internet access are covered in the sampling frame. Such households are eligible for recruitment and, if successfully recruited, are provided with free Internet access in exchange for joining the panel. (Panelists who already have Internet access are rewarded in other ways; see Appendix F, page 9.) Households do not even need computer access to participate, nor much computer literacy. If necessary, those who agree to become panelists are provided an Internet device (MSN-TV, more commonly known as web TV), web access, email account, and ongoing technical support (Berrens *et al.*, 2003). Ignoring for the moment some other potential pitfalls, the probability basis of a KN sample makes it suitable for inferential statistics, and for producing results that may reasonably be generalized beyond the population of (pre-recruitment) Internet users to the wider U.S. population (Couper, 2000).

For the present project, KN's probability-based approach to recruiting a web panel is used. A KN panel has potentially the same coverage and sampling frame as an RDD telephone survey (Berrens *et al.*, 2003). By using KN for our web panel, while also obtaining our phone sample through RDD and drawing our mail sample by RDD with reverse lookup, we come as close as possible to holding constant the sampling frame.

All three of our samples tend to miss the growing segment of the population with no landline phone. For our methodological purposes, however, the under-coverage due to cell-only and no-phone households is not of great concern. This source of under-coverage should be roughly constant across the three modes, because all three samples are drawn from the same RDD frame of landline telephone exchanges. Coverage differences by mode that remain can be further minimized by matching (i.e., by excluding households in any mode with no landline phone or no deliverable reverse-lookup address), by weighting to a common demographic profile, and by using statistical controls. In consequence, we can assess *mode effects* as such (response differences due solely to the medium by which a person receives a question), largely in isolation from *frame effects* (due to sampling from different populations).

3.2. Mode Effects and Other Potential Mode Differences

One well-established mode effect is the *social desirability bias* that results from the interaction between the respondent and a live interviewer, whether by phone or in person (Dillman, 2007). In answering a self-administered questionnaire, by mail or on the web, the respondent is less likely to distort responses toward the socially "right" answer. The absence of social interaction with a self-administered questionnaire may also generate more forthright responses on sensitive or private matters. And there may be differences in the cognitive processing that ensues when a question is heard aurally and answered orally, compared to the same question read from a computer screen and answered by mouse-click, or read from a paper questionnaire and answered with a pencil.

For our study, we followed the principles of unimodal questionnaire design (Dillman, 2007), using nearly identical wording and response choices across all three modes. For example, we did not use an explicit "don't know" category on any of the three modes; instead, in each questionnaire we embedded an initial instruction and subsequent reminders telling respondents that they could skip any question if they did not know the answer or preferred not to answer. We also strived for a similar visual appearance on the web and mail questionnaires.

Our goal in choosing unimodal design was to avoid confounding true mode effects with differences in question wording or appearance. Some slight variations were unavoidable, especially in transposing a read-and-respond, self-administered questionnaire into a hear-and-reply telephone interview. Differences are identified in Appendix A (where variant wording in the phone script is italicized), and are visible by comparing the web questionnaire (from the screenshots in Appendix E) to the mail questionnaire (reproduced in Appendix G).

Even within a single mode, differences in design and layout can affect respondents' answers (Dillman & Bowker, 2001; Dillman *et al.*, 1998). The appearance of a survey on the Internet is difficult to control because respondents may have their own user preferences, browser settings and hardware configurations. KN minimizes this issue by providing standardized equipment if needed, and by using a simplified interface intended to give the same look and feel, whether the survey is completed via KN's hardware or on the respondent's own computer. Together, these features of a KN web panel make for consistent presentation as well as reliable delivery of survey instruments.

Another concern with a web panel has been termed "panel conditioning." This refers to the possibility that, over time, respondents in a panel may change their behaviors, attitudes, or responses as a consequence of repeatedly being surveyed. Participants with longer tenure on the panel might respond differently compared to those who have a shorter tenure (Chatt and Dennis, 2003; Dennis, 2001). Invitations for KN surveys come to panelists at the rate of up to six per month, in the expectation that, on average, each panelist will complete about four surveys per month (see Appendix F). With the growth of marketing research and political polling, a more general effect (which we call *survey fatigue*) could apply to repeat respondents by other modes, as well. Frequent survey participants might become so accustomed to typical response choices (e.g., "strongly agree to strongly disagree") that they do not carefully consider all possible choices, the presence or absence of a neutral category, etc.

Our study allows us to explore such issues. We obtained information from KN on how long the individuals in our sample had been panelists. In addition, our questionnaire includes some meta-questions, asking respondents in all three modes how many phone, mail, and web surveys they completed in the past year.

3.3. Non-response Issues

Web panels (as well as mixed-mode surveys with a web component) have garnered increasing interest in part because of rapidly declining response rates for traditional approaches (see Swoboda *et al.*, 1997; Yun & Trumbo, 2000). A low rate of survey response increases the risk of *non-response bias*, which occurs when non-respondents differ in systematic ways from those who do respond to a survey. A low response rate is not problematic if the differences between respondents and non-respondents are small. However, when the magnitude of the difference is unknown, a high response rate limits the possible impact that the (unknown) degree of non-response bias can have on conclusions derived from the survey responses.

Differences between respondents and non-respondents are difficult to estimate because, by definition, those who do not respond to a survey do not provide any information about themselves. Demographic differences can sometimes be inferred by comparing the distribution of respondents on certain variables to known population values obtained from the U.S. Bureau of the Census or elsewhere. In turn, the effect of non-response bias on those comparison variables can be minimized by sample balancing (weighting the sample to approximate the known population distributions). Unfortunately, differences in other variables, such as attitudes, cannot typically be corrected in this way, unless they can be assumed highly correlated with the demographics. Population values on attitudinal questions are usually unknown.

A virtue of RDD sampling is that some information about non-respondents is available from the sampling frame, which permits an assessment of non-response (see Couper, 2000). For example, in a landline RDD sample, geographic information is known about every case (even those that decline to answer the survey) from the telephone number alone. The telephone "exchange" (area code plus first three digits of the phone number) of a landline phone maps with reasonably good reliability to a specific geographic area, about which certain aggregate characteristics can be determined from Census or other sources. Comparing respondents to non-respondents on these exchange-level, aggregate variables sheds light on the likely nature, degree, and direction of non-response bias.

Panel-based Internet surveys can often achieve impressively high *completion rates*. However, the completion rate alone can be a misleading indicator of the potential for non-response bias. Respondents in a KN web panel go through several stages before being sampled to participate in a particular survey: (1) initial RDD panel recruitment, (2) web device installation, (3) profile survey completion, and (4) post-profile panel retention. All of these stages, as well as actual survey participation, are susceptible to attrition in the potential respondent pool (see Appendix F, pages 74-77). Having remained in the panel through all of these stages, the KN panelists selected for a particular survey are quite likely to complete that survey in high proportion. This will yield a high completion rate.

In the typical phone or mail survey, on the other hand, non-response includes (indeed, is mainly) attrition in the recruitment stage, through non-contact of a selected household or the household's refusal to participate if contacted. Therefore, comparisons of "response rate" between a web panel survey and a phone or mail survey must distinguish carefully between overall response and post-recruitment completion.

It is generally inappropriate to assume that the reasons for failing to respond by one survey mode are the same as those for failing to respond by another mode. One topic may lend itself to telephone interviewing, while another topic yields a better response from letter contact or on the web. A mail survey (like a web panel) allows its participants to respond to the survey at times that are convenient for them. However, a questionnaire received by email (or mail) is easily put aside and forgotten, whereas a phone call (if answered) may be harder to ignore if there is a well-trained and personable human interviewer on the line. Telephone survey samples face their own specific sources of attrition, such as poor landline connections, interruptions of family activities, and failure to reach a willing but initially unavailable respondent on a call-back. Our study allows some comparisons in the outcomes of these distinct but parallel processes across the three modes.

3.4. Estimating Willingness to Pay for Environmental Quality

In all three modes, we used an approach known in the economics literature as contingent valuation (CV), in order to assess how much the public is willing to pay for environmental improvement. The CV method for estimating economic value for non-market goods is one of a broader category of valuation methods called *stated preference* approaches (for a review, see Adamowicz, 2004). Choice experiments and contingent ranking fall in the same general category.

These methods have the common feature that they are all based on surveys in which the public is directly questioned about willingness to pay (WTP) for certain hypothetical changes in access to natural resource use or environmental quality, or about choices between different "packages" of environmental quality and the price of each package (e.g., Herriges & Shogren, 1996). The contingent valuation method is the most common of these approaches in practice (see e.g., Bateman & Willis, 1999).

The CV method originated in the 1960s; it developed rapidly between the 70s and 80s; and it began to mature in the 1990s. We note that economists from the University of Wyoming, in part funded by EPA, have been influential in the development of CV methods (e.g., Crocker *et al.*, 1979) and in their continuing application to environmental issues (e.g., Aadland and Caplan, 2003). CV has been sanctioned for use in government decision-making and in the courts. By 2000, Carson *et al.* had identified over 1,600 CV-style studies, and its use continues to grow worldwide. For example, a recent study estimated the public's WTP for visibly cleaner air in national parks and wilderness areas to be $4.3 billion per year (Hill, 2000).

The CV method is straightforward, in theory. Since the absence of market prices for environmental goods is due to the absence of a market, CV asks people how they would behave *if* there were in fact a mechanism through which they could pay to purchase some quantity of environmental improvement. Ideally, a CV study starts with some focus groups, in which different hypothetical market scenarios and question formats are tested to refine the questionnaire. The next step is to conduct a survey, whether by phone, by mail, through group administration, with face-to-face interviews, or on the Internet. A final stage is to undertake a "bid curve" analysis, where WTP responses are statistically related, generally using some form of regression analysis, to social and economic variables thought likely to influence it.

In practice, there are several concerns that must be addressed in designing a CV questionnaire: (1) People should be given a plausible reason why they might pay for something that they currently do not see themselves as having to purchase. (2) A "bid vehicle" (the imaginary means by which respondents pay in the hypothetical market) should be credible. Respondents must think that the bid vehicle could reasonably be applied in practice. (3) Respondents should be given adequate, unbiased information on the good and its hypothetical market, to let them make an informed judgment. (4) The CV payment question should be asked in a way that minimizes incentives for respondents to behave strategically. For example, some respondents may be "free-riders" who try to under-bid in the expectation of enjoying the environmental good cheaply if others will bear the cost. (5) An effort should also be made to minimize "hypothetical bias" – the well-established empirical tendency to state WTP values that are greater than those revealed in real-market transactions (see Diamond & Hausman, 1994; Harrison & Rutström, 2006). For example, the questionnaire could include some "cheap talk," explicitly reminding respondents that they have other ways to spend their money and/or to obtain utility (e.g., Banzahf *et al.*, 2006). (6) "Protest bidders" should be identified. When respondents are asked how much they would pay, a fraction will give a zero response. For some people, this is because they do not value the good. For others, a zero bid might be because they are protesting about being asked the question in this way, or because the hypothetical market is not credible. Protest bidders are often separated out before analysis.

In addition to dealing with these practical concerns, the designer of a CV questionnaire must decide how to ask the WTP question. This can be done using an open-ended format ("What is the most you would be WTP?") or a dichotomous-choice format. In the latter format, people are asked to say whether they would pay a specific amount, known as the *bid price*. This bid price is then varied across people, which yields yes/no responses to different amounts. The dichotomous-choice format (also known as referendum or "take it or leave it") is cognitively less challenging than requiring the respondent to state a specific dollar amount. It has the further value that it mimics how real-life purchasing decisions are usually made (Cameron and James, 1987). In most actual market transactions, a good is offered at a certain price, and the consumer decides whether or not to buy it at that price.

The next section describes how these practical considerations in using the CV method were taken into account for the present study.

4. Designing the Questionnaire

4.1. The Environmental Problem: Ozone Pollution in National Parks

WYSAC's research team has a long-standing research interest in issues affecting national parks, as evidenced in past and current studies conducted on behalf of NPS (e.g., Taylor and Grandjean, 2009). That interest, plus EPA's role in protecting air quality in national parks, determined our choice of air pollution in the parks as the substantive focus for examining mode effects in panel-based Internet surveys.

Ordinarily, researchers designing a WTP survey could choose a single mode of administration, to allow optimizing the design of the questionnaire for that survey mode alone. For instance, a study of air pollution might use photographs of smoggy skies in a mail or web questionnaire, but could not readily do so in a telephone interview. However, our objective was to compare three survey modes (phone, web, and mail), and hence we had to devise a valuation scenario that does not require visual aids or other questionnaire elements that would not work equivalently in all three modes.

Given these constraints and our knowledge of the air quality issues facing national parks, we decided to survey how much the public would value reductions in ground-level ozone concentrations in the parks. Ground-level ozone is a form of air pollution that is not apparent to the eye. Yet it is one of the most widespread pollutants affecting vegetation and public health throughout the world.

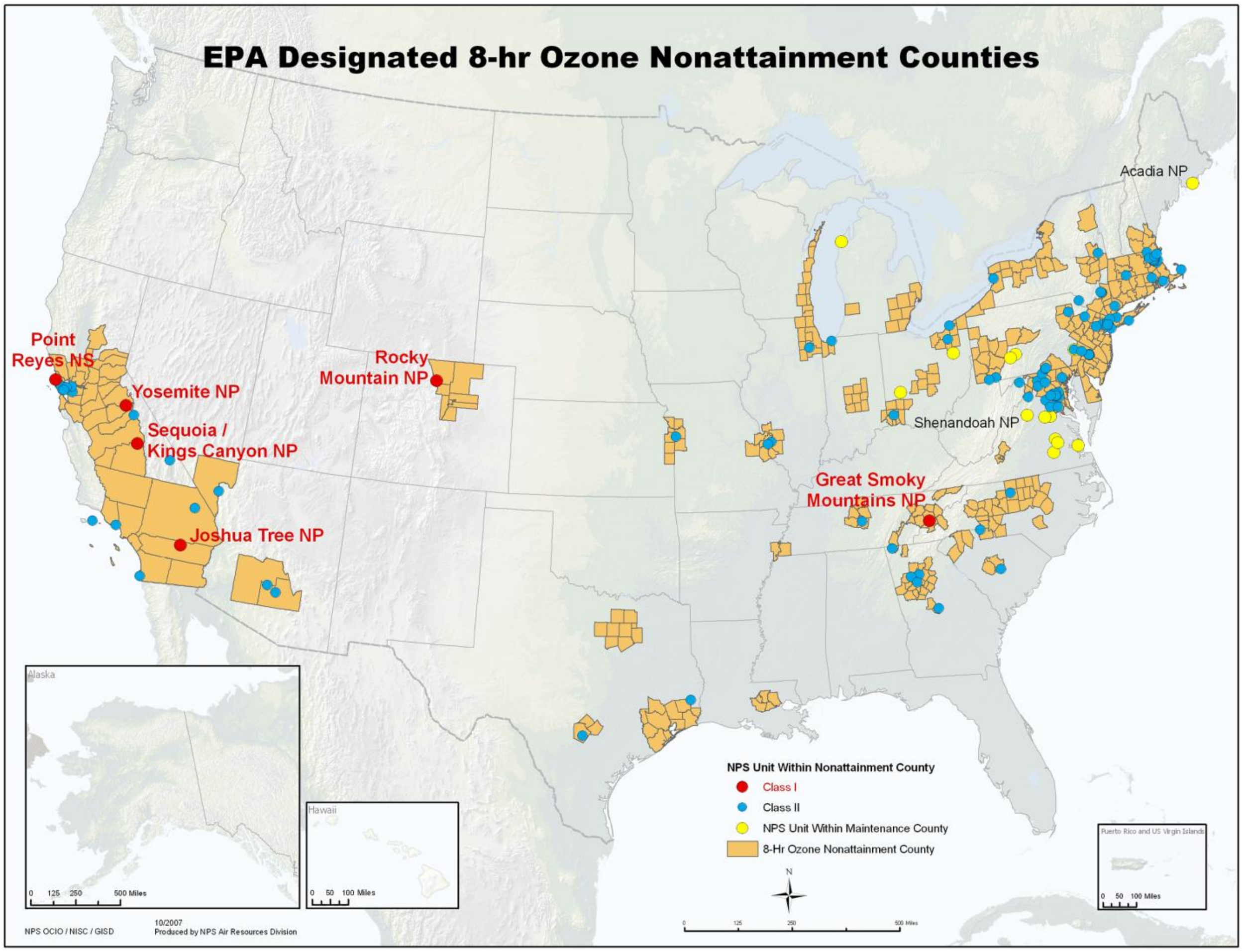
Produced by the reaction of nitrogen oxides and volatile organic compounds under sunlight, ozone pollution endangers human health in a variety of ways. It can cause chest pain, congestion, and lung irritation. It can also trigger episodes of bronchitis, emphysema, and asthma, and may permanently scar lung tissue. Indirect effects of ground level ozone on human well-being, through damage to vegetation, include slowed growth and reduced survival of tree seedlings, and increased susceptibility to pests and diseases for both forests and farm crops (EPA, n.d.a).

As mapped in Figure 4.1, there are national parks in the Southeast, the Northeast, and the Pacific Coast that experience with some frequency a level of ozone concentration in excess of standards set by EPA to protect human health. Indeed, as illustrated in Figure 4.2, concern for the health and safety of visitors and employees has led NPS to adopt an ozone advisory system in several parks where levels are likely to approach or exceed EPA ozone standards (NPS, n.d.b).

Therefore, ozone pollution at national parks held promise as the topic for our surveys. It is an environmental problem that respondents should find plausible and that can be described without the need for visual aids.

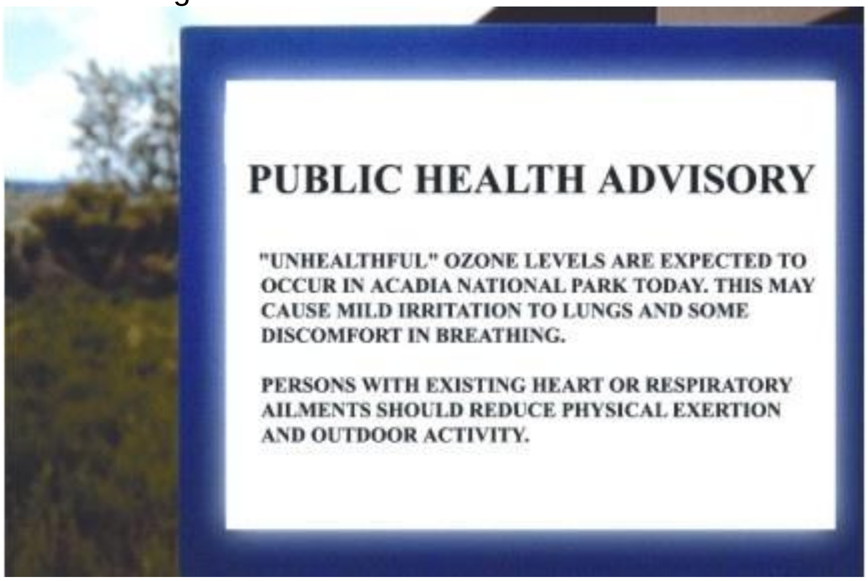
On the other hand, the word "ozone" carries a dual meaning. While ground-level ozone is a health hazard, ozone high in the atmosphere is necessary for human health as a protection from ultraviolet radiation (EPA, n.d.a). Lest respondents confuse ground-level ozone with their ideas about the "ozone hole" in the upper stratosphere, we decided against using the word "ozone" in our questionnaire. Instead we used the generic phrase "invisible air pollution," paired with descriptions of health effects from EPA's air quality guide for ozone (see section 4.3, below).

Figure 4.1. NPS Units within Counties that Exceed EPA Standards for Ozone



Source: NPS, n.d.a.

Figure 4.2. Health Warning for Ozone Posted at Acadia National Park



Source: NPS, n.d.b.

4.2. Structure of the Questionnaire

The questionnaire can be considered in three parts. (The phone, web, and mail questionnaires are presented in full in Appendices A, E, and G, respectively.) The first part helped to establish a context for the WTP scenario, by asking about a respondent's experience with units of the National Park System (national parks, national historic and cultural sites, and national monuments). Respondents were also asked about their outdoor activities and their opinions on several policy issues facing the larger national parks. This first part of the questionnaire was designed to be similar to (although much shorter than) the 2008 Comprehensive Survey of the American Public, a year-long, nation-wide RDD telephone survey that WYSAC completed in the spring of 2009 for the National Park Service. Most questions in this part of the mode-test questionnaire have identical counterparts in the NPS survey, providing opportunities for comparisons in future research.

The second part of the questionnaire began by presenting information on air pollution in national parks, and then described a hypothetical program to convert park vehicles to non-polluting electric or solar power. Using (without attribution) the EPA ozone standards that are discussed in the next section of this report, the questionnaire described the potential health effects associated with varying levels of "invisible air pollution." Then the valuation question asked respondents whether or not they would be willing to pay a specified additional entrance fee to fund the hypothetical program for reducing air pollution in some (unnamed) national park they were visiting.

The valuation question was immediately followed by a request for any information that would help to explain the respondent's answer. In addition to allowing qualitative analysis of their perceived motivations (e.g., health problems in the family), the respondents' open-ended explanations can be used to identify protest bids. Another WTP question was asked next, specifying a greater quantity of the environmental good (i.e., a greater improvement in air quality). The answers to the second valuation question provide the basis for a "scope test" (see Arrow, *et* al., 1993; Smith & Osborne, 1996). Respondents who are responding rationally to the valuation exercise should be willing to pay more (or at least the same) for the larger benefit received. If not, then the credibility of some aspect of the WTP scenario is suspect. The elicitation format for the second WTP question was open-ended.

The third part of the questionnaire consisted of demographic questions to measure factors that may affect an individual's WTP. The survey ended with some meta-questions about the respondent's survey behaviors, to obtain information relevant to methodological issues such as survey conditioning.

4.3. The Valuation Scenario

EPA has developed the Air Quality Index, or AQI, for ozone (and other pollutants) as a way of easily communicating to the public the health effects of ozone levels in a community. The AQI is a tool that state and local agencies use to issue public reports of actual levels of ground-level ozone. It is thus a familiar indicator to many people who live in areas with chronically poor air quality.

Table 4.1 presents the established air quality categories for ground-level ozone, the corresponding numerical ranges for the AQI and for ozone concentration, and EPA's verbal statements of the associated health advisories. In March 2008, after we had developed our survey instrument but before we had fielded the questionnaire by any mode, EPA strengthened the primary National Ambient Air Quality Standard for 8-hour ozone exposure, reducing the allowable amount from 0.084 to 0.075 part per million (ppm) (EPA, 2008; EPA, n.d.b). The AQI for ground-level ozone was subsequently updated and that is what we report in Table 4.1 (EPA, n.d.c). Since our questionnaire did not associate the verbal health advisories with specific numerical values of either the AQI or ozone concentration, the change by EPA had no adverse effect on our study. If anything, the change confirmed the factual accuracy of the questionnaire's description of the potential for health problems from ground-level ozone in national parks.

After reviewing current and historical ozone levels in the national parks we decided to use three levels of air pollution for our valuation scenario. Our descriptions of high, medium and low levels of "invisible air pollution" in the valuation scenario relate directly to EPA's air quality levels of unhealthy, unhealthy for sensitive groups, and good-to-moderate, respectively. The first (dichotomous) WTP question asked about a reduction in pollution from medium to low. The second (open-ended) question asked about a reduction from high to low, for use in our scope test. Table 4.2 provides a matching of our questionnaire descriptions to EPA's AQI and air quality categories.

Table 4.1. EPA Air Quality Guide for Ozone

Air Quality	Air Quality Index	Ozone Level (ppm)	Health Advisory
Good	0-50	0.000 – 0.059	No health impacts are expected when air quality is in this range.
Moderate	51-100	0.060 – 0.075	Unusually sensitive people should consider limiting prolong outdoor exertion
Unhealthy for Sensitive Groups	101-150	0.076 – 0.095	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
Unhealthy	151-200	0.096 – 0.115	Active children and adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else, especially children, should limit prolonged outdoor exertion.
Very Unhealthy (alert)	201-300	0.116 – 0.374	Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.

Source: AIRNow (n.d.); EPA (n.d.c)

Table 4.2. EPA Air Quality Categories vs. Levels of Air Pollution in the Questionnaire

EPA		Questionnaire*	
Air Quality	**AQI**	**Pollution Level**	**Extent of Health Concerns**
Good	0-50	LOW	“When invisible air pollution is LOW, it will not cause these health concerns. There will be no reason for anyone to limit outdoor activities.”
Moderate	51-100		
Unhealthy for Sensitive Groups	101-150	MEDIUM	“When invisible air pollution is MEDIUM, it will cause health concerns for some people. Active children and adults, and also inactive people with breathing problems like asthma, should limit their outdoor activities.”
Unhealthy	151-200	HIGH	“When invisible air pollution is HIGH, it will cause health concerns for everyone. All children and adults should limit or even avoid outdoor activities.”

Prefatory wording: “Please consider invisible air pollution at three levels: LOW, MEDIUM, or HIGH. Depending on the level, it can cause health concerns like lung irritation, painful breathing, sore throat, coughing, and shortness of breath. When invisible air pollution is LOW …”

4.4. Focus Groups and Other Feedback

WYSAC's research team drafted the WTP section of the questionnaire with the assistance of Jason Shogren, Distinguished Professor Economics at the University of Wyoming. When a completed draft was in hand, WYSAC conducted two focus-group sessions intended mainly to clarify the valuation scenario and to improve the credibility of the bid vehicle. The focus groups also provided helpful suggestions for rewording or reordering certain questions on the survey instrument. Changes were made to the survey instrument after each focus group that reflected the feedback received.

At the beginning of each meeting participants completed the then-current draft of the mail questionnaire, which served as a focal point for the ensuing discussion. Two of WYSAC's lead researchers on this project (Nelson and Taylor) served as the facilitators for both discussions. The focus groups, each about two hours in duration, were held in Scotts Bluff, Nebraska (with 15 participants, in a private room of a local restaurant) and in Cheyenne, Wyoming (with 13 participants, in a meeting room of a local hotel) on the evenings of November 14 and December 6, 2007, respectively. Potential participants were recruited through advertisements in local newspapers serving their respective communities. The final list of participants was selected for demographic diversity based on a short set of screening questions asked of those who responded to the ad by phoning WYSAC's toll-free number. Light refreshments were served at both sessions, and participants were compensated $50 each.

The first part of the questionnaire, on recreational activities and opinions about national parks, also benefitted from two other focus groups that WYSAC had convened to obtain input on the NPS Comprehensive Survey of the American Public. The first of these was held with a group of nine African-American participants on February 20, 2007, at the Blair Caldwell African American Research Library in Denver; the second was held with a group of nine Hispanic Americans on October 24, 2007, at the Jesus Rodarte Community Center in Greeley, Colorado.

More generally, that NPS questionnaire was developed under the guidance of James Gramann, Chief Social Scientist for the National Park Service. WYSAC sought written input on a draft of the NPS survey from a number of stakeholder organizations, obtaining responses from the American Recreation Coalition, America Outdoors, the National Park Hospitality Association, and the National Parks Conservation Association. We note, too, that the 2008 NPS questionnaire was modeled on an instrument designed and fielded for NPS in 2000 by the Social Research Laboratory at Northern Arizona University. All of these sources helped in crafting the 2008 NPS survey, which in turn determined the final wording of most of the items in the first part of the mode-test questionnaire as well as several of the demographic questions.

4.5. Pretests

WYSAC completed two nation-wide pretests of the WTP survey instrument: one by phone in January, 2008 (n=80 completed interviews) and one on a KN web sample in March (n=106 completions; see Appendix F). In addition to checking for programming flaws and any last-minute wording changes that might be needed, the major purpose of these two pretests was to finalize the "bid vector," the range of prices to be used in the valuation scenario. Pretests help to ensure that the bid vector covers an appropriate range of values for the environmental good under consideration.

After trying bids from $2 to $40 in the pretests, the final bid vector was set at $2, $5, $10, $15, and $25. The bid amount was randomly varied across respondents. In the final survey, unequal probabilities were used in randomizing the bids, so as to produce the following distribution of bid amounts within each mode: 30% of those surveyed got the middle bid of $10; there were 20% each at $5 and at $15; and 15% got the extremes of $2 and $25.

This non-uniform distribution for the bid vector was selected so as to put more respondents near the middle of the bid range, and thereby to improve the statistical precision of the WTP estimates. If most people in a sample say Yes to a particular bid amount, or most say No, the sample does not need to be very large to estimate with high confidence the proportion saying Yes in the population. A larger sample is needed if the Yes/No split is closer to 50/50. Based on the pretests, we expected mostly Yes at the $2 bid, mostly No at the $25 bid, and about a 50/50 split at the $10 bid. The non-uniform distribution of bid probabilities reflects these expectations (which were in fact confirmed in the full survey results).

For the phone and web questionnaires, the unequally randomized bid vector was written into the programming code. For the mail survey, five separate versions of the questionnaire were printed and mailed. Cases in the RDD mail sample were randomly assigned to receive one of the five versions, with the different bid amounts ($2/5/10/15/25) being printed in unequal percentages (15/20/30/20/15%). The mail returns were tracked using a bar code on the questionnaire, to make certain that re-mailings to non-responders would enclose the same version of the questionnaire originally assigned at random to that case.

Items in the questionnaire with counterparts in the NPS survey benefitted from two additional nation-wide pretests. In the first of these, cognitive interviews were conducted by telephone on a small national sample in November, 2007. Three specially trained WYSAC interviewers asked the 28 respondents a subset of questions from the draft NPS survey, and then asked them about their interpretations of those questions. When the NPS questionnaire was in nearly final form, WYSAC conducted a national telephone pretest with 81 respondents in March, 2008, to check for programming errors and the like. The few minor changes in question wording that resulted were made not only in the NPS survey, but also in the corresponding items of the phone, web, and mail mode-test questionnaires.

5. Administering the Three Surveys

5.1. Sampling

The potential universe of contacts in each mode consisted initially of all landline telephone numbers in the United States with an area code, three-digit prefix (the exchange), and working 100-bank (the next two digits).

The latest available estimates indicate that this conventional sampling frame for Random Digit Dialing (RDD) covered about 85 percent of all U.S. households in the spring of 2008. The list-assisted RDD frame covers both listed and non-listed telephone numbers (though only 100-banks with at least two listed residential phone numbers were included). Incomplete coverage is mainly attributable to the large and growing segment of the population with a cell phone but no residential landline. Fortunately for our purposes, under-coverage due to cell-only households should be roughly constant across the three modes, because all three samples are drawn from the same RDD frame of landline telephone exchanges. Hence, despite under-coverage, the three samples can be compared to assess mode effects as such, largely in isolation from frame effects.

Because households for which no address could be obtained by reverse lookup of the RDD phone number must (necessarily) be omitted entirely from the mail survey, we anticipated limiting key comparisons across mode to cases with valid addresses. As a cost-effective way of increasing the usable sample size for such matched comparisons, we under-sampled by a factor of one-half those households in the phone and web surveys with no reverse-lookup address. When analyzing only the phone or web sample (so that matching to the mail survey is not at issue), the no-address respondents can be weighted by a factor of 2 to compensate for the intentional under-sampling.

Details of KN's sampling and recruitment for the RDD-based web panel are described in Appendix F. For the phone and mail surveys, WYSAC purchased a single RDD sample of 40,000 landline phone numbers from a national vendor (Marketing Systems Group). The vendor screened that initial sample to eliminate, insofar as possible, disconnected numbers, businesses, duplicates, and other known ineligibles. By coordinating with KN, the sample vendor also purged our phone/mail sample of any phone numbers that appeared in our web sample. After all screening, the sample provided 24,041 potentially eligible phone numbers, for about 60% of which the vendor was able to provide reverse-lookup mailing addresses.

WYSAC randomly split the 24,041 pre-screened phone numbers into two unequal pools: about two-thirds for use in the phone survey, and one-third for use in the mail survey. That unequal split was based on the relative response rates we expected from our prior experience with national phone and mail surveys. In the phone split, a random one-half of the numbers with no reverse-lookup address were held out of the calling (i.e., we intentionally under-sampled this segment). In the mail split, all such cases were held out (because a mailing address was essential for the mail survey). Ultimately, we made 13,475 numbers available for calling in the phone survey, and used 4185 cases with reverse-lookup addresses for the mail survey. No data collection efforts were pursued on the remaining 6381 phone numbers (the unused half of the no-address numbers in the phone split, plus all of the no-address numbers in the mail split).

Along with the RDD sampling frame, the method for within-household sampling of one adult respondent was also held constant. The logistics of a mail survey led to the choice of the "last birthday" method for within-household selection in all three modes. In the web survey, KN could have easily selected an adult panelist in each household by strictly random means. The phone questionnaire could also have been programmed to randomly identify a selected respondent (e.g., by asking how many adults live in the household and then using the software's randomizer to select, say, the "second oldest" adult to be interviewed). For the mail survey, however, the within-household selection mechanism had to be simple enough that whichever household member opened our mailings could be expected to understand the selection instructions and to apply them with reasonable accuracy.

Previous research has shown that the birthday method gives an acceptable approximation to pure random selection, as long as children are not part of the target population (see Grandjean *et al.*, 2005). All three modes therefore used within-household selection of the adult with the most recent birthday. In the mail survey, instructions in the cover letter and also on the questionnaire itself asked that it be completed by "the adult in your household who had the most recent birthday." The phone interviewers used the same phrasing, and sought to schedule a callback appointment if the selected respondent was not available at the initial contact. KN used birth date information from its database on panel members to determine which adult panelist in the household should receive an email invitation to participate in the survey. Thus, for the web survey only, the within-household selection mechanism was transparent to the respondent.

5.2. Field Periods, Survey Effort, and Undeliverable Addresses

5.2.1. Phone survey

Calling for the phone survey began on April 30, 2008, and ended on July 31.

The questionnaire was programmed for Computer-Aided Telephone Interviewing (CATI) by WYSAC's Survey Research Center personnel, and the interviewing was conducted using the Wyoming Call Center's fully equipped CATI facility. (See Appendix A for the complete interview script.) Telephone numbers were typically dialed automatically by the CATI software, but with a live interviewer on the line from the outset of each call so that potential respondents were unaware of the automated calling. However, numbers flagged (by the sample vendor, or through the early calling) as likely cell phones were then dialed manually, as required by federal law.

Households in the RDD landline sample that were reached on a cell phone were not presumed to be ineligible, because a valid household landline number might be temporarily call-forwarded, or even permanently ported, to a non-landline phone. As discussed in Section 5.4, below, whenever the analysis was limited to matched subsamples we excluded the few phone respondents who reported, on a question near the end of the survey, that the household had no landline phone.

The phone sample was randomly subdivided into waves, and each wave was thoroughly worked in the calling (initially, 12 or more callbacks per un-reached number) before deciding whether the next wave was needed. This step reduced the potential over-representation of households easiest to reach by phone. Nevertheless, when the initial target of 1000 completed phone interviews was reached in early June, a preliminary analysis of the data suggested the possibility of some early-responder bias. (For example, females and senior citizens were over-represented.) Therefore, the field period was re-opened for another six weeks; the callback criterion was increased to at least 15; more than 15 callbacks were attempted on numbers for which the call history looked promising (any contact identifying a working household number); and refusal-conversion efforts were completed on all outstanding "soft" refusals.

By the end of July, some completions had required more than 20 callbacks to achieve, and a few phone numbers had been attempted unsuccessfully more than 50 times. Among completed interviews, the mean number of attempts was 4.1, and the 75th percentile was 5 calls. Among eventual non-respondents, the mean was 9.0 attempts and the 75th percentile was 15. At least one refusal conversion attempt was made on every terminated call (hang-up) or other soft refusal, but not for irate refusals or requests to be added to the Do Not Call list.

Households where the interviewers encountered a language barrier were called back by a refusal converter and then, if confirmed as non-English speaking, were coded as ineligible for the phone survey. Like the web and mail questionnaires, the phone survey was only provided in English. Because our primary interest was to isolate mode effects, we decided to hold language of survey administration constant and to accept some loss in national representativeness as a result.

In August, 2008, WYSAC sent thank-you postcards to the reverse-lookup address (if available) of each completed telephone interview. Those returned by the U.S. Postal Service as undeliverable were identified as such in the data file so that these cases could be excluded from analyses based on matched comparisons with the mail survey.

5.2.2. Web survey

In the web survey, KN's email invitation to participate went to the selected sample of panelists on April 22, 2008. The field period, planned for 21 days, was ultimately extended through May 27.

During the first three days in the field, nearly half the contractual target of 1000 web completions had already been achieved. (For the text of all communications between KN and the panelists in our survey, see Appendix E.) On the fourth day, KN sent its standard email reminder to non-respondents. A week later, the target for completions was surpassed.

To minimize any early-responder bias, and to keep the level of survey effort similar across modes, WYSAC then amended our agreement with KN so as to keep the survey open and pursue additional completions. On May 5, KN sent all non-respondents a longer, customized email reminder, authored and signed by a WYSAC researcher. On May 14, the household of each remaining non-respondent received a phone call from KN with an automated reminder message. When the survey closed on May 27, a total of 1162 web completions had been received.

In August, 2008, KN sent thank-you postcards to the reverse-lookup address (if available) of each web respondent. The objective, as in the similar thank-you mailings after the phone survey, was to identify undeliverable addresses. Because of an error by KN staff in tracking some of the initial return notices from the Postal Service, KN mailed a second set of postcards in October. Postcards returned to KN as undeliverable, from either of these mailings, were identified as such in the data file. To ensure comparability with deliverability as measured in the phone and mail surveys, these web thank-you mailings did not rely on the addresses on file in KN's panelist database. Instead, they used addresses obtained by KN in August from WYSAC's sample vendor (Marketing Systems Group) by reverse lookup of the respondents' phone numbers.

5.2.3. Mail survey

The field period for the mail survey ran from April 14 through August 31, 2008.

Correspondence for the mail survey began with a pre-survey contact letter to all reverse-lookup addresses cases in the mail sample (see Appendix G). For efficiency in the associated clerical work, these letters went out in batches over several days, with the first batch being mailed on April 14, 2008. A few days after each batch of contact letters, the corresponding batch of questionnaires (with cover letters) was mailed. The cover letter included a P.S. requesting a very brief reply, using the pre-paid envelope enclosed, to indicate if the mailing had reached a business or government office, or if the household was declining to participate in the survey.

About a week after sending the questionnaires, reminder postcards went out (in batches) to non-respondents, excluding cases that had already produced either (a) a reply declining to participate or identifying the address as business or government, or (b) a return from the Postal Service identifying the address as undeliverable. However, if the return from the Postal Service could be traced to a correctable problem in the address, such as a typographical error or a missing zip code, we fixed that problem in our address database and sent subsequent mailings to the corrected address. By the end of the entire field period, the mailing efforts had identified 112 explicit refusals, 6 ineligible addresses (business/government), and 563 undeliverable addresses.

Two to three weeks after the reminder postcard, a second copy of the questionnaire (with a somewhat different cover letter) was sent to each remaining non-respondent, except for the exclusions denoted as (a) and (b), just above. The last batch of these replacement questionnaires went out on May 16, 2008.

By the end of June, 749 mail questionnaires had been received. Allowing for the restriction to deliverable addresses in this mode, about 750 mail completions was roughly in line with the original target of 1000 completions (with or without address) in the other two modes. To keep the level of survey effort similar across modes, WYSAC decided to keep the survey open and to initiate reminder phone calls. During July 10-31, 2008, WYSAC called the RDD phone numbers associated with all non-responding addresses in the mail sample, except for the business/government cases and those that had explicitly declined to participate. The field period for accepting returned questionnaires was extended through August 31, to allow time for any returns that might be generated by the reminder calls.

If the phone call reached an answering machine or voicemail, a short reminder message was left, including WYSAC's toll-free phone number for requesting a new copy of the questionnaire. If the call reached a person, our interviewer provided a similar reminder and offered to send another copy of the questionnaire. In that case, the replacement questionnaire was sent within a day or two by Priority Mail, in a red-white-and-blue, 9 by 12 inch cardboard mailer. The last of these was mailed on August 1, 2008.

Once a household had received either kind of phone message (machine, or live), it was not called again. No refusal-conversion attempts were made in the reminder calling for the mail survey. The remaining mail non-respondent households were called back up to 5 times each. Ultimately, this additional survey effort yielded 146 returns of the third copy of the questionnaire (sent out by Priority Mail), plus 3 responses from households where only a machine message was left and 6 from households where a live message was left but the offer of a replacement questionnaire was declined. The total number of mail questionnaires received was therefore 904.

5.2.4. Comparing survey effort across modes

To compare the level of survey effort across modes, Table 5.1 imposes a common metric: a rough ordinal scale with four gradations. In the phone survey, "initial" effort (a single phone call) was sufficient to generate almost a third of the eventual completion total. In the mail survey, initial effort could not possibly have generated completions, because the pre-survey contact letter did not include a copy of the questionnaire. Nevertheless, such pre-notification is a well-established and relatively inexpensive way of increasing response rates (Dillman, 2007).

In the web survey, as well, the initial effort could not generate any completions on our particular survey. In contrast to the mail survey, however, that part of the web survey effort is both indispensible and very expensive. It involves the essential steps of recruiting panel members by RDD, installing web equipment, obtaining profile information for the database, and retaining panel members once they have enrolled. All of this work is undertaken by KN on a continuing basis, but is not part of the measurable survey effort for our particular web survey.

With the "minimal" additional effort of a single email invitation to participate in our web survey, KN generated more than a third of the eventual completions by that mode. Minimal effort in the mail survey (simply mailing the questionnaire, followed by a reminder postcard) generated almost two-thirds of the eventual completions. With minimal effort in the phone survey (no more than 5 calls per phone number), the original target of 1000 phone completions was exceeded – more than four-fifths of the eventual total for that mode.

Table 5.1. Number of Completions, by Survey Effort and Mode

Survey Effort	Phone	n	Web	n	Mail	n
Initial	1st call	392 (392)	KN panel recruitment, Setup, Profiling	--	Contact letter	-- (7)
Minimal	1 – 4 callbacks	632 (1,023)	KN email invitation	425 (425)	Cover letter + survey, Reminder postcard	586 (593)
Ordinary	5 – 9 callbacks	111 (1,134)	KN email reminder	630 (1,055)	2nd cover letter + survey	156 (749)
Concerted	10+ callbacks	138 (1,273)	WYSAC email, KN automated phone call	107 (1,162)	Phone reminder, Priority mailing	155 (904)

Totals in parentheses are cumulative; grand total across all three modes = (3,339). In the mail survey, the 7 respondents shown in the first row obliterated the identification code when returning the questionnaire, so the level of survey effort to generate those completions is unknown.

With the "ordinary" (and cheap) effort of a short email reminder, KN brought the web completions well past the original target, adding more than half of the eventual web total. A second mailing of the questionnaire brought the mail total from two-thirds to five-sixths, while a few more callbacks on the phone survey added less than 9% of total completions by that mode.

What we have termed "concerted" effort generated about the same number of additional completions for each mode, but made a bigger proportional difference in the mail survey. Mail completions increased by more than 20% as a result of the telephone reminder calls. The phone total grew more than 12% due to the tenth and subsequent callback attempts. The web total increased about 10% after the WYSAC-authored email and KN's automated phone calls.

These results indicate sharply diminishing returns for the web-panel survey, beyond ordinary survey effort (a simple email reminder). For the phone survey, returns diminished even sooner, past 4 callbacks or so. For the mail survey, more concerted effort– going beyond a second mailing of the questionnaire to include telephone reminder calls – continued to yield good returns.

We note a possible avenue for additional survey effort that was not pursued in this research because it could not have been used equivalently in all three modes. A token incentive payment (perhaps $2 or $5) would likely have increased response. In the mail survey, such a payment would ordinarily have been included with the first mailing of the questionnaire, because a non-contingent incentive is best for improving the response rate (Dillman, 2007). The web survey could also have included a non-contingent incentive, in a mailing to the addresses that KN has on file for those in our web sample. However, the logistics of an incentive payment for the phone survey would have been highly problematic.

In our phone sample, a pre-survey notification letter (with or without an incentive) could only have been sent to cases with reverse-lookup addresses. That would have introduced a likely source of non-response bias, by increasing the response rate only among the address-obtainable households. Alternatively, the incentive could have been offered to everyone reached by phone, during the initial phone contact. But under that approach, only those who agreed to stay on the phone and provide their mailing address could be sent the payment. As a result the confidentiality of the survey would be compromised for the phone mode (only), and the phone incentive would have become partly a contingent reward for providing an address. To ensure the comparability of our phone data with the web and mail results, we decided not to use a monetary incentive for any of the modes.

5.3. Rates of Survey Participation

Survey participation rates can be defined in various ways. The basic idea is to form a ratio of the number of respondents divided by the number of potential respondents. The number of respondents is usually unambiguous, but the number of potential respondents may not be so clear-cut. A narrow definition of potential respondents puts a smaller number in the denominator of the ratio. That yields a larger numerical value for the participation rate than when potential respondents are defined more broadly.

One popular measure of survey participation is the (post-selection) "completion rate": the number of finished questionnaires, as a proportion of all cases in which an eligible adult is known to have been selected as the prospective respondent. This rate tends to be a high numerical value, because much of the non-response in a survey occurs before selection of a specific respondent (indeed, even before eligibility is known with certainty).

A somewhat more stringent measure is the (pre-selection) "cooperation rate": the number of finished questionnaires divided by the number of cases known to be eligible. By including in the denominator all cases known to be eligible (whether or not a particular respondent was selected), this rate yields somewhat lower values than the post-selection completion rate.

Response rates tend to be lower still, but even a so-called "response rate" can be calculated in several ways. A formula that is useful for the present comparative purposes is what the American Association for Public Opinion Research (AAPOR) identifies as RR3. The formula for RR3 includes in the denominator not only the cases known to be eligible, but also a fraction of the cases with unknown eligibility. RR3 assumes that the fraction of eligible cases, among those where eligibility could not be determined, is the same as the proportion eligible among cases that were definitively identified as either eligible or ineligible.

To be eligible for our phone survey, the RDD number must have reached a private U.S. household, not a business or government office and not a disconnected number or a dedicated data line. In addition, the number must at some point have reached an adult member able to converse in English. Cases of unknown eligibility included phone numbers that were continuously busy, never answered, or otherwise unidentifiable as residential. With multiple callbacks spread over several weeks, most such non-contacts were probably not working phones. Therefore, the RR3 formula is likely to give a conservative value for the "response rate."

In our mail survey, the reminder calling allowed us to apply the same eligibility criteria as in the phone survey (but with the additional requirement of a deliverable mailing address). Prior to the reminder calls, the mail sample had more than 2700 cases of unknown eligibility (no return of any sort from the mailings). The reminder calling identified most of these as either eligible or ineligible, leaving only 602 of unknown eligibility. Those identified as eligible included new completions and firm refusals, plus any other contacts that sufficed to verify the case as a private household. The ineligible category included non-residential and non-working phone numbers, plus non-English speaking households.

The same eligibility criteria also apply to the KN panel, but they are implemented by KN at the telephone recruitment stage. Within our sample of web panelists, everyone had previously been identified by KN as eligible, so the number ineligible at the surveying stage was zero. But the proportion eligible at the recruitment stage can still be taken into account. KN calculates RR3 at the recruitment stage, then the post-recruitment profiling rate, the study-specific completion rate, and the panel retention rate. These can be combined in various ways to estimate the overall response rate (see Appendix F, pages 75-77, as well as Section 6.4, below).

What KN calls the Cumulative RR1 is the product of the recruitment, profiling, and completion rates (Appendix F, page 77). KN's Cumulative RR2 is the product of their RR1 multiplied by the retention rate. For our purposes, the Cumulative RR2 seems most comparable to the AAPOR RR3 that we use for our phone and mail surveys.

The definition of refusal to participate was similar across the three modes, but with some necessary differences. In the phone and mail surveys, pre-selection refusals constituted by far the majority of all refusals. Phone survey calls (and reminder calls for the mail survey) that never progressed through the full set of introductory items (the call was answered, but was abruptly terminated by hanging up or declining to proceed) were temporarily coded as pre-selection refusals. If no subsequent refusal-conversion attempt was undertaken (because the initial refusal was irate, because it included a request to be added to the Do Not Call list, or because it came in the mail reminder calling), then the temporary refusal code automatically became the final code. Phone numbers that were called again for refusal-conversion attempts received a final disposition code as pre-selection refusals if those callbacks were ultimately unsuccessful.

In the web survey, the number of pre-selection refusals is unknown. All such refusals would have occurred during KN's panel recruitment, profiling, and retention efforts, not during our particular web survey. All non-respondents to the web survey were therefore treated as post-selection refusals.

For our phone survey, post-selection refusals occurred only in households where someone stayed on the phone long enough to confirm the household's eligibility to participate, and then reached the item where the interviewer asked if the person on the line was the adult in the household with the most recent birthday. A terminated call (a break-off) at any point thereafter and before the end of the second WTP question, or a failure to speak with the selected birthday adult on that call and all subsequent callbacks, was coded as a post-selection refusal on the phone survey.

In the mail survey, post-selection refusals included any case that returned a written response (to the P.S. in the cover letter) indicating that the household declined to participate. In getting to that P.S. at the bottom of the page, the household member would presumably have been exposed to the sentence in the body of the letter requesting that the survey be completed by the adult in the household with the most recent birthday. In that sense, the 112 written refusals received were all post-selection. In the reminder calling, cases were coded as post-selection refusals if the person on the line said that the birthday adult would complete and return the mail questionnaire, but no completed survey was ever received from that household. Among the returned mail questionnaires, there were no mid-questionnaire break-offs.

Using these definitions, Table 5.2 compares participation rates across the three survey modes. Not surprisingly, the post-selection completion rates yield high numerical values: around 90% for the phone survey, and over 75% by mail and on the web.

Table 5.2. Participation Rates, by Mode

	Phone	**n**	**Web**	**n**	**Mail**	**n**
Starting Pool	2/3 of new RDD	26,700	KN's on-going RDD	?	1/3 of new RDD	13,300
Sample Drawn	Post-screening; under-sampled ½ if no address	13,475	Post-profiling; under-sampled ½ if no address	1507	Post-screening; not drawn if no address	4185
Completions, C	Phone interviews	1273	Web submissions	1162	Mail returns	904
Total Refusals, R	All unconverted	4140	All non-responding	345	Mail plus calls	1291
(after selection, Ra)	(Selectee break-offs/not reached)	(134)	(No web submission)	(345)	(Mail: declined) (Calls: promised)	(112) (183)
(before selection, Rb)	(Unconverted abrupt refusals)	(4106)	[Need recruiting data]	?	(Calls: abrupt refusals/hang-ups)	(996)
Other eligible, O	Priv. ans. mach./ failed appointment	1872	[None]	0	Calls: private ans. mach./appntmnt.	352
Not eligible, N	Business/govt./ non-working/ fax/language	4778	[Already eliminated during recruitment]	0	Business/govt./ non-working/ fax/language	1036
Unknown, U	No contact: always busy, etc.	1412	[Already eliminated during recruitment]	0	No mail return; no call contact	602
Completion Rate	C*/*(C+Ra)	90.5%	C*/*(C+Ra)	77.1%	C*/*(C+Ra)	75.4%
Cooperation Rate	C*/*(C+Rb)	23.7%	[Need recruiting data]	?	C*/*(C+Rb)	47.6%
Response Rate	AAPOR's RR3	15.6%	KN's CUMRR2	3.9	AAPOR's RR3	29.8%

Response rates for phone and mail are calculated from the Response Rate 3 formula from the American Association for Public Opinion Research. See text for discussion of the response rate calculation for the web.

A pre-selection cooperation rate on the web survey cannot be calculated from the data available to us. In its early years, KN was credited with "a cooperation rate of about 56%" (Huggins and Eyerman, 2001), but that figure is of limited comparative value here. KN's cooperation rate has almost certainly declined since then, consistent with a national trend of declining willingness to respond to telephone surveys. Indeed, the current KN cooperation rate is probably closer to what we found in our phone survey (23.7%) than that in our mail survey (47.6%). We base this speculation in part on the similarity between KN's phone recruitment by RDD and a straight phone survey. Also, we note that KN's household recruitment rate (Appendix F, page 76) is not much better than the overall response rate in our phone survey (22.0% versus 15.6%). KN's recruitment does not require within-household selection of a specific respondent, and it includes a substantial economic incentive in the form of free Internet service. Since those two advantages yield only modestly higher response for the web survey at the recruitment stage than in our phone survey, the web cooperation rate is also likely to be similar to that in the phone survey.

Our mail survey achieved the best response rate, at 29.8%, followed by the phone survey at 15.6% and the web survey at a strikingly low 3.9%. As noted above, even more intensive survey effort, such as an incentive payment, would likely have improved response, but could not have been applied equivalently across all modes.

On the web survey, no amount of additional survey effort on our part could have made much difference. If all 1507 web panelist who were included in our sample had responded (i.e., if we had achieved a web completion rate of 100%), the response rate would still have been only 5.1% (since .039/.771 = .051). The Cumulative RR2 value for the web survey is mainly determined by the recruitment response rate, the rate of successful profiling, and the panel retention rate, all of which are outside our control.

KN's alternative calculation (their CUMRR1) yields a somewhat higher number (9.3%). Whatever the precise value, a response rate of less than 10% (or even 30% as in our mail survey) raises major concerns about potential non-response bias. At the very least, a response rate in that range suggests the need to adjust the demographic distribution of survey respondents, via weighting, to mirror the demographic characteristics of the U.S. population.

5.4. Weighting, Matching, and Benchmarking

KN's weighting process for the web sample is detailed in Appendix F. Although weighting was less complex for our phone and mail surveys, the essential approach is the same.

Our weighting began by addressing known under- and over-sampling inherent in our design. No-address households in the phone (and web) surveys had been intentionally under-sampled by a factor of ½, and therefore the no-address respondents were up-weighted by a factor of 2 to compensate. Households with multiple landline phone numbers (as determined within the questionnaire) had a proportionally greater chance of being drawn in the RDD sample than single-line households. Such respondents were therefore down-weighted by the reciprocal of the number of landlines. Conversely, within-household selection of one adult per household meant that any given adult in a multi-adult household had a reduced probability of being selected. Respondents were therefore up-weighted by the number of adults in the household.

The next step in weighting was post-stratification, "to reduce the sampling variance ... and ... reduce bias due to survey non-response" (Appendix F, page 13). Respondents in all three modes were weighted to reflect the distribution of the U.S. adult population on key demographic characteristics: gender, age, racial/Hispanic identification, education, geographic region, and metropolitan residence. For the web sample, KN also weighted by the pre-recruitment Internet access of the household, but this variable was not available (and unnecessary) for our phone and mail samples.

The population benchmarks used for post-stratification were obtained from the U.S. Census Bureau. Adjusting the distributions of respondents in each of our three survey modes to the same set of benchmark demographics should reduce difference across modes that are attributable to coverage error or differential non-response. This is in keeping with our objective of isolating true mode effects, attributable to the medium through which the survey questions were posed. Using Census data for the common benchmarks has the additional advantage of facilitating generalization to the U.S. population. To the extent that the demographic characteristics used for weighting are correlated with other demographics, behaviors, and attitudes, the weighted samples should closely resemble the population of interest on the variables in this study.

With the post-stratification weighting, each sample of respondents should be representative of the entire population of U.S. adults.

To ensure comparability across modes in the application of post-stratification weights, WYSAC contracted with KN to weight the phone and mail respondents using the same algorithm (iterative proportional fitting), the same Census benchmarks, and the same weight-trimming criteria that KN used for the web sample (see Appendix F). The final step in that process was to scale the weights so that the weighted sample sizes within each mode equaled the actual numbers of respondents.

By definition, the mail survey covered only those households for which a valid, deliverable address was obtained through reverse-lookup of the RDD telephone number. Therefore, while initially the RDD sampling frame had identical coverage for all three survey modes, the effective coverage for our mail survey is unavoidably quite different. Weighting alone may not be sufficient to compensate for a coverage difference of that magnitude.

Another, smaller coverage difference is that the web and mail samples, especially, were found to include some households that reported being without a landline telephone, whereas almost all of the phone respondents had at least one landline. While the RDD frame targeted only landline phone numbers, the web and mail samples were not actually surveyed by phone. That meant more households reachable by those modes could have recently given up their landlines for cell phones (or for no phone service at all). The handful of phone respondents who reported no landline were likely reached on a landline number that had been forwarded or ported to a cell.

Table 5.3. Number of Cases, by Mode, for Matched and Total Samples

Sample	Phone	Web	Mail
All completed questionnaires	1273	1162	904
Fully matched respondents	1038	836	816
Partially matched respondents	1126	937	889
All non-completions	12,202	345	3281
Partially matched non-respondents	6195	248	2245
Non-matching non-respondents	1229	97	0
Known ineligibles	4778	0	1036
Total cases drawn	13,475	1507	4185

Fully matched cases consist of all eligible respondents who have a valid reverse-lookup address that was not determined to be undeliverable, provided that they also have at least one landline phone in the household. Partially matched cases include both respondents and non-respondents with a seemingly valid reverse look-up address, provided that they were not definitively identified as having an undeliverable address or as being otherwise ineligible.

To reduce the effect of any such differences in effective coverage across modes, much of our analysis is reported using matched subsamples. The matching excludes cases from any mode for which no deliverable reverse-lookup address was obtained (as determined by the mailings, including thank-you mailings to the phone and web respondents). Matching also excludes all respondents who reported on the questionnaire that there was no landline phone in their household. The matched subsamples therefore come even closer to holding the sampling frame constant across all three modes than does the shared RDD design alone. As shown in Table 5.3, matching reduced the number of cases available for analysis to 1038 in phone sample, 836 on the web, and 816 by mail.

Although the full samples had been weighted to a common set of demographic distributions, excluding cases to create the matched subsamples introduced some cross-mode variation. The next step was therefore to re-weight the matched subsamples to a single, shared set of benchmarks. However, the matched subsamples represent a very specific segment of the population – all U.S. adults in households with a deliverable reverse-lookup address and at least one landline phone. No Census benchmarks are available for that specific population.

A reasonable alternative to using Census benchmarks was to adjust the distributions for two of our three matched subsamples to correspond to the distributions of the third. Choosing which of the three to use for the matched benchmarks was arbitrary; we used the web distributions for that purpose. Accordingly, the matched phone and mail subsamples were re-weighted to reflect the (Census-weighted) distribution of the matched web subsample on gender, age, racial/Hispanic identification, education, region, and metropolitan residence. As the starting point for this re-weighting, the first set of post-stratification weights comprised the base weights.

The re-weighting should minimize the effects of any between-mode differences in coverage and/or in non-response that matching did not eliminate. Here again, we contracted with KN to perform the re-weighting by iterative proportional fitting. And again, the new weights were scaled so that the weighted subsample sizes equal the actual numbers of cases within each matched subsample.

After matching and re-weighting, each matched subsample of respondents should be representative of the population of all U.S. adults who have a deliverable, reverse-lookup address and at least one landline, residential phone.

The full matching required information about address status (obtained from the thank-you mailings that went only to survey respondents) and about number of landlines (obtained from responses to an item on the questionnaire). Therefore, full matching is possible only when comparing respondents across the modes, not for any analysis involving non-respondents. Analyses comparing respondents to non-respondents, by mode, can at best rely on what we term "partial matching."

As with full matching, the objective is to reduce the effect of coverage differences across modes. Partial matching does so less thoroughly than full matching, because it uses the less precise information that is available not only on respondents but also on non-respondents. Partially matched cases in all three modes are those with a seemingly valid reverse look-up address, provided that they were not definitively identified as having an undeliverable address or as being otherwise ineligible. As shown in Table 5.3, partial matching yields 1126 respondents (plus 6195 non-respondents) for analysis in the phone sample, 937 (plus 248) on the web, and 889 (plus 2245) by mail.

Of our three survey modes, the partially matched phone sample (pooling respondents with non-respondents) gives us the most accurate picture of the corresponding segment of the RDD sampling frame. In the phone sample, unlike the web sample, no cases have been lost due to panel attrition. Also, in the phone sample most of the ineligible cases among the non-respondents have been culled out through the survey calling. We can therefore pool the partially matched respondents and non-respondents from the phone sample and use their demographics as yet another set of benchmarks.

Without any weighting, the pooled, partially matched phone sample of respondents and non-respondents should be representative of all eligible U.S. landline telephone numbers with reverse-lookup addresses.

When examining non-respondents in any mode, the partially matched samples cannot be weighted demographically, because there is no demographic information about non-respondents on which to weight. But it may be of interest to compare the partially matched respondents only (without the non-respondents), against the pooled-phone sample benchmarks. If so, then weighting is needed, but the weights used for other purposes should be fine-tuned. For the pooled phone-sample benchmarks, the relevant level of analysis is households (or, strictly speaking, phone numbers), not individual adults. Therefore to properly weight the partially matched respondents, the previous (Census-based) post-stratification weights must be partially un-weighted by a factor of (j/k), where j is the reported number of landline phones and k is the number of adults in the household. That partial un-weighting converts the respondents back from a sample of adults to its original status as a sample of phone numbers, while retaining the adjustment for Census demographics.

After partial unweighting, each partially matched sample of respondents should be representative of the population of all eligible U.S. phone numbers with reverse-lookup addresses.

6. Comparing the Results across Modes

6.1. Statistical Methods

Appendix A provides unweighted frequency counts and percentages for every variable and each survey mode, as well as percentages that reflect the weighting adjustments just discussed. In this section we summarize and highlight the key findings from those analyses.

Except as otherwise noted, all results in Section 6 (and in Section 7 that follows) are based on the fully matched respondents, weighted to correspond to the Census-weighted demographics of the web sample. These results may therefore be generalized to the population of all U.S. adults who have a deliverable, reverse-lookup address and at least one landline, residential phone. With sampling frame held virtually constant, any remaining differences should be attributable to mode effects as such, and/or to differential non-response not completely counteracted by the weighting.

Chi-square tests of statistical significance are reported in Appendix A for all of the matched, weighted cross-tabulations summarized in this section. Each Pearson's chi-square gives an overall test that is sensitive to any kind of differences in the distribution of survey responses across all three modes. In addition, p-values are also indicated for tests comparing the survey modes two at a time (phone/web, phone/mail, and mail/web).

In the two-mode comparisons, when the response being analyzed is a set of unordered categories (e.g., racial identification), the test reported is the usual Pearson chi-square. When the response being analyzed is ordinal (e.g., income groups, from lowest to highest), the test is the Mantel-Haenszel chi-square for linear trend (hereafter, MH). The MH test has greater statistical power than the Pearson for detecting a significant association if the relationship is indeed ordinal (that is, if one survey mode has a fairly consistent tendency to score higher on the response variable than another mode). Conversely, the MH test will identify an association as non-significant, even when there are some differences between the two modes, if those differences do not form a fairly consistent ordered pattern (see Agresti, 1996).

Because the total sample size is large (2690 cases after matching), statistical significance is rather easy to achieve in the overall tests. The two-mode tests, especially the MH, highlight associations that, because they are more specific, are more readily interpretable. The MH tests for ordinal association are therefore the major focus of attention in this section, but the overall tests are also summarized.

Given the large number of statistical tests in Appendix A, caution is warranted in drawing conclusions from borderline "significant" results. About 5 out of 100 comparisons will test as significant at the .05 level, by chance alone, when there is no association at all in the population. In the results below, no formal adjustment for multiple comparisons (such as the Bonferroni) has been made. Less formally, we seek to avoid capitalizing on chance by labeling significance at the .01 level as a "notable" difference. We refer to significance between .01 and .10 as "marginal," and differences that are not even significant at the .10 level as "negligible." All of the p-values reported assume two-tailed tests.

6.2. Demographic Differences

We begin this discussion near the end of the questionnaire, by comparing the three matched and weighted subsamples on their demographic characteristics. Table 6.1 summarizes the fully detailed results from Appendix A, Tables Q11.1 through Q21.2.

Using the overall chi-square test, six of the 11 possible demographic differences by survey mode are notably significant ($p < .01$), and another is marginally so. Perhaps surprisingly, two of the notable overall associations are with variables that were used in the post-stratification weighting (education and age). Weighting should have produced very similar distributions across modes on each of the weighting variables. The explanation for these two anomalies is that only four categories of education and age were used for the weighting, whereas the cross-tabulations and statistical tests in Appendix A used more fine-grained categories. Though the weighted distributions are indeed very similar across modes using the broad, 4-category classification, they are not so nearly identical when the 4 categories are subdivided.

Importantly, however, the mode differences on education, age, and most of the other demographics do not form strong and consistent ordinal trends. For both education and age, the MH test for linear trend is only marginally significant comparing mail to web respondents, and does not approach significance in either of the other two-mode comparisons. With some notable exceptions to be considered next, matching and weighting seem to have controlled adequately for demographic differences across the modes.

The variables that show the most significant ordinal associations in the two-mode MH tests are membership in an environmental group, household size (both number of adults and number of children), breathing problems like asthma, and income. The web panelists are less likely than either the phone or the mail respondents to belong to an environmental organization. The invitation to participate in our survey described its topic as "issues facing national parks, like air quality." We suspect that people reached by phone or mail may be more interested in that topic, and hence more likely to respond, if they hold membership in an environmental group. The web panelists, on the other hand, had agreed during KN's recruitment process to participate in surveys on a variety of topics. Their rate of response to any particular survey is likely to be less sensitive to topic. If so, then phone and mail surveys will be more susceptible to non-response bias due to self-selection for interest in the topic of the questionnaire.

The web panelists in our survey tend to live in smaller households – with fewer adults and, especially, fewer children – than phone or mail respondents. With smaller households, it is not surprising that web respondents are less likely to have anyone in the household with breathing problems. Web panelists have lower income than the other two samples, and mail respondents have the highest income, on average, among the three modes.

In sum, it appears that matching and weighting have not completely eliminated demographic differences across the modes. Apparently, the demographic variables used in weighting are strongly correlated with some but not all other demographics. Notable differences remain in the weighted distributions of some variables not used explicitly for weighting. With less environmental activism, lower income, and fewer household members who suffer breathing problems, web panelists might be less willing to pay for improving air quality in national parks. Statistical controls for such variables will be essential in our econometric modeling. These results do not specify the source of the remaining demographic variations, but differential non-response by mode is a possibility that merits exploration.

Table 6.1. Mode Differences on 11 Demographic Variables for Matched Subsamples

Difference	Overall	Phone vs. Web	Phone vs. Mail	Mail vs. Web
Notable	Education, q11			
(p < .01)	Age, q12			
	Membership, q13	**Membership** +		**Membership** +
	Adults, q15		Adults (-)	Adults +
	Children, q16	**Children** +		**Children** +
				Asthma +
	Income, q21		Income (-)	Income +
Marginal		Adults +		Education (-)
(.01 < p < .10)				Age +
		Phones (-)		**Phones** (-)
	Asthma, q18	Asthma +		
		Income +		
Negligible		Education	Education	
(p > .10)		Age	Age	
			Membership	
	Gender, q14	Gender	Gender	Gender
			Children	
	Phones, q17		Phones	
			Asthma	
	Hispanic, q19	Hispanic	Hispanic	Hispanic
	Race, q20	Race	Race	Race

Source: Appendix A. Overall p-values are from Pearson chi-square tests across all three survey modes. Two-mode p-values are from Mantel-Haenszel tests for ordinal association between the pair of modes identified in the column headings. Signs indicate the direction of the significant ordinal associations. For example, in the column headed Phone vs. Web, a positive sign in any cell means that the phone sample tends to have higher scores on the row variable than does the web sample. Variable names are **bolded** if there is a significant ordinal difference between the web sample and both of the other two samples. Variable numbers (q11, q12, etc.) correspond to the numbering in Appendix A.

6.3. Geographic Differences

Next we consider some data obtained from the sampling company (Marketing Systems Group), rather than directly from the survey respondents. These variables (numbered as x1 through x10 in

Appendix A) record characteristics of the telephone exchange (area code plus first three digits of the RDD phone number), as estimated by the sampling company. For narrative convenience only, we refer to respondents as "living in" their respective exchange, as if the exchange were an easily recognized geographic area or neighborhood. Results are summarized in Table 6.2, and provided in detail in Appendix A (after the last of the survey questions, q26).

Consistent with some of the individual-level demographic differences just discussed, the responding web panelists live in exchanges with a lower density of persons per household, relatively fewer children, and more senior citizens than mail respondents. Web and phone respondents are quite similar to each other in these respects.

Table 6.2. Mode Differences on 9 Geographic Variables for Matched Subsamples

Difference	Overall	Phone vs. Web	Phone vs. Mail	Mail vs. Web
Notable	Density, x4			
(p < .01)	Kids, x5		Kids (-)	Kids +
			Seniors +	Seniors (-)
	Whites, x8			
	Poverty, x9	**Poverty** +		**Poverty** +
	Affluence, x10		Affluence (-)	
Marginal		Region	Density (-)	Density +
(.01 < p < .10)	Young, x6		Young +	
	Seniors, x7			
		Whites +	Whites +	
		Affluence (-)	Poverty +	
Negligible	Region, x2		Region	Region
(p > .10)	Metropolitan, x3	Metropolitan	Metropolitan	Metropolitan
		Density		Whites
		Kids		
		Young		Young
		Seniors		Affluent

Source: Appendix A. Overall p-values are from Pearson chi-square tests across all three survey modes. Two-mode p-values are from Mantel-Haenszel tests for ordinal association between the pair of modes identified in the column headings(except for the unordered row variable Region, for which the standard Pearson chi-square is used). Signs indicate the direction of the significant ordinal associations. For example, in the column headed Phone vs. Web, a positive sign in any cell means that the phone sample tends to have higher scores on the row variable than does the web sample. Variable names are **bolded** if there is a significant ordinal difference between the web sample and both of the other two samples. Variable numbers correspond to the numbering in Appendix A. For Phone vs. Web only, a tenth geographic comparison is available: address deliverability (x1); p > .30 for that comparison. Deliverability is a constant on the mail survey, and so no other tests involving this characteristic are possible.

However, web panelists who responded live in exchanges with less poverty (i.e., their exchanges have a lower proportion of households with annual household incomes less than $10,000) than either phone or mail respondents. This contrasts with the individual-level results, in which the web panelists tend to have lower income than respondents from either of the other modes. While KN's offer of free Internet service may be appealing to people of modest means, it appears that surveys by phone or mail generate more responses from areas of extreme poverty.

The rest of the exchange-level variables show no strong or consistent mode differences. The phone respondents tend to be in exchanges where non-minorities (non-Hispanic whites) predominate, to an even greater extent than in the other two modes, but those mode differences are only marginally significant. Overall, the exchange-level data do not raise major new concerns about geographic or associated demographic differences across modes.

6.4. Mode Differences in Non-response

Our research was designed to focus primarily on mode effects in survey *responses*. The design decisions that we made with that goal uppermost present some difficulties for any comparative analysis of *non*-response across modes.

First, we have only limited information about the non-responding cases. In all three modes, we have exchange-level geographic data on non-respondents as well as on the respondents. We also know which responding and non-responding cases had a reverse-lookup address and (usually) whether or not the RDD phone number reached an eligible private household. Beyond that, we know essentially nothing else about the phone and mail non-respondents.

For the web survey, we do have some additional information from KN's profiling database. However, only "end-stage" non-response is visible to us in our web data. We have no information about web non-response at the recruitment stage, which corresponds to the main point of non-response in our phone and mail surveys. And we also have no information from the web survey concerning panel attrition between the time of initial recruitment and eventual respondent selection into our web sample. Thus, comparing the pattern of end-stage non-response in the web survey to total non-response in the other two modes cannot definitively identify mode differences in non-response, because stage in the response process cannot be held constant across modes.

Between-mode comparisons face the further difficulty of unequal total sample sizes. In the phone and mail modes, the number of non-responding cases is very large, whereas there are relatively few (end-stage) non-respondents to the web survey. Considering only partially matched cases (as described in Section 5.4), there are more than 6000 phone non-respondents and over 2000 mail non-respondents, but fewer than 250 non-responding web panelists. In statistical tests, even trivial differences between respondents and non-respondents in the phone or mail samples are likely to be "significant," and hence to be mistakenly taken as substantively important. Yet a substantively meaningful difference between those two subgroups of the web sample might not achieve statistical significance because of the much smaller number of non-respondents.

In the same grant program under which EPA supported the present study, other research teams were funded to undertake investigations focusing on non-response, including web-panel recruitment and attrition. We therefore defer to those other researchers on the issue. While acknowledging the importance of differential non-response by mode, we forgo an elaborate analysis here. Instead, we provide a short summary of some exploratory comparisons that are included in Appendix A.

Our exploratory analyses follow the template laid out in Tables X2.3 and X2.4 in the appendix. Within each mode, Table X2.3 compares partially matched respondents versus non-respondents on geographic region. Table X2.4 compares the respondents from each mode to the regional distribution of the pooled benchmark sample. There are two similar tables for each of the other exchange-level variables.

We note that the partially matched sample sizes in these tables differ somewhat from those presented in Table 5.3, due to a late refinement in our definition of partial matching. For our present exploratory purposes, the differences are negligible. Also, the tables in section 4.3 of Appendix A (those involving the exchange-level variables) have not been adjusted to correct for a programming error by KN that we identified shortly before finalizing our report. As a result of that error, KN had misclassified 22 non-respondents as respondents in the web data file. We have corrected KN's error throughout our descriptive and econometric analyses below, including the tables in Appendix A, except those involving the exchange-level variables.

We also made the appropriate corrections in our presentation of response rates above (Section 5.3). KN did not correct the error in the frequency tabulations that they provided as part of their revised field report (Appendix F, pages 57-72). They corrected some but not all of the figures in their calculations of response rates (Appendix F, page76), where the corrected figures are in the bolded headings but the uncorrected figures remain in plain text immediately beneath each heading.

With these caveats we turn to Tables X2.3 and X2.4 in Appendix A (along with the corresponding subsequent tables), which provide the results of significance tests for the differences between respondents and non-respondents (within mode), and for the differences between respondents for each pair of modes. These tests were run two ways: once, using the actual subsample sizes, and again, with an adjustment setting all subsample sizes to 897, the number of mail respondents. The adjustment lets us use the p-values for comparing across modes the relative strength of the association between (say) region and response status. Although it is *ad hoc*, this simple adjustment suffices for the exploratory purposes here.

By that rough measure, the differences between respondents and non-respondents are greatest for the web panel, and least for the mail survey. In the web survey there is a notably significant difference ($p < .01$, setting each n to 897) between respondents and non-respondents on 6 of the 9 exchange-level variables. In the mail survey, only 2 of the 9 comparisons reach that level of significance, and in the phone survey, it is 4 of 9.

Regional differences between respondents and non-respondents are notably significant only in the mail survey. With region categorized into 9 Census divisions, responses to the mail survey were heaviest from the West North Central region. Overall, however, the regional distributions of respondents and non-respondents look pretty similar, regardless of mode.

In all three surveys, respondents are much more likely than are non-respondents to be in ethnically white, non-Hispanic exchanges. Besides region and racial/ethnic composition, there are no notably significant non-response differences in the mail survey on any of the other 7 exchange-level variables (with the non-respondent n set to 897).

In both the phone and the web surveys, there are highly significant differences between respondents and non-respondents, not only in racial composition of the exchange but also in metropolitan status, persons per household, and proportion of children. Compared to non-respondents, both phone and web respondents come disproportionately from non-metropolitan exchanges with lower household density and fewer children. Web respondents are also more likely than web non-respondents to live where there are more senior citizens and less poverty.

These differences between web respondents and non-respondents are quite similar to the between-mode differences (among respondents only) identified in the two preceding sections. Though merely exploratory, this analysis therefore suggests that the geographic mode differences that we found can be largely attributed to non-response bias in the web survey.

With respect to the geographic variables on which we have data, end-stage non-response bias in the web panels seems to be greater than total non-response bias in our phone survey. Total non-response bias on these variables appears to be lowest of all in our mail survey. On the other hand, the mode differences in environmental membership (combined with differences in recreational behavior to be discussed in Section 6.6) suggest the potential for self-selection bias by phone and, especially, by mail. Work by the other research teams should prove more definitive on this issue.

6.5. Differences in Survey-taking Behavior

Our survey generated information about several different survey-taking behaviors. Some survey behaviors can be inferred quite simply, from whether or not the respondents answered certain substantive items on the questionnaire. Other behaviors can be measured using meta-questions about the respondents' recent survey participation. The results are summarized in Table 6.3.

Table 6.3. Mode Differences on 6 Survey-taking Behaviors for Matched Subsamples

Difference	Overall	Phone vs. Web	Phone vs. Mail	Mail vs. Web
Notable	Suggestion, q4	**Suggestion** +	Suggestion +	**Suggestion** (-)
(p < .01)	Phone surveys, q22	**Phone surveys** +	Phone surveys +	**Phone surveys** (-)
	Mail surveys, q23			
	Web surveys, q24	**Web surveys** (-)	Web surveys +	**Web surveys** (-)
	Birthday, q25	**Birthday** +		**Birthday** +
Marginal			Protest bid (-)	
(.01 < p < .10)			Mail surveys (-)	
			Birthday (-)	
Negligible	Protest bid, q9	Protest bid		Protest bid
(p > .10)		Mail surveys		Mail surveys

Source: Appendix A. Overall p-values are from Pearson chi-square tests across all three survey modes. Two-mode p-values are from Mantel-Haenszel tests for ordinal association between the pair of modes identified in the column headings. Signs indicate the direction of the significant ordinal associations. For example, in the column headed Phone vs. Web, a positive sign in any cell means that the phone sample tends to have higher scores on the row variable than does the web sample. Variable names are **bolded** if there is a significant ordinal difference between the web sample and both of the other two samples. Variable numbers correspond to the numbering in Appendix A. For Phone vs. Web only, a seventh survey behavior is available: concluding comment (q26); p < .01 for that comparison, with phone respondents more likely than web respondents to provide a comment when invited to do so. Q26 was not asked on the mail survey.

On open-ended questions, phone respondents were notably more expressive than web respondents, with mail respondents being the most reticent. This is hardly surprising: a spoken answer, taken down by the phone interviewer, requires less respondent effort than entering text at a computer keyboard, and a hand-written answer is the most laborious of all. Hence, when asked on item q4 what NPS could do to encourage more park visitation, almost 90% of phone respondents offered some kind of suggestion, compared to 80% on the web and only 60% by mail. Also, when invited to give a concluding comment upon completion of the questionnaire, almost half of phone respondents but less than a quarter of web respondents did so. (The mail questionnaire did not include that invitation.)

A similar pattern is evident in the open-ended item (q9) that asked for an explanation of the respondent's answer to the willingness to pay question. Phone respondents were most likely to provide any explanation at all, and mail respondents least likely. Comfortingly, however, there was not much difference across modes in the content of those explanations, at least as regards the key issue of protest bidding that was the main purpose of q9. Based on our coding of the open-ended answers, phone respondents were in fact marginally *less* likely to be protest bidders than mail or web respondents, who differed hardly at all from each other in that respect. Therefore, mode differences in item non-response on the open-ended q9 should not adversely affect the identification of protest bidders.

Item non-response on the closed-ended questions is not addressed in Table 6.3, but was generally low across all modes. An exception is household income (q21), on which item non-response ranged from about 9% among web panelists to more than 20% by phone. As expected, the greater privacy of the self-administered web and phone surveys yields fewer refusals on this sensitive item. Some of the attitude items discussed in Section 6.7, below, are also exceptions, in that many respondents in all three modes provided no answer if they were uncertain, rather than choosing the neutral category. On one of those items, in particular, that was especially true of phone respondents; see Section 6.7.

On the meta-questions, Table 6.3 shows that phone respondents reported significantly more frequent participation in phone surveys than either web or mail respondents, with web respondents also significantly higher on this behavior than mail respondents. Conversely, mail respondents reported marginally more frequent participation in previous mail surveys. And of course, web panelists had completed far more Internet surveys than either phone or mail respondents.

All this suggests an element of self-selection in each mode: our phone survey was completed disproportionately by people who do other phone surveys more than the average person does; our mail survey was completed by people who do other mail surveys more often; and our web survey was necessarily completed by panelists who do frequent web surveys.

As a check on the within-household selection of the quasi-random adult with the most recent birthday, the last question on the survey instrument explicitly asked whether the survey had been completed by the adult with the most recent birthday. The web panelists are significantly less likely than phone or mail respondents to confirm that they had the most recent birthday. However, this seeming anomaly is mainly due to an inconsequential methodological artifact. KN drew the sample for our web survey, using their panelists' dates of birth for the within-household selection, a full six weeks before the invitation to participate in our web survey went out, and ten weeks before the field period ended. As a result, some of the correctly selected respondents in our web sample were no longer the most recent birthday by the time they completed the questionnaire. But the time lag is of absolutely no consequence for the representativeness of the original within-household selection.

More meaningful is that over 85% of both phone and mail respondents confirmed, at the end of the questionnaire, that they did indeed have the most recent birthday. This percentage may be somewhat inflated by the respondents' natural inclination to say they complied with the survey's initial instructions, even if they did not. But discounting the 85% figure somewhat, the within-household selection of a specific, quasi-random respondent still appears to have succeeded in the great majority of households. The difference in success between phone and mail modes is small, with a marginally higher proportion of the mail respondents confirming their status as the correct "birthday" respondent.

6.6. Differences in Recreational Behavior

As summarized in Table 6.4, there are clear and consistent differences on all of the items in the first part of the questionnaire concerning park visitation and outdoor activities. By any of three different measures, the mail respondents are notably more likely to visit a national park site than either phone or web respondents. And both mail and phone respondents are notably more likely than web panelists to engage in each of the five outdoor activities listed on the questionnaire.

Table 6.4. Mode Differences on 8 Recreational Behaviors for Matched Subsamples

Difference	Overall	Phone vs. Web	Phone vs. Mail	Mail vs. Web
Notable	Ever visited, q1		Ever visited (-)	Ever visited +
(p < .01)	Recent visits, q2		Recent visits (-)	Recent visits +
	Planned visit, q3		Planned visit (-)	Planned visit +
	View nature, q6a	**View nature** +		**View nature** +
	Hike or jog, q6b	**Hike or jog** +		**Hike or jog** +
	Snow sports, q6c	**Snow sports** +		**Snow sports** +
	Water activities, q6d	**Water activities** +		**Water activities** +
	Hunt or fish, q6e	**Hunt or fish** +		**Hunt or fish** +
Marginal		Recent visits +		
(.01 < p < .10)			View nature (-)	
Negligible		Ever visited	Hike or jog	
(p > .10)		Planned visit	Snow sports	
			Water activities	
			Hunt or fish	

Source: Appendix A. Overall p-values are from Pearson chi-square tests across all three survey modes. Two-mode p-values are from Mantel-Haenszel tests for ordinal association between the pair of modes identified in the column headings. Signs indicate the direction of the significant ordinal associations. For example, in the column headed Phone vs. Web, a positive sign in any cell means that the phone sample tends to have higher scores on the row variable than does the web sample. Variable names are **bolded** if there is a significant ordinal difference between the web sample and both of the other two samples. Variable numbers correspond to the numbering in Appendix A.

Either or both of two mechanisms could account for these differences. Both mechanisms are related to self-selection of survey respondents. On one hand, interest in outdoor recreation and national parks may determine whether a household contacted by phone or (especially) by mail decides to participate in the survey. Hence frequent park visitors and other outdoor enthusiasts (along with members of environmental organizations) would be over-represented in the phone and mail responses. The web panelists had already agreed during KN's recruitment to participate in a variety of surveys, so the survey topic should make less difference in their decision to complete any particular one.

On the other hand, agreement to participate in the web panel may have been easiest for KN to secure from people whose preferred leisure activities are sedentary. To a notable degree, the KN panelists are web surfers, not wave surfers or snowboarders. Hence park visitors and outdoor enthusiasts would be under-represented in the web responses.

There are no Census figures on recreational activities that could serve as benchmarks for the variables in Table 6.4. Thus, while the results provide strong evidence of differential non-response by mode, they do not specify which mode most closely resembles the population on those variables. They do underscore the necessity of statistical controls for such factors in the econometric modeling. Frequent visitors and outdoor recreationists are likely to value clean air in national parks quite differently from non-visitors and the sedentary.

6.7. Attitudinal Differences

Like the recreational behaviors, all of the opinion items measured in the questionnaire show significant mode differences in the overall chi-square test. However, unlike the recreational behaviors, few of the attitudinal differences are notably significant in the more specific MH test for

Table 6.5. Mode Differences on 10 Opinion Questions, for Matched Subsamples

Difference	Overall	Phone vs. Web	Phone vs. Mail	Mail vs. Web
Notable	Satisfied, q5	**Satisfied** +		**Satisfied** +
(p < .01)	Restore wildlife, q7a			Restore wildlife +
	Remove animals, q7b			
	Basic facilities, q7c	**Basic facilities** +		**Basic facilities** +
	Major facilities, q7d		Major facilities (-)	Major facilities +
	Limit vehicles, q7e			
	Snowmobiles, q7f			
	Air pollution, q7g			
	Willing to pay, q8	Willing to pay +	Willing to pay +	
	Maximum, q10		Maximum +	
Marginal		Restore wildlife +	Restore wildlife (-)	
(.01 < p < .10)		Remove animals +		
		Snowmobiles +	Snowmobiles +	
Negligible		Major facilities	Satisfied	
(p > .10)			Remove animals	Remove animals
			Basic facilities	Snowmobiles
		Limit vehicles	Limit vehicles	Limit vehicles
		Air pollution	Air pollution	Air pollution
				Willing to pay
		Maximum		Maximum

Source: Appendix A. Overall p-values are from Pearson chi-square tests across all three survey modes. Two-mode p-values are from Mantel-Haenszel tests for ordinal association between the pair of modes identified in the column headings. Signs indicate the direction of the significant ordinal associations. For example, in the column headed Phone vs. Web, a positive sign in any cell means that the phone sample tends to have higher scores on the row variable than does the web sample. Variable names are **bolded** if there is a significant ordinal difference between the web sample and both of the other two samples. Variable numbers correspond to the numbering in Appendix A.

linear trend. In other words, with a few exceptions, the attitudes tend not to differ across mode in a systematic, ordered way. The results are summarized in Table 6.5.

Regarding the systematic differences, both phone and mail respondents are more satisfied with the National Park Service than the web panelists, and more in favor of having basic visitor facilities in the parks, such as roads, trails, and restrooms. Consistent with their more frequent visitation, mail respondents are the most supportive of major facilities such as lodging, restaurants, and stores. Consistent with their greater membership in environmental organizations, mail respondents are also the most supportive of bringing back animals that were formerly native to the parks.

We note that item non-response is somewhat higher on the opinion items (q7, parts a through g) than elsewhere in the questionnaire, primarily among telephone respondents. This is especially true on the last such item, concerning the severity of air pollution in national parks. Anecdotal reports from the telephone interviewers suggest that many respondents tried to engage the interviewers in a conversation about their uncertainty on that question (and to a lesser extent, on some of the other opinion items in this part of the questionnaire). In such cases, rather than push the respondents to pick a category, the interviewers followed their training, coded the item as no-answer, and moved on. In the web and mail questionnaires, with no opportunity to converse about their uncertainty, some respondents skipped the question but others probably selected the neutral response, "neither agree nor disagree." For purposes of comparing the results across modes, the results in Appendix A and Table 6.5 are based on recoding all missing data to the neutral category on items q7a through q7g, in all three modes.

Compared to both mail and web, the phone respondents are notably more likely to accept whatever bid they received on the first willingness to pay question. Similarly, on the open-ended willingness to pay questions, the phone respondents report a higher maximum than the mail respondents. However, as documented in Sections 6.2 through 6.7, weighting and matching the three samples have not eliminated all potentially relevant demographic, behavioral, and attitudinal differences. Therefore, statistical controls for these other mode differences are needed before drawing conclusions about mode effects in willingness to pay. For that purpose, we turn next to econometric modeling of the WTP data.

7. Econometric Analysis

In this section, we build and estimate an econometric model of willingness to pay (WTP) for improved air quality in national parks. The empirical model describes WTP for air-quality improvement from low to medium, which is elicited through a dichotomous-choice question with bid, *b*.

7.1. Three Kinds of Mode Differences in Willingness to Pay

We estimate three different impacts of survey mode on WTP. First, survey mode is allowed to directly affect WTP through intercept dummy variables that shift the level of WTP up or down. Second, we allow for mode-specific heteroscedasticity so that the modes may exhibit different degrees of unexplained variation in WTP. Third, we model the three modes separately to allow for differential effects of the explanatory variables across survey modes (interactions). Another potential impact of survey mode (not modeled here) is self-selection into our sample (Heckman, 1979).

7.1.1. Modeling the Additive Effects of Survey Mode

The baseline model (homoscedastic, without interactions) is

$$WTP_i = X_i'\beta + \beta_M Mail_i + \beta_P Phone_i + \varepsilon_i \quad (1)$$

where X_i is a vector of control variables; β is a vector of coefficients; $Mail_i$ and $Phone_i$ are dummy variables capturing how these two survey modes differ in WTP from *Web* (the omitted, reference category); ε_i is an error term with a mean-zero normal distribution and variance σ (initially, assumed constant across modes); and $i = 1, \ldots, N$ indexes respondents. To maximize the number of observations in the estimation, the vector of control variables includes dummy variables constructed to account for missing data. A variable indicating the absence of a response was created for each substantive explanatory variable. Respondents with missing data on an explanatory variable were assigned a value of zero on that variable. The missing-data dummy variables are included as controls in estimating the model, but omitted from the tables reported here.

Given expression (1), the probability of accepting the bid for improved air quality is:

$$P_i = \Pr(y_i = 1) = \Pr(WTP_i \geq b_i) = \Phi((b_i - X_i^{'}\beta - \beta_M Mail_i - \beta_P Phone_i)/\sigma), \qquad (2)$$

where $y_i = 1$ if the person accepted the offered bid and $y_i = 0$ if the person did not accept the bid. The offered bids are chosen from the following bid vector (developed using focus groups and pre-tests):

$$\text{b} = (\$2, 5, 10, 15, 25), \qquad (3)$$

with probability 0.15, 0.2, 0.3, 0.2 and 0.15, respectively. The non-uniform bid distribution yields a more even division of respondents between those accepting and those rejecting the proposed fee for improved air quality.

Table 7.1 shows the percentage of respondents that accepted the various bids. As expected, the percentage responding “YES” declines as the bids increase. Table 7.2 presents the definitions and sample means for the dependent and explanatory variables. Notice that the percentage of respondents accepting the offered bid (57%) is slightly higher than those refusing the bid (43%). Also notice that, by design, there is also a roughly equal split of respondents across mail (30%), phone (39%) and web (31%) survey modes.

Table 7.1. Percent Responding “YES” to Offered Bid

Bid	Percent “YES”
$2	87.6%
$5	75.7%
$10	56.3%
$15	40.0%
$25	27.9%

Table 7.2. Variable Definitions and Descriptive Statistics

Variables	Definitions	Type	Mean	N
Bid	Referendum chosen from set {$2,$5,$10,$15,$25}	C	11.08	2288
y	Accepted bid	D	0.57	2288
Mail	Mail survey mode	D	0.30	2288
Phone	Phone survey mode	D	0.39	2288
Past Visit	Visited National Park but not in the past two years	D	0.38	2234
Low Visit	Visited once or twice in the last two years	D	0.31	2234
High Visit	Visited three or more times in the last two years	D	0.19	2234
View	Viewing or photographing animals, birds, or plants	D	0.67	2260
Hike	Hiking or jogging for at least 30 continuous minutes	D	0.59	2248
Snow	Snow sports (e.g., skiing, snowmobiling, or sledding)	D	0.19	2224
Water	Outdoor water activities (e.g., swimming or boating)	D	0.60	2242
Hunt	Hunting or fishing	D	0.34	2249
Native Animals	Agree: Native animals should be brought back	D	0.79	2179
Alien Animals	Agree: Alien animals should be removed	D	0.41	2151
Basic Services	Agree: Basic visitor services should be provided	D	0.87	2246
Major Services	Agree: Major visitor services should be provided	D	0.46	2238
Limit Vehicles	Agree: Number of private vehicles should be limited	D	0.72	2206
Snowmobiles	Agree: Jet-skiing and snowmobiling should be allowed	D	0.20	2174
Air Pollution	Agree: Air pollution is causing breathing trouble	D	0.33	1849
College	Undergraduate degree or higher	D	0.48	2267
Env. Group	Member of environmental group	D	0.13	2266
Young	18 – 40 years old	D	0.23	2254
Middle-age	41 – 65 years old	D	0.54	2254
Female	Female	D	0.58	2288
Adult	Number of adults in household	C	1.97	2274
Child	Number of children (Age ≤ 17) in household	C	0.53	2277
Lung	Household member has a breathing problem	D	0.27	2278
Hispanic	Hispanic	D	0.04	2241
White	White	D	0.85	2241
Medium Income	$25K ≤Household income < $75K	D	0.47	2013
High Income	Household income ≥ $75K	D	0.35	2013
ENC	States = (NJ, NY, PA)	D	0.13	2285
WNC	States = (IL, IN, MI, OH, WI)	D	0.17	2285
GP	States = (IA, KS, MN, MO, ND, NE, SD)	D	0.10	2285
SE	States = (DC, DE, FL, GA, MD, NC, SC, VA, WV)	D	0.18	2285
WSC	States = (AL, KY, MS, TN)	D	0.07	2285
SW	States = (AR, LA, OK, TX)	D	0.11	2285
RM	States = (AZ, CO, ID, MT, NM, NV, UT, WY)	D	0.08	2285
PC	States = (AK, CA, HI, OR, WA)	D	0.11	2285
Metro	Lives in metropolitan area	D	0.80	2288
Ozone	Fails to meet EPA's ozone standards within FIPS code	D	0.28	2288

C = Continuous variable; D = Dummy variable. Sample sizes less than N = 2288 indicate missing data.

The interval regression model described above (Wooldridge, 2002) involves selecting the coefficients to maximize the following log likelihood function:

$$lnL(\beta, \beta_M, \beta_P, \sigma) = \sum_{i=1}^{N}[y_i lnP_i + (1 - y_i)\ln(1 - P_i)]. \tag{4}$$

Because the cut points (i.e., bids) are known and vary across respondents, we are able to identify the mode-effect coefficients β_M and β_P. These coefficients are measured in terms of dollar WTP for improved air quality in national parks.

Equation (1) was estimated using the matched and weighted data set, combining 2288 respondents from all three modes. Recall that matching holds the sampling frame approximately constant across modes, by excluding respondents who did not have a deliverable address and at least one landline phone in their household. The weighting further adjusts the demographic distribution of respondents in the matched mail and phone samples to correspond, insofar as possible, with the demographics of the matched web respondents. We excluded 124 respondents (3.7% of the initial sample) who said "Yes" to the bid for the medium improvement in air quality but then stated a lower WTP for the larger improvement (see Section 7.2, on our scope test). We also excluded 194 respondents (7.4% of the initial sample) who responded to the WTP question with a protest bid. In addition, all significance tests in our econometric analyses have been corrected to reflect the inflation in standard errors due to weighting (see Dorofeev and Grant, 2006).

Results from the baseline model are shown under the "Baseline Model" heading of Table 7.3. However, we defer substantive discussion of the results to first introduce more elaborate models for mode differences in WTP.

7.1.2. Modeling Unequal Variances across Modes

The second potential effect of survey mode on WTP is through mode-specific heteroscedasticity. We allow the error variance to follow:

$$\sigma_i^2 = \exp(\alpha_0 + \alpha_1 Mail_i + \alpha_2 Phone_i), \tag{5}$$

where α_1 and α_2 capture the effect of survey mode on error variance. For example, if $\alpha_2 < 0$ then (all else equal) phone surveys have less model uncertainty than web surveys. This could happen if, for example, the interpersonal communication of phone surveys helps respondents converge toward their true WTP. On the other hand, if we cannot reject the null hypothesis that $\alpha_1 = \alpha_2 = 0$ then the baseline econometric model is appropriate and the error variance is approximately constant across survey modes. The results from this model are shown under the heading "Heteroscedastic Model" in Table 7.3.

7.1.3. Modeling Interactions of Mode with the Explanatory Variables

We also estimate the baseline models separately, allowing respondent characteristics to influence WTP differently across modes. The results of the models for mail, phone and web surveys are shown under the headings "Mode-Specific Models" in Table 7.3. A likelihood ratio test for overall interaction by mode examines whether the coefficients on all the explanatory variables are equal across all modes. For each explanatory variable, we also conducted pairwise tests comparing the modes (not shown in the table).

Table 7.3. WTP Estimates for Improved Air Quality in National Parks

Explanatory Variables†	Baseline Model		Heteroscedastic Model		Mode-Specific Models					
					Mail (M)		Phone (P)		Web (W)	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Constant	10.951***	3.983	11.107***	4.022	11.379**	6.755	0.328	6.246	12.027**	6.794
Mail	0.745	1.359	0.269	1.448	--	--	--	--	--	--
Phone	3.670***	1.438	2.961**	1.548	--	--	--	--	--	--
Past Visit	-2.598*	1.564	-2.459*	1.591	-3.883	4.109	-0.704	2.294	-4.941**	2.490
Low Visit	-3.080**	1.697	-2.975**	1.715	-3.127	4.116	-1.282	2.635	-4.996**	2.873
High Visit	-5.139***	1.978	-4.888***	1.965	-5.760	4.614	-3.195	2.974	-8.889***	3.709
View	0.156	1.178	0.392	1.180	-0.454	2.614	1.959	1.888	-1.959	2.042
Hike	1.957**	1.187	1.804*	1.194	-2.212	2.564	3.391**	1.882	4.623**	2.072
Snow	2.074*	1.471	2.157*	1.447	1.771	3.000	2.578	2.300	-0.613	2.831
Water	0.665	1.260	0.541	1.257	-0.355	2.681	0.643	1.934	3.154*	2.190
Hunt	-2.616**	1.179	-2.655**	1.165	-3.015	2.391	-1.695	1.832	-3.011*	2.138
Native Animals	4.203***	1.350	4.002***	1.385	10.396***	2.998	2.533	2.036	5.006***	2.146
Alien Animals	-0.403	1.112	-0.482	1.104	-1.837	2.335	-0.017	1.724	0.117	2.000
Basic Services	-3.395**	1.685	-3.323**	1.720	-6.431**	3.507	0.774	2.584	-2.551	2.586
Major Services	-0.961	1.110	-1.072*	1.105	-1.856	2.386	-0.302	1.750	-0.208	1.963
Limit Vehicles	4.902***	1.198	5.120***	1.215	3.077**	2.363	6.378***	1.942	3.585**	1.957
Snowmobiles	-1.659	1.340	-1.731*	1.320	-3.958	3.117	-1.101	1.941	0.559	2.614
Air Pollution	4.532***	1.328	4.607***	1.326	4.984**	2.823	5.186***	1.969	4.458**	2.333
College	-0.174	1.198	-0.117	1.198	3.730*	2.462	-2.079	1.943	-1.145	2.147
Env. Group	1.565	1.841	1.734*	1.778	2.177	3.525	3.048	2.771	1.677	4.007
Young	1.450	1.706	1.477	1.711	7.084**	3.770	-0.357	2.661	1.383	2.956
Middle-age	0.123	1.520	-0.134	1.510	-1.346	2.907	0.240	2.415	3.354	2.714
Female	0.425	1.068	0.337	1.070	0.128	2.276	-0.452	1.839	0.324	1.833
Adult†	0.222	0.547	0.296	0.549	0.079	1.132	1.419**	0.883	-0.213	0.961
Child†	-0.747	0.518	-0.609	0.514	-0.615	1.141	-0.103	0.777	-2.836**	1.079
Lung	1.433	1.143	1.423	1.136	0.538	2.367	1.164	1.846	0.746	2.016
Hispanic†	4.282*	2.351	4.109*	2.337	1.130	5.312	2.151	3.544	8.459**	4.327
White†	0.421	1.523	0.460	1.528	-0.600	3.030	1.139	2.314	2.051	2.633
Medium Income	0.263	1.476	0.496	1.499	3.564*	2.765	2.578	2.029	-2.978	2.414
High Income	1.511	1.767	1.621	1.770	5.355*	3.266	2.234	2.365	-1.167	3.107
ENC†	-1.602	2.563	-1.631	2.529	-2.661	4.844	1.326	4.129	-2.426	4.696
WNC†	-2.767	2.648	-2.841	2.610	-2.820	4.934	0.301	4.229	-3.637	4.886
GP†	-1.581	2.934	-1.873	2.898	-3.658	5.518	0.369	4.710	0.663	5.476
SE†	-2.190	2.628	-2.150	2.585	-1.586	4.896	-0.302	4.085	-5.676	4.952
WSC†	-1.957	3.100	-2.037	3.058	4.956	6.574	-1.774	4.731	-4.530	5.742
SW†	-0.877	2.759	-0.814	2.723	-1.126	5.181	1.513	4.424	-2.955	5.116
RM†	-2.054	3.015	-2.289	2.992	-1.714	5.644	-0.070	4.949	0.353	5.571
PC†	-3.658	2.680	-3.748	2.633	1.492	5.119	-4.923	4.222	-4.076	5.143
Metro	-0.571	1.386	-0.832	1.382	-2.517	2.979	-0.816	2.201	1.996	2.491
Ozone	-0.448	1.318	-0.280	1.311	2.388	2.752	-0.342	2.150	-2.826	2.391

Table 7.3. WTP Estimates for Improved Air Quality in National Parks (continued)

Explanatory Variables††	Baseline Model		Heteroscedastic Model		Mode-Specific Models					
					Mail (M)		Phone (P)		Web (W)	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Variance: Constant	4.850***	0.148	5.234***	0.273	4.865***	0.313	4.698***	0.248	5.049***	0.252
Variance: Mail	--	--	-0.560*	0.365	--	--	--	--	--	--
Variance: Phone	--	--	-0.585**	0.336	--	--	--	--	--	--
Likelihood Ratio Tests										
Null Hypothesis	Statistic	P-val.	Statistic	P-val.	Statistic	P-val.	Statistic	P-val.	Statistic	P-val.
H_0: $\beta = 0$	154.15	0.000	156.86	0.000	64.73	0.000	61.74	0.000	55.04	0.000
H_0: $\beta_M = \beta_P = \beta_W$	--	--	--	--	--	--	60.38	0.000	--	--
Summary Statistics										
Sample Size	2288		2288		685		890		713	
Pseudo R^2	0.123		0.125		0.195		0.143		0.107	
% Correct y=1's	61.11%		61.04%		50.27%		60.42%		68.49%	
% Correct y=0's	69.10%		69.40%		76.32%		68.52%		68.09%	
% Correct overall	64.51%		64.60%		62.48%		63.37%		68.30%	
Mean WTP	$13.51		$13.48		$12.04		$15.16		$13.05	
Median WTP	$8.52		$8.52		$6.38		$8.87		$9.46	

Notes. (***), (**), and (*) refer to statistical significance at the 1, 5 and 10 percent levels.

†Indicates two-tailed significance tests.

††Although not explicitly listed as an explanatory variable, the bids are incorporated through the probabilities (equation 2). See Cameron and James (1987) for further details.

The estimation was carried out using the Constrained Maximum Likelihood (CML 2.0) package in Gauss version 8.0. The nonlinear optimization routine was Newton-Raphson with a convergence criterion of $1{\times}10^{-5}$ for the gradient of the coefficients. The estimates for "missing" dummy variables are not shown.

7.1.4. Results of Estimating Mode Effects on WTP: Additive Effects

In the baseline model the variables of primary interest for the mode test are *phone* and *mail*, with "web" as the reference category for these two variables. The coefficient for *phone* is positive and significant ($p < .01$) in the baseline model (3.67), and remains positive and significant ($p < .05$) in the heteroscedastic model (2.96).

Thus, controlling statistically for all the other explanatory variables (while also matching and weighting to a common demographic profile), the phone respondents state a WTP for clean air in national parks that is about $3 to $4 higher than the WTP for web respondents. This strong effect can be generalized confidently to the respective (matched) populations. The coefficient for *phone* is also much more positive than that for *mail*, although the difference between them is not significant ($p > .05$, calculated from the standard errors in Table 7.3). The coefficient for *mail* does not depart significantly from zero in either model, indicating virtually no net difference (after all controls) between web and mail estimates of WTP for clean air in national parks.

A likely interpretation of these findings is that social desirability bias leads phone respondents to assert a WTP value that is higher than they would actually pay if given the chance. Reducing pollution and supporting national parks both tend to be viewed as desirable, "good citizen" behaviors. The social aspects of an interview, even over the phone, may therefore elicit a higher stated WTP from a phone respondent than in other, more impersonal modes of survey administration. The web and mail versions of the survey were both self-administered, and the WTP estimates for these two modes are very similar.

These results are consistent with the survey literature on social desirability effects in interviews as compared to self-administered questionnaires. In light of that literature, our results suggest the conclusion that a WTP estimate derived from a probability-based Internet-panel survey is no less accurate than that obtained from a well-designed mail survey, and is probably more accurate than from a comparable telephone survey. Estimating WTP for an environmental improvement by using a panel-based Internet survey (or a mail survey) will produce a more conservative dollar value than using a phone survey, *ceteris paribus*. The "all else equal" qualifier is essential; failing to control adequately for differences across the modes in sampling frame and/or demographic characteristics could produce quite different findings.

We note that our models do not explicitly control for hypothetical bias. Incorporating hypothetical bias in the modeling would require information from each respondent about some equivalent real-market transaction ("revealed preference" data), against which to compare their hypothetical survey responses ("stated preference" data (see Aadland and Caplan, 2003; Aadland *et al.*, 2009; Aadland *et al.*, 2006; Boyle, 2003). Most WTP studies, this one included, have no such real-market transaction against which to compare. However, if degree of hypothetical bias is unrelated to survey mode, then the bias in estimating absolute WTP drops out when we look at differences in WTP between survey modes. Or, if the degree of hypothetical bias does vary by mode, then our estimated mode effects include the differences in hypothetical bias. Either way, our main conclusion (that phone respondents report higher WTP) is sound.

7.1.5. Results of Estimating Mode Effects on WTP: Heteroscedasticity

Our second model allows for mode-specific heteroscedasticity. We find that the assumption of constant error variance across modes in the baseline model is not valid. The estimates for both α_1 and α_2 in Table 7.3 (labeled *variance: mail* and *variance: phone*) are marginally significant and negative, indicating that all else equal, surveys completed on the web have greater model uncertainty than both mail and phone surveys.

Our data cannot definitively explain why there is greater model uncertainty associated with web surveys. One possibility is that the web panelists may provide off-the-cuff answers. Members of KN's web panel are assigned up to six surveys a month with an expectation that on average four surveys will be completed. The regularity with which KN panel members complete surveys could result in respondents not pausing to reflect on the answers they provide. Another possible explanation for this result is self-selection for completing the survey. Respondents to the phone and mail surveys may have been willing to complete the questionnaire in part because they have an interest in, and well formed opinions about, national parks or air quality or both. In contrast, web panelists completed the survey mainly to fulfill their agreement with KN; they may not have fully formed opinions about parks and/or air quality. Recall that the web panelists engage less in outdoor recreation and are less frequent park visitors than mail and phone respondents.

7.1.6. Results of Estimating Mode Effects on WTP: Interactions

A highly significant likelihood ratio test in Table 7.3 shows that the estimated coefficients are not all equal across modes. However, with very few exceptions the coefficients for any given variable have the same sign in all three modes, and also tend to be roughly similar in magnitude. For every explanatory variable, we tested each of the three pairwise mode comparisons (phone/web, mail/web, and phone/mail), for a total of 112 tests. Only three of these differences (*hike*, *native animal*, and *child)* are even marginally significant ($p < .05$), which is about what would be expected by chance alone. Only one of those (*hike)* involves a change in sign (positive in the phone and web surveys, but negative in the mail survey). We conclude that interactions between mode of survey administration and the explanatory variables in the model are of little substantive importance in determining WTP for clean air in national parks.

Accordingly, we focus the remainder of our discussion of Table 7.3 on the coefficients in the heteroscedastic model. Generally similar conclusions would follow from either the baseline model or any of the three mode-specific models.

7.1.7. Other Predictors of WTP

The other factors that significantly influence an individual's stated WTP for improved air quality in national parks are the number of visits to national parks, engaging in certain outdoor activities, opinions about some of the issues faced by the major parks, membership in an environmental group, and being Hispanic or Latino. Visiting a national a park in one's lifetime (*past visit*) and the number of visits in the past two years (*low visit* and *high visit*) have an increasingly significant and negative effect on WTP. Since our valuation scenario involved a fee per visit, the economically rational frequent visitor should have a lower WTP, and that is what we find.

Among the specific outdoor activities considered in the questionnaire, hiking or jogging for at least 30 continuous minutes (*hike*), snow sports (e.g., skiing, snowmobiling, sledding) (*snow*), and hunting or fishing (*hunt*) have marginally significant effects on WTP (all $p <.10$). Both the hiking/jogging and snow sports variables have positive effects overall while hunting/fishing has a negative effect on WTP. The negative coefficient for this variable is quite plausible: hunting is generally not allowed in national parks (except in Alaska), and fishing need not be a particularly strenuous physical activity. Hence hunters and anglers may value air quality in the national parks less than people with other recreational interests.

Agreement that air quality in some parks is getting more serious (*air pollution*) has a significantly positive effect on WTP, as does approval of limits on private vehicles (*limit vehicles*) and endorsing the reintroduction of native animals (*native animals*) (all $p < .01$). People who believe air pollution in the parks is a growing problem ought to be more willing to pay to reduce it, and they are. Those who approve of one way to address the problem – a limit on private vehicles – should be more willing to support another way – non-polluting park vehicles – and they are. Similarly, people who want to restore the parks' former wildlife species ought to endorse another way of mitigating human impact on the park environment, and they do.

Conversely, acceptance of jet-skiing and snowmobiling as allowable park activities is (marginally) negatively related to WTP ($p < .10$). Those who oppose a ban on gas-powered recreational equipment should be less supportive of electric-powered park vehicles, and they are. Agreeing that visitor facilities should be provided (*basic services* or *major services*, either possibly indicating a pro-development stance) has a negative effect on WTP ($p < .10$).

Membership in a local, state or national organization whose main purpose is to protect the environment (*env. group*) has a marginally positive effect on WTP ($p<0.10$). Not surprisingly, people who are interested in protecting the environment are more inclined to pay an additional fee to improve air quality in national parks. A second demographic variable to approach significance is being Hispanic or Latino; Hispanic respondents have a higher WTP than non-Hispanics ($p < .10$ for a two-tailed significance test).

7.2. Scope Test

A second valuation question was included in the survey to test respondents' sensitivity to scope. The first question asked about WTP for reducing invisible air pollution from *medium* to *low*. The second WTP question asked respondents to report their maximum WTP for a greater amount of the same good (i.e., pollution would be reduced from *high* to *low*). This sequencing of WTP questions allows us to perform a scope test (Arrow *et al., 1993*).

Of those who accepted the first bid amount, 96.3% stated an equal or greater WTP on the second question. This result supports the validity of our design. We suspect that most of those who did not state an equal or higher amount on the second question found the second question problematic, not the first. As noted earlier, an open-ended WTP question presents a greater cognitive challenge to the respondent than a dichotomous choice question. Indeed, in each mode the number of people who were unable or unwilling to state a dollar value in the open-ended format is about triple the number of missing cases on the dichotomous question (see Appendix A). This strongly suggests that many respondents found the open-ended question too challenging.

Also, some respondents may have misinterpreted the second question as additive: Given the fee mentioned in the first question, how much more than that would one pay for an even greater reduction? Misunderstanding the second question as additive with the first could well lead to replying with an amount less than the first bid price. We were alerted to this possible confusion by the focus groups, and we did our best to forestall it with a revised wording of the second question, but we may not have succeeded completely.

We therefore view the results of the scope test as conservative with regard to the validity of the first WTP question. Nevertheless, as noted above, we excluded from the econometric analysis of that dichotomous item the respondents who said "Yes" to the bid for the medium improvement in air quality but then stated a lower WTP for the larger improvement on the open-ended item.

7.3. Tests for Survey Fatigue/Panel Conditioning

A separate analysis was completed for survey fatigue and/or panel conditioning. An additional dummy variable was constructed for each of the three modes. The variable is based on respondents' answers to our meta-questions about their survey behaviors. In the case of web panelists we obtained from KN each respondent's tenure (in weeks) on the panel. A description of the *survey experience* dummy variable for each mode is presented in Table 7.4. Estimation results are shown in Table 7.5.

We find no compelling evidence for an effect on WTP from survey fatigue (*survey experience*) for mail or phone respondents. The point estimate for the effect of survey experience is modestly positive for phone respondents, and substantially negative for mail respondents, but neither comes close to statistical significance at any conventional level of confidence.

Table 7.4. Definition of *Survey Experience*, by Mode

Mode	Description
Mail	Completed 4 or more mail surveys in the past 12 months
Phone	Completed 4 or more phone surveys in the past 12 months
Web	Completed 52 or more web surveys in the past 12 months OR completed more than 52 weeks on the web panel

The number of web surveys completed in the past year or the number of weeks serving on the KN web-panel (*survey experience*) has a marginally significant effect on stated WTP for better air quality in national parks ($p < .10$). Web panelist who had completed at least one survey a week over the past year or had web tenure of at least a year had a somewhat lower WTP.

In broad outline, the other factors that influence WTP for cleaner air in national parks, with the additional control for survey experience, are similar to those presented above.

Table 7.5. Estimation Results for Survey Experience

Explanatory Variables	Mode-Specific Baseline Models					
	Mail (M)		Phone (P)		Web (W)	
	Coef.	SE	Coef.	SE	Coef.	SE
Constant	11.248**	6.736	0.1334	6.2507	13.964**	6.933
Past Visit	-3.980	4.102	-0.7295	2.2896	-4.917**	2.479
Low Visit	-3.122	4.105	-1.3935	2.6339	-5.092**	2.864
High Visit	-5.769	4.598	-3.43	2.9817	-9.101***	3.697
View	-0.574	2.609	1.9306	1.8897	-1.924	2.035
Hike	-2.096	2.551	3.3532**	1.8778	4.519**	2.064
Snow	1.811	2.989	2.7119	2.3038	-0.353	2.820
Water	-0.428	2.670	0.8017	1.9356	3.006*	2.181
Hunt	-3.020	2.379	-1.7361	1.8302	-3.071*	2.133
Native Animals	10.485***	2.989	2.6909*	2.0441	4.836**	2.138
Alien Animals	-1.913	2.349	0.0249	1.7229	0.045	1.995
Basic Services	-6.257**	3.490	0.9407	2.5809	-2.210	2.588
Major Services	-1.938	2.379	-0.2335	1.7518	-0.176	1.956
Limit Vehicles	3.152*	2.354	6.4221***	1.9415	3.737**	1.956
Snowmobiles	-3.690	3.112	-1.1318	1.9371	0.491	2.604
Air Pollution	4.847**	2.814	5.0367***	1.9698	4.326**	2.327
College	3.560*	2.453	-2.0843	1.9407	-1.144	2.142
Env. Group	2.585	3.546	2.9752	2.7633	1.437	3.984
Young	6.915**	3.754	-0.4049	2.6546	1.066	2.948
Middle-age	-1.518	2.900	0.0547	2.4185	3.213	2.727
Female	0.132	2.269	-0.4964	1.8384	0.406	1.826
Adult	0.010	1.129	1.3415*	0.8861	-0.131	0.960
Child	-0.598	1.136	-0.0923	0.7779	-2.877***	1.077
Lung	0.624	2.361	1.2208	1.8479	0.530	2.013
Hispanic	1.156	5.278	2.2259	3.5366	7.809**	4.329
White	-0.488	3.023	0.9708	2.3162	2.575	2.658
Medium Income	3.505	2.753	2.4345	2.0305	-2.873	2.406
High Income	5.366**	3.249	2.2402	2.3648	-1.229	3.099
ENC	-2.326	4.833	1.4862	4.1408	-2.169	4.707
WNC	-2.545	4.919	0.1804	4.2364	-3.486	4.886
GP	-3.461	5.490	0.6236	4.7197	0.774	5.470
SE	-1.450	4.872	-0.3155	4.092	-5.357	4.951
WSC	5.007	6.545	-1.5234	4.7413	-4.553	5.740
SW	-1.018	5.157	1.5384	4.4361	-2.986	5.113
RM	-1.626	5.612	0.042	4.9586	0.313	5.561
PC	1.596	5.100	-4.7926	4.2257	-3.919	5.135
Metro	-2.338	2.973	-0.7084	2.2037	2.021	2.479
Ozone	2.341	2.743	-0.3431	2.147	-2.862	2.385
Survey Experience	-12.554	12.885	3.8643	3.9791	-3.082*	2.379

Notes. (***), (**), and (*) refer to statistical significance at the 1, 5 and 10 percent levels.

8. Cost Comparisons

Precise cost comparisons across the three modes are not possible, but reasonable approximations can be developed. For precision, the costs of questionnaire development and pre-testing should be eliminated from the comparison, because development efforts for any of the modes also benefitted the others. However, pre-testing is a standard part of KN's services, and its cost was therefore folded into the total price on the KN invoice. Similarly, data cleaning, data analysis, and reporting ought to be excluded from the cost comparisons, because most of that work gets applied to a merged data file for all three modes. But KN's standard services include the delivery of a data file and a field report covering the web panel only (see Appendix F) and again, the cost of that was not itemized in the bill.

Some of WYSAC's costs are also difficult to estimate. For example, three of WYSAC's lead researchers on this project are University of Wyoming faculty members, with a research obligation as part of their job descriptions. Some of the time they spent on this project was therefore contributed labor (implicit cost-sharing); the cost of that time was included in their standard university paychecks, not billed to the project. These faculty members were mainly involved in design and analysis, rather than in data collection activities as such, but their labor on that phase of the research was not entirely negligible.

In addition, other WYSAC researchers are salaried employees who often put in uncompensated overtime. As a result, the "billable hours" for salaried staff that WYSAC tracks with care, project by project, may understate the true costs of any given project. Fortunately for present purposes, the great majority of the cost for data collection in a mail or phone survey covers materials, outsourced services, and hourly wages rather than salaries (e.g., purchasing envelopes, printing questionnaires, and paying mail handlers and telephone interviewers). Costs such as those can be attributed much more precisely to a particular project.

Even with these caveats, it seems clear that the cost of data collection alone (excluding survey development, pre-testing, data analysis, and reporting) was highest by mail and lowest by phone. The web costs were closer to the high end (mail) than the low; see Table 8.1. The calculations supporting these conclusions follow.

Formatting the mail questionnaire, materials, printing, postage, clerical labor, and supervision for the mail survey (plus all costs of telephone reminders to non-respondents, but excluding the focus groups) totaled about $38,500, and generated 904 completed questionnaires. This amounts to a cost per completion of $42.59. Programming the phone questionnaire, long-distance calling, interviewer labor, and supervision for the phone survey (plus all costs of the thank-you mailing to identify deliverable addresses, but excluding the phone pre-test) totaled about $33,500 for 1273 completed interview. This equates to $26.32 per completion.

Table 8.1. Data Collection Costs, by Mode

	Phone	Web*	Mail
Approximate Cost	$33,500	$49,200	$38,500
Completions	1273	1268	904
Cost per Completion	$26.32	$38.80	$42.59

*Including the web pre-test.

The KN bill for the web survey (including not only a thank-you mailing but also the web pre-test) was $49,198. To make this figure roughly comparable to those for the phone and mail surveys (which do not include the related pre-testing), we divide by the combined completions in both the main web survey and the web pre-test (1162 + 106 = 1268). This yields an approximate cost per completion for the web survey of $38.80. We have not further adjusted that figure to remove KN's costs for producing the field report, costs that are unknown but likely to be relatively small.

9. Conclusions

The goal of this research was to assist the U.S. Environmental Protection Agency (EPA) in assessing different modes of administering a questionnaire to estimate the public's willingness to pay (WTP) for environmental quality. A relatively new method for eliciting WTP is the use of standing panels of respondents for surveys administered via the Internet. In particular, Knowledge Networks (KN) recruits its panelists through probability-based sampling. KN web surveys therefore offer a promising alternative to other, more traditional survey modes.

The present study compared phone and mail surveys to a web survey of KN panelists. Each sample was administered a questionnaire on the public's valuation of improved air quality in national parks. By design, the questionnaire was nearly identical for all three modes, as was the sampling frame.

The response rate was much lower for the web survey than by phone or by mail. Response was best in the mail survey, which also showed the greatest yield from additional survey efforts aimed at encouraging response. Exploratory analyses gave indications of differential non-response bias by mode, apparently due to mode-related variation in the mechanisms of self-selection as a survey participant. Phone and mail surveys may involve more self-selection of respondents interested in the topic of a particular survey, whereas a web panel may self-select for those with sedentary lifestyles.

Weighting and matching the respondents did not eliminate significant demographic, behavioral, and attitudinal differences across modes. For example, web respondents were less likely than either phone or mail respondents to have children in the home, to be members of any environmental organizations, or to participate in various kinds of outdoor recreation. Our econometric models therefore incorporated statistical controls for variables likely to be correlated with both mode of survey administration and WTP.

Results showed that using either a panel-based Internet survey or a mail survey produces a more conservative dollar value for WTP than using a phone survey. Communication with a live interviewer over the phone seems to yield over-statement of true WTP. Though face-to-face interviewing was not part of our research design, the apparent upward bias on WTP due to the effects of social desirability in a phone survey would also be expected in a face-to-face survey.

We found, further, that the variance in WTP left unexplained by our model was higher for the web panel than for either of the other two survey modes. There was a slight negative effect on WTP from duration of panel membership or from number of web surveys completed, whereas survey fatigue had no consistent effect in the other two modes. Statistical interactions between mode of survey administration and other explanatory variables were of little substantive importance. We conclude that, with appropriate controls, a WTP estimate derived from a KN web survey should be no less accurate than that obtained from a well-designed and well-executed mail or phone survey.

The cost of data collection proved to be highest by mail and lowest by phone. The web survey was in the middle on cost, but closer to the high end than the low. Any of these three modes would be much cheaper than face-to-face interviewing if the goal is to obtain a large representative sample.

Interest in web panels as an alternative mode of data collection has grown in part as a reaction to recent trends such as declining response rates to phone surveys, the increase in cell-only households, and other cultural and technological changes in telephone use in the U.S. However, it should be noted that KN's recruitment process is itself reliant on a fairly traditional telephone survey approach. In that sense, a KN web panel survey does not solve the methodological problems associated with changes in telephony, even if it makes them less conspicuous.

For example, the low cumulative response rate in our web survey (well below 10%), with the attendant risk of non-response bias, results mainly from KN's low response rate during telephone recruitment of panelists. To keep pace with cultural and technological changes in telephony, KN will need to adopt the same kinds of tactics for recruitment that all telephone surveyors are now using to improve the representativeness of their samples. These tactics include dual-frame sampling (cell and landline), multi-mode initial contacts (mail, phone, and email), non-contingent incentives, and multi-mode follow-up with non-respondents. But such tactics can be expensive, and their use could well close the fairly narrow gap between the cost of a KN web panel survey and that of a thoroughly designed and implemented mail survey.

Meanwhile, advances in address-based sampling are improving the potential for representativeness of mail surveys. The techniques for maximizing response and minimizing non-response bias in a mail survey are also being refined through systematic research (e.g., Dillman 2007). Like a web survey, a mail survey is well-suited to the use of photos or other visual aids that may be especially helpful when asking about environmental quality. And unlike phone-recruited web panels, mail surveys are not much affected by cultural and technological changes in telephone use.

Thus, mail surveys and probability-based web surveys both merit consideration as cost-effective alternatives to phone or face-to-face interviewing in studies of willingness to pay for environmental quality.

10. References

Aadland, D., B. Anatchkova, B. D. Grandjean, J. Shogren and P. A. Taylor. (2006). *Analysis to assist with pricing the new recreation pass: Final project summary*. Laramie: University of Wyoming.

Aadland, D., B. Anatchkova, B. D. Grandjean, J. Shogren & P. A. Taylor. (2009). Valuing access to US public lands: A unique pricing experiment. Under review.

Aadland, D. & A. J. Caplan. (2003). Willingness to pay for curbside recycling with detection and mitigation of hypothetical bias. *American Journal of Agricultural Economics*, 85, 492–502.

Adamowicz, W. (2004). What's it worth? An examination of historical trends and future directions in environmental valuation. *Australian Journal of Agricultural and Resource Economics*, 48, 419-443.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons, Inc.

AIRNow. (n.d.). *Air quality index, guide to air quality and your health*. Retrieved June 8, 2007, from http://airnow.gov/index.cfm?action=aqibroch.aqi#aqioz.

Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso Books.

Andrews, D., B. Nonnecke, & J. Precce. (2003). Conducting research on the Internet: Online survey design, development and implementation guidelines. *International Journal of Human-Computer Interaction*, 16, 185-210.

Arrow, K., & R. Solow (1993). Report of the NOAA Panel on contingent valuation. [S.I.]: National Oceanic and Atmospheric Administration.

Banzahf, S. H., D. Burtraw, D. Evans and A. Krupnick. (2006) Valuation of natural resource improvements in the Adirondacks. *Land Economics*, 82, 445-464.

Bateman, I. J. & K. G. Willis (Eds). (1999). *Valuing environmental preferences: Theory and practice of the contingent valuation method in the US, EU, and developing countries*. Oxford: Oxford University Press.

Berrens, R., A. Bohara, H. Jenkins-Smith, C. Silva, & D. Wiemer. (2003). Advent of Internet surveys for political research: a comparison of telephone and Internet samples. *Political Analyst*, 11, 1-23.

Boyle, K. (2003). Contingent valuation in practice. In P. Champ, K. Boyle & Tom Brown (Eds.), *A primer on nonmarket valuation.* New York: Kluwer Academic Publishers.

Cameron, T. A & M. D. James. (1987). Efficient estimation methods for "close-ended" contingent valuation surveys. *The Review of Economics and Statistics*, 69, 269-276.

Carson, R. T., N. E. Flores & N. F. Meade. (2001). Contingent valuation: Controversies and evidence. *Environmental and Resource Economics*, 19, 173-210.

Carson, R. T., T. Groves & M. J. Machina. (2000). *Incentive and informational properties of preference questions*. Unpublished manuscript.

Chatt, C. & J. M. Dennis. (2003). Data collection mode effects controlling for sample origins in a panel survey: Telephone versus Internet. Paper presented at the 2003 Annual Meeting of the Midwest Chapter of the American Association for Public Opinion Research, Chicago, Illinois.

Combs, J., R. Kirkpatrick, J. Shogren, & J. Herriges. (1993). Matching grants and public goods: A closed-ended contingent valuation experiment. *Public Finance Quarterly*, 21, 178-195.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-481.

Crocker, T. D., D. S. Brookshsire, R. C. d'Arge, S. Ben-David, A. V. Kneese, & W. D. Schulze. (1979). *Methods development for assessing air pollution control benefits*. EPA Document No. EPA-600/5-79-001a. Washington, DC: US Environmental Protection Agency.

Cummings, R. G., S. Elliot, G. W. Harrison, & J. Murphy. (1997). Are hypothetical referenda incentive compatible? *Journal of Political Economy*, 105, 609-621.

Cummings, R. G., G. W. Harrison, & E. E. Rutstrom. (1995). Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive compatible? *American Economic Review*, 85, 260-266.

Dennis, J. M. (2001). Are Internet panels creating professional respondents? *Marketing Research,* 13, 35-38.

Diamond, P. A., & J. A. Hausman. (1994). Contingent valuation: Is some number better than none? *Journal of Economic Perspectives*, 8, 45-64.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method.* (2nd ed., with 2007 update). New York: John Wiley & Sons.

Dillman, D. A. & D. K. Bowker. (2001). The web questionnaire challenge to survey methodologists. In U. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science*. Lengerich, Georgia: Pabst Science Publishers.

Dillman, D. A., G. Phelps, R. D. Tortora, K. Swift, J. Kohrell, and J. Berck. (2001). *Response rate and measurement differences in mixed mode surveys using mail, telephone, interactive voice response and the Internet*. Retrieved November 13, 2006 from http://www.sesrc.wsu.edu/dillman/papers/Mixed%20Mode%20ppr%20_with%20Gallup_%20POQ.pdf.

Dillman, D. A., R. Tortora, J. Conradt, & D. Bowker. (1998). Influence of plain vs. fancy design and response rates for web surveys. Paper presented at the joint statistical meetings of the American Statistical Association, Dallas, TX.

Dorofeev, S., & P. Grant. (2006) *Statistics for real-life sample surveys; Non-simple-random samples and weighted data* . Cambridge, UK: Cambridge University Press.

Gordon, R. (2005). *Pondering panels*. Retrieved November 16, 2006, from http://www.naa.org/Resources/Articles/Digital-Edge-Pondering-Panels/Digital-Edge-Pondering-Panels.aspx

Grandjean, B. D., M. Leighty, P. A. Taylor, & Y. Xu. (2005). Is target selection by last birthday 'random enough'? A split-ballot test. *Proceedings of the Joint Statistical Meetings,* American Statistical Association.

Hanemann, W., J. Loomis & B. Kanninen. (1991). Statistical efficiency of double bounded dichotomous choice contingent valuation. *American Journal of Agricultural Economics*, 73, 1255-1263.

Harrison G. W. & E. E. Rutström. (2006). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. Plott and V. L. Smith (Eds.), *Handbook of results in experimental economics*. New York: Elsevier Science.

Herriges J. & J. Shogren. (1996). Starting point bias in dichotomous choice valuation with follow-up questioning. *Journal of Environmental Economics and Management*, 30, 112-131.

Hill, L. B. (2000). *Out of sight: Haze in our national parks. How power plants cost billions in visitor enjoyment.* Retrieved November 16, 2006 from http://www.catf.us/publications/reports/Out_of_Sight.pdf.

Hill, L. B., W. Harper, J. M. Halstead, T. H. Stevens, I. Porras, & K. D. Kimball. (2000). Visitor perceptions and valuation of visibility in the Great Gulf Wilderness, New Hampshire. *USDA Forest Service Proceedings*, RMRS-P-15-VOL-5, 304-310.

Huggins, V. & J. Eyerman. (2001). Probability based Internet surveys: A synopsis of early methods and survey research results. Paper presented at the Research Conference of the Federal Committee on Statistical Methodology, November 14-16.

Kehoe, C. M. and J. E. Pitkow. (1996). Surveying the territory: GVU's five www user surveys. *The Worldwide Web Journal*, 1, 77-84.

National Telecommunications and Information Administration [NTIA]. (2004). *A nation online: Entering the broadband age*. Washington, DC: U.S. Department of Commerce. Retrieved November 19, 2006 from http://www.ntia.doc.gov/reports/anol2004/NationOnlineBroadband04.pdf

Swoboda, S. J., N. Muehlberger, No., R. Weikunat, & S. Schneeweiss. (1997). Web-based surveys by direct mailing: An innovative way of collecting data. *Social Science Computer Review*, 15, 242-255.

Smith, K. & L. Osborne. (1996). Do contingent valuation estimations pass a 'scope' test? A meta analysis. *Journal of Environmental Economics And Management*, 31, 287-301.

Taylor, P. A. & B. D. Grandjean. (2009). Visitor satisfaction and support for park fees: Front country and back country experience at Rocky Mountain National Park. *The George Wright Forum* (forthcoming).

Taylor, P. A., B. D. Grandjean, & B. Anatchkova. (2006). *Analysis to Assist with Pricing the New Recreation Pass: Task #5: The Survey* (WYSAC Technical Report No. SRC-615). Laramie: Wyoming Survey & Analysis Center, University of Wyoming.

U.S. Environmental Protection Agency [EPA]. (2006). *2006-2011 EPA Strategic Plan: Charting Our Course* (pre-publication copy). EPA Document No. EPA-190-R-06-001. Washington, DC: US Environmental Protection Agency. Retrieved November 20, 2006, from http://www.epa.gov/ocfo/plan/2006/entire_report.pdf.

U.S. Environmental Protection Agency [EPA]. (2008). National Ambient Air Quality Standards for Ozone; Final Rule. *Federal Register* , 40 CFR Parts 50 and 58. Washington, DC: National Archives and Records Administration.

U.S. Environmental Protection Agency [EPA]. (n.d.a) *Ozone good up high bad bearby*. Retrieved April 20, 2008, from http://www.epa.gov/air/oaqps/gooduphigh/bad.html#7.

U.S. Environmental Protection Agency [EPA]. (n.d.b). *Fact sheet final fevisions to the national ambient air quality standards for ozone*. Retrieved April 20, 2008, from http://www.epa.gov/air/ozonepollution/pdfs/2008_03_factsheet.pdf.

U.S. Environmental Protection Agency [EPA]. (n.d.c). *Revisions to EPA's ozone air quality index fact sheet*. Retrieved April 20, 2008, from http://www.epa.gov/air/ozonepollution/pdfs/2008_03_aqi_changes.pdf.

U.S. National Park Service [NPS]. (n.d.a). *NPS Units within counties that exceed EPA standards for ozone concentration*. Retrieved June 12, 2007, from http://www.nature.nps.gov/air/Maps/AirAtlas/gisprojects.cfm

U.S. National Park Service [NPS]. (n.d.b). *Human health effects of air pollution*. Retrieved June 12, 2007, from http://www.nature.nps.gov/air/AQBasics/human.cfm.

Wilson, Edward O. (2000). *Natural resource stewardship*. Keynote address to Discovery 2000, St. Louis, Missouri, September 12, 2000.

Yun, G. W. & C. W. Trumbo. (2000). Comparative response to a survey executed by post, e-mail, and web form. *Journal of Computer Mediated Communication*, 6.

11. Appendices

The following appendices are provided as separate documents:

Appendix A. Phone, Web, and Mail Frequencies, with Phone Script

Appendix B. Open-ended Responses by Mode on Q4 (Why Visit More)

Appendix C. Open-ended Responses by Mode on Q9 (WTP Explanation)

Appendix D. Open-ended Responses by Mode on Q26 (Final Comment)

Appendix E. Web Survey Correspondence, with Screenshots

Appendix F. Web Survey Field Report from Knowledge Networks

Appendix G. Mail Survey Correspondence, with Questionnaire