Comparing an Internet Panel Survey to Mail and Phone Surveys on Willingness to Pay for Environmental Quality: A National Mode Test

Burke D. Grandjean[1], Nanette M. Nelson[1], and Patricia A. Taylor[1]
[1]Wyoming Survey & Analysis Center, University of Wyoming, Dept. 3925,
1000 E. University Ave., Laramie, WY 82071

Abstract
Recently, the U.S. Environmental Protection Agency sought to determine whether a recruited Internet panel of respondents can be used to produce reliable estimates of willingness to pay for goods or services for which no formal market exists. Here we report the results of a national mode test comparing results from an Internet panel survey on air quality in national parks with the same questionnaire administered by mail and by phone. Sampling frame was held nearly constant by using national Random Digit Dialing samples with reverse-lookup of addresses for all three surveys. Weighting and matching the respondents did not eliminate significant demographic, behavioral, and attitudinal differences across modes. Results indicate a willingness to pay estimate derived from a probability-based Internet-panel survey is likely to be as accurate as that obtained from a well-designed mail survey.

Key Words: Internet panel survey, contingent valuation, willingness to pay, air pollution, ozone, national park

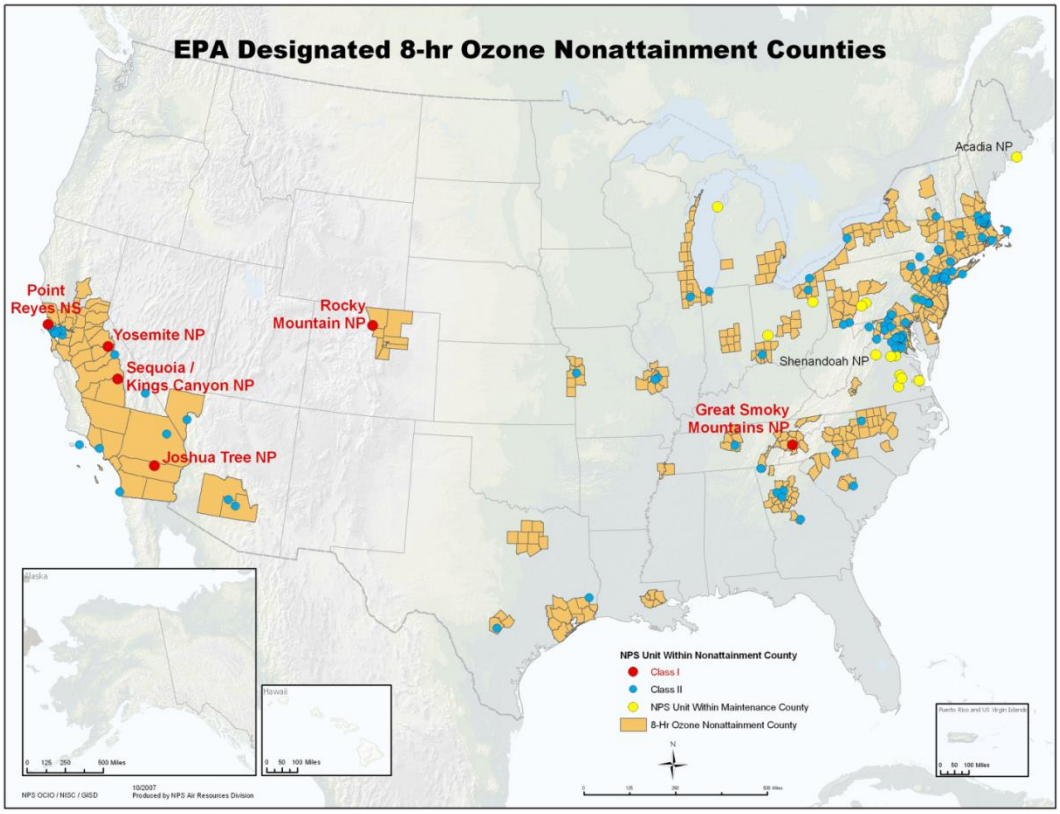
1. Introduction

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. Our aim here is to examine differences between Internet panel surveys and studies conducted by two other modes of survey administration, telephone, and mail. Our main focus is on mode effects *per se* (differences attributable to the medium through which the respondent is questioned). A full technical report of the study is available on the EPA website: http://yosemite.epa.gov/EE/Epa/eerm.nsf/vwRepNumLookup/EE-0519?OpenDocument

2. The Policy Context

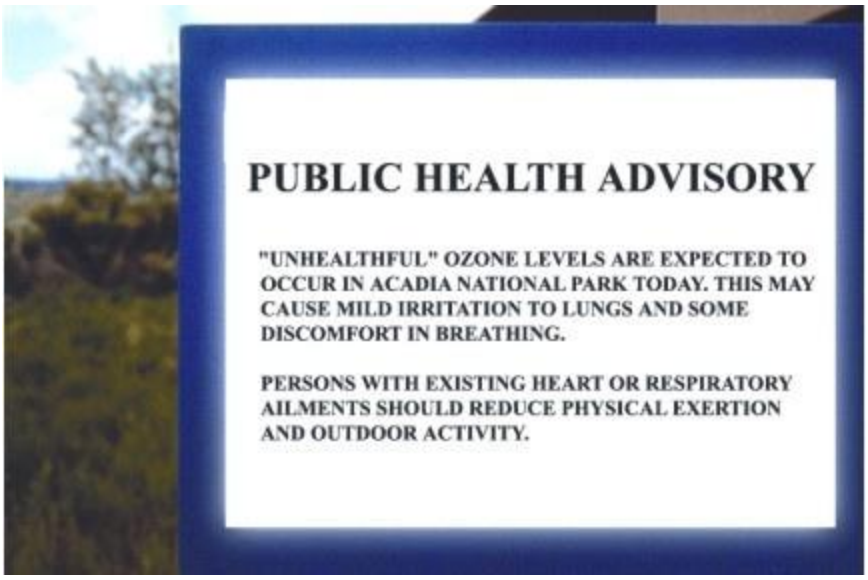
We apply our study objective (the comparison of survey modes) to the problem of deteriorating air quality in natural areas. Specifically, we address health concerns regarding increasing *ground-level ozone concentrations in national parks* in the Southeast, the Northeast, and the Pacific Coast. Figure 1 maps the parks that are experiencing with some frequency a level of ozone concentration in excess of the standards set by the EPA to protect human health (NPS, n.d.a). Indeed, as illustrated in Figure 2, concern for the health and safety of visitors and employees has led the National Park Service (NPS) to adopt an ozone advisory system in several parks (NPS, n.d.b). Moreover, air quality in national parks has been shown to have positive economic value. 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).



Source: NPS, n.d.a.

Figure 1: NPS Units within Counties that Exceed EPA Standards for Ozone

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 forests (EPA, n.d.a).



Source: NPS, n.d.b.

Figure 2: Health Warning for Ozone Posted at Acadia National Park

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 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.

3. Web Surveys and Web Panels

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, Nonnecke, & Precce, 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, Bohara, Jenkins-Smith, Silva & Wiemer, 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 chosen at random, perhaps specified so as to include pre-determined proportions in various demographic subgroups (e.g., half men and half women). 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. 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 example "household recruitment rate" of about 33% and even lower response rates (Callegaro & DiSogra, 2008).

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. 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, an email account, and ongoing technical support (Berrens, *et al*., 2003). Ignoring non-response issues and 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).

4. Research Design

4.1 Sampling Frame and Coverage

For the present project, KN's probability-based approach to recruiting a web panel was 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 and those with 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).

4.2 Mode Effects and Unimodal Questionnaire Design

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 on a self-administered questionnaire may also generate more forthright responses for 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 sought 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.

4.3 The Contingent Valuation Scenario

We used contingent valuation (CV) to assess how much the public is willing to pay for improved air quality in national parks. 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). These methods 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). By 2000, Carson, *et al*. had identified over 1,600 CV-style studies, and its use continues to grow worldwide. Furthermore, CV has been sanctioned for use in government decision-making and in the courts.

For the present application, we relied on the EPA's Air Quality Index 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 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. These verbal statements are crucial for our purposes, because they allow us to describe air pollution equivalently in all three modes, without the use of visual aids.

Table 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 prolonged 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)

However, 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" in our valuation scenario. Feedback from two focus groups validated the phrasing.

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. Table 2 provides a matching of our questionnaire descriptions to EPA's AQI and air quality categories.

Table 2: EPA Air Quality Categories vs. Levels of Air Pollution in the Questionnaire

EPA Air Quality	AQI	Questionnaire Pollution Level	Questionnaire Description of Health Concerns
Good to Moderate	0-100	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."
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 to Very Unhealthy	151-300	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."

4.4 The Survey Instrument

In developing our questionnaire to elicit WTP for improved air quality in national parks, we applied guidelines set forth in the Report on the NOAA Panel on Contingent Valuation (Arrow, *et al.*, 1993). Focus group sessions and pre-testing helped to clarify the scenario, refine the bid vector, and improve the credibility of the bid vehicle.

The questionnaire can be considered in three parts. 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 national parks.

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, 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. Respondents were reminded just prior to the fee question that they should consider their income and other costs of visiting a park when deciding how much they could really afford.

We elicited WTP with a "referendum" format. In this 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, referendum format (also known as "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 *bid vector* (the range of prices used in the valuation scenario) was finalized by conducting focus groups and two nation-wide pretests. After trying bids from $2 to $40 in the pretests, the final bid vector was set at $2, $5, $10, $15, and $25.

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 are used to identify protest bids. 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 something external to the valuation exercise (e.g., the government is wasteful), or because the hypothetical market is not credible. The few protest bidders so identified were eliminated from our econometric analyses.

A second 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 provided 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. Results confirmed the validity of our valuation scenario (see full report).

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 (addressed in our full report).

4.5 Administering the Three Surveys

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 82% of all U.S. households at the time of the surveys (spring 2008) (Blumberg & Luke, 2009). 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).

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. 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, Leighty, Taylor, & Xu, 2005).

4.6 Matching, Weighting, and Benchmarking

To hold sampling frame constant across all three modes, all of our analysis presented in Table 5 (below) are limited to 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. All respondents were weighted to a common set of marginal distributions on gender, age, racial/ethnic identification, education, region, and metropolitan residence, using Census benchmarks.

5. Results

5.1 Survey Effort

To compare the level of survey effort across modes, Table 3 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 not only indispensible but also 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.

Table 3: Number of Completions, by Survey Effort and Mode

Survey Effort	Phone	n	Web	n	Mail	n
Initial	1st call	392 (392)	Panel recruitment, Setup, Profiling	--	Contact letter	-- (7)
Minimal	1 – 4 callbacks	632 (1023)	Email invitation	425 (425)	Cover letter + survey, Reminder postcard	586 (593)
Ordinary	5 – 9 callbacks	111 (1134)	Generic email reminder	630 (1055)	2nd cover letter + survey	156 (749)
Concerted	10+ callbacks	138 (1273)	Customized email, Automated phone call	107 (1162)	Phone reminder, Priority mailing	155 (904)

Totals in parentheses are cumulative; grand total across all three modes = (3339). 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 "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.

With the "ordinary" (and cheap) effort of a short and generic 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 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 an additional, customized email reminder plus 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.

5.2 Rates of Survey Participation

We used the formula that the American Association for Public Opinion Research ([AAPOR], 2006) 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.

KN calculates RR3 at the recruitment stage, then computes the post-recruitment profiling rate, the study-specific completion rate, and the panel retention rate. What KN calls the Cumulative RR1 is the product of the recruitment, profiling, and completion rates. 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.

As shown in Table 4, the mail survey achieved the best response rate, at 30%, followed by the phone survey at 16% and the web survey at a strikingly low 4%. A response rate of less than 10% (or even 30% as in our mail survey) raises 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.

Table 4: Participation Rates by Mode

		Phone	Web	Mail
Completion Rate	After respondent selection	91%	77%	75%
Response Rate	Unweighted	16% RR3	4% CUMRR2	30% RR3

5.3 Statistical Tests

Mantel-Haenszel chi-square tests for linear trend (comparing the modes two at a time) are reported in the following subsections. The Mantel-Haenszel test has greater statistical power than the Pearson chi-square 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 Mantel-Haenszel 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). In Table 5, we only report differences that are significant at the .01 level in a two-tailed test. In all of these comparisons except one, web panelists tend to score lower than other respondents; the exception is marked (+).

Table 5: Significant Mode Differences for Matched Subsamples ($p < .01$)

	Web vs. Both Phone & Mail	Web vs. Phone Only	Web vs. Mail Only
Demographics	Membership Children		Adults Breathing problem Income
Survey-taking Behaviors	Web surveys (+)	Suggestions Phone surveys	
Recreational Behaviors	View nature Hike or jog Snow sports Water activities Hunt or fish		Ever visited Recent visits Planned visits
Opinions	Satisfied Basic facilities	Willing to pay	Restore wildlife Major facilities

5.4 Demographic Differences

The demographic variables that show the most significant ordinal associations in the two-mode Mantel-Haenszel 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 also have lower income than the other two samples.

5.5 Differences in Survey-taking Behavior

Our survey also generated information about several different survey-taking behaviors. 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. Thus, when asked what NPS could do to encourage more park visitation (Suggestions), almost 90% of phone respondents offered some kind of suggestion, compared to 80% on the web and only 60% by mail. Also, phone respondents reported significantly more frequent participation in phone surveys than web respondents, and of course the web panelists reported completing the most web surveys.

5.6 Differences in Recreational Behavior

As summarized in Table 5, there are clear and consistent differences on all of the items in the first part of the questionnaire concerning park visitation and outdoor activities. The mail respondents are notably more likely to visit a national park site than 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.

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 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, and 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 that are sedentary. To a notable degree, the KN panelists are web surfers, not wave surfers or snowboarders. Hence outdoor enthusiasts could be under-represented among the web respondents.

5.7 Attitudinal 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. Compared to web panelists, the phone respondents are notably more likely to accept whatever bid they received on the willingness to pay question.

5.8 Differences in WTP Estimates

Using a maximum likelihood procedure and controlling statistically for all other explanatory variables (while also matching and weighting to a common demographic profile), the phone respondents state a willingness to pay for cleaner 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. We found virtually no net difference (after all controls) between web and mail estimates of WTP for cleaner air (see full report).

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.

6. Conclusions

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. 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. 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 survey.

Acknowledgements

This study was supported by the U.S. Environmental Protection Agency, under Cooperative Agreement # USEPA 83359101. The present report is entirely the responsibility of the authors, and does not represent an official position of the EPA. We thank our Project Officer at the EPA, Alan Carlin, for his assistance throughout the study, and our colleagues at the University of Wyoming for their input, including especially David Aadland, Bistra Anatchkova, Michael Dorssom, and Brian Harnisch.

References

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.

American Association for Public Opinion Research. (2006). Standard definitions: Final dispositions of case codes and outcome rates for surveys. Retrieved June 18, 2007, from www.aapor.org/uploads/standarddefs_4.pdf.

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

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., & Schuman, H. (1993). "Report of the NOAA Panel on contingent valuation." *Federal Register*, 58:10 (Jan. 15, 1993), 4601-4614.

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

Blumberg, S. J. & Luke, J. V. (2009, May 6). Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2008. National Center for Health Statistics. Retrieved May 15, 2009 from http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless200905.pdf

Callegaro, M. & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, 72, 1008-1032.

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

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

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

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

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

Herriges J. & Shogren, J. (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.

Huggins, V. & Eyerman, J. (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. & Pitkow, J. E. (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

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

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*, 70:60 (March 27, 2008), 16436-16514.

U.S. Environmental Protection Agency [EPA]. (n.d.a) *Ozone good up high bad nearby*. 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 revisions 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.