Morbidity and Mortality: How Do We Value the Risk of Illness and Death?

PROCEEDINGS OF SESSION III: PANEL DISCUSSION ON THE USE OF THE INTERNET IN VALUATION SURVEYS

A WORKSHOP SPONSORED BY THE U.S. ENVIRONMENTAL PROTECTION AGENCY'S NATIONAL CENTER FOR ENVIRONMENTAL ECONOMICS AND NATIONAL CENTER FOR ENVIRONMENTAL RESEARCH

April 10 – 12, 2006

National Transportation Safety Board
Washington, DC 20594

Prepared by Alpha-Gamma Technologies, Inc.
4700 Falls of Neuse Road, Suite 350, Raleigh, NC 27609

ACKNOWLEDGEMENTS

This report has been prepared by Alpha-Gamma Technologies, Inc. with funding from the National Center for Environmental Economics (NCEE). Alpha-Gamma wishes to thank NCEE's Maggie Miller and the Project Officer, Cheryl R. Brown, for their guidance and assistance throughout this project.

DISCLAIMER

These proceedings have been prepared by Alpha-Gamma Technologies, Inc. under Contract No. 68-W-01-055 by United States Environmental Protection Agency Office of Water. These proceedings have been funded by the United States Environmental Protection Agency. The contents of this document may not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Table of Contents

U.S. EPA NCER/NCEE Workshop
Morbidity and Mortality: How Do We Value the Risk of Illness and Death?

Washington, DC
April 10-12, 2006

Session III: Panel Discussion on the Use of the Internet in Valuation Surveys

Nathalie Simon, U.S. EPA, National Center for Environmental Economics

Will has asked me to sort of set things up for the panel discussion, so I'll talk through it a little bit and then present the charge questions. The way I see it, there are three kinds of internet surveys. There are those in which you recruit individuals into the sample using standard random probability sampling—and then you ask people to actually complete the survey over the internet, using a link that you provide. Then there are internet surveys using standing panels, and there are two kinds of standing panels: those in which individuals self-select into the panel and those in which the panel is created using sampling techniques.

It seems to me that there are benefits to all those types of web-based surveys. Generally speaking, once the survey is administered, you tend to have quicker turnaround on the results. In addition, you often have lower costs with these types of surveys and lower respondent burden. You can have greater accuracy as well—there's no interviewer bias or data-entry mistakes to worry about. Generally, individuals are entering the data the way they want to, and then they submit the results to you.

There is also greater flexibility in how the information is presented. You can have complicated skip patterns programmed directly into the survey instrument—you can have extensive use of graphics and color, which would be expensive or difficult to do using other modes. You can also have more interactive questions and can basically tailor the survey to individuals as they're going through it. Especially in the case of the standing panels you have the availability of some unique information that has been collected prior to the survey being administered.

You can also get information on time for question, and you can have extensive variable-tracking information if you need it. In some cases, you also have the possibility of using a voice-over, which can be very helpful in getting people to understand the questions that are being asked and to take the time to listen to the questions as well as reading them.

Of course there are a number of problems associated with web surveys as well. With the panel-based surveys you can often have low response rates. In fact, those of you who are users of Knowledge Networks, if you start looking at the response rate from the time that people are initially contacted to join the panel, the response rate is rather low. Given this low response rate, non-response bias then becomes an issue. In other venues and at other conferences, I've also heard a concern expressed that panels run the risk of creating

"expert survey takers"—I believe Reed [F. Reed Johnson] referred to them as "trained seals" at one point. That is a concern, as well. If you're going to the same individual repeatedly with surveys, do you create this expert survey taker?

There are other issues as well, especially with those surveys in which individuals are self-selecting into the panel. Really, these result in little more than convenience samples. It's often difficult to tell whether you're getting more than one individual from a household and things like that. You may have problems with actually downloading the information from the internet, and you may run into technology constraints as well.

Regardless of these problems, we are intrigued by the benefits associated with these web-based surveys, especially given that telephone surveys are becoming more and more difficult to do and mail surveys are also somewhat difficult—these modes pose problems when we're dealing with complicated questions that do involve complicated skip patterns or where we would like to use more complicated graphics.

As an agency though, to my knowledge, I think we've managed to use web-based surveys only on a very limited basis, and generally these have been surveys that have been done for research purposes. You'll hear more about one of these tomorrow, "Eliciting Risk Tradeoffs for Valuing Fatal Cancer Risks." This was work done by Chris Dockins, Melonie Sullivan, who is no longer with our office, and George Van Houtven, who will be presenting the paper tomorrow. Two other surveys looked at willingness to pay for water improvements, one designed by Kip Viscusi looking at eliciting willingness to pay estimates for improvements to fresh water, and then a survey looking at coastal water improvements.

But, again, these were surveys that were either couched in terms of pure research or testing of survey instruments—so, they were pilot surveys. We have yet, at least to my knowledge, to get approval for a web-based survey that would feed directly into a policy analysis for one of our rules and regulations.

Faced with the problems associated with web surveys, but trying to balance those against the benefits that we could exploit, I wonder whether there is a way to actually address some of these issues. One of the most important ones, perhaps, is this issue of non-response bias. It seems to me that non-response bias is perhaps more of an issue if you are dealing with low response rates, but it seems that you have a potential for non-response bias regardless of what the response rate is, unless of course you're dealing with a survey where you have 100 percent compliance. So, it seems to me that one question is "How do we address that?"—How do we go about trying to improve the representativeness of the sample or "How do we test for sample representativeness?"

Thinking about all these things, we had asked our panelists here to think about several questions as they were looking back over their own research and what they've done in this field. [referring to a slide] We have the questions for the panelists up on the monitor here. Basically, we've asked people to think about their experience with using the internet as a survey mode and to think about the choice of the survey mode in their

research and to consider the tradeoffs between convenience, cost, and bias and to comment on the key issues. Specifically, we've asked them to address these questions:

- What special issues must be considered when using the internet as a survey mode?
- Are there special circumstances where it makes sense to use the internet for stated-preference surveys?
- What conditions or circumstances does the internet provide and under what circumstances should the mode be avoided?
- What specific follow-up analysis or testing should be conducted when using the internet?

Brian Harris-Kojetin, Office of Management and Budget

I'm going to take a very "10,000-foot-level perspective" here and then focus in a little bit and touch on some of the things that Nathalie mentioned, but I suspect that others in the panel will focus even more deeply into the specific issues. For those of you who have ever wondered, "Why does OMB review our surveys?" it is because we are required to by law. The Paperwork Reduction Act requires that any information collection that is sponsored or conducted by a federal agency go through this review, and the purpose of this is to improve the quality and practical utility of information that is gathered by federal agencies.

I want to make you aware of some new guidance that we recently issued in January of this year. It's entitled "Questions and Answers When Designing Surveys for Information Collection," and it covers a broad swath of things, but I'll just provide a brief overview of that. Just so you know, the intended audience for this guidance is very broad. It's intended to be used by people implementing the Paperwork Reduction Act—Chief Information Officers, Program Managers, survey folks, who are out there in the front lines doing this—and it covers a wide variety of different topics—everything from what do you have to do in terms of some basic process issues in terms of submitting the information collection requests or "OMB clearance packages," as they're more popularly known to what kinds of different issues you need to address and explain and justify and document here. I'm going to focus on a few of these that are related to some of the things you're interested in here in terms of internet surveys.

Specifically, we have questions on when should agencies consider designing a survey. Obviously, a survey is just one method in a social scientist's arsenal—it's appropriate some kinds of questions and issues and not so much for others. We have a whole section on sampling covering probability samples, coverage issues, sampling frames, . . .We bring in the issue that Nathalie raised, too, in terms of non-probability samples—that there are these internet panels out there that are essentially convenience samples. Even though these panels often boast of their numbers, which can reach over a million people, what you have is still an entirely self-selected convenience sample of "1.2 million" people who had nothing better to do than stumble across a web site and say, "Sure."

I'll also touch on several points about the mode of data collection . . . also the Government Paperwork Elimination Act (GPEA), which OMB is also in charge of implementing. This required agencies to allow citizens electronic options for reporting to the federal government. Although this law was not really written with surveys in mind, it can be applied that way—it's more for people who are applying for benefits or things like that or for businesses conducting transactions with the government. . . .

There's also a short section here in the guidance on stated preference methods. For those of you familiar with OMB Circular A-4, there's nothing really new here.

Generally, agencies are being encouraged to do a lot of electronic reporting, but there are some important stipulations in GPEA—agencies are encouraged to do this, "as practical." So, if you're doing a very small-scale survey or if you're doing anything with fewer than 5,000 respondents you don't even really need to consider it—or if it's otherwise just not practical or cost-effective.

Most federal agencies that use the internet for their surveys use it as one option in a multi-mode survey. Looking across agencies, it's being used more and more for establishment surveys or business surveys or surveys of organizations or institutions like hospitals or schools, and sometimes it is being used as the sole mode for those or for some specialized survey, such as a web site satisfaction survey. It's being used as the sole mode pretty much exclusively in cases where your target population or some sub-population of that has nearly universal web access. Not all businesses have that, but in certain industry sectors you can really count on that. There are several post-secondary school surveys that are now based exclusively on web collection.

One other thing that I want to point out is that web surveys are sometimes touted as convenience, and with some of the things I've seen from agencies it's not clear if they're thinking about the respondent or themselves. It can be very convenient for the *researcher* sometimes to use a web survey, but not so much for the respondent. The worst case scenario that I've seen in this regard is where the agency sends out a request to a respondent saying, "Please do our survey on the web. What you can do is download this, print out the PDF file, go fill it out, and then get on the web and put all the information in." This is *not* more convenient for the respondents. Why not just mail them the survey and let them mail it back? Why do they have to go through these extra steps?

In terms of cost, web surveys are often portrayed as being less costly. This is true under some circumstances, especially if it's a very simple survey that doesn't require much complex programming or testing. We have a lot of government surveys that very quickly become very complicated

In terms of bias and error reduction, we're looking for agencies to take these things into account and talk about how they are dealing with them in terms of why they've chosen the mode or modes that they're using.

In terms of choosing an internet survey as the preferred mode or one to be avoided, in reviewing packages we're basically looking for a good understanding and justification of how the agency is balancing some of the advantages and disadvantages—and Nathalie mentioned a number of these. Email reminders are certainly cheap and convenient for prompting respondents, especially if you can include that hyperlink in the email message that will take them right there. That has advantages over sending them a postcard with a long address that they have to type in. You do get faster data collection without delays in receiving the data. For instance, respondents can't tell you, "Oh, I mailed that last week,"—you know whether it's completed or not. Using visual aids and sometimes even multimedia is another advantage, as is the ability to build in some of these edit-and-consistency checks.

Disadvantages: Again, reflecting some of those issues mentioned earlier in terms of coverage and non-response and measurement error. What is the sampling frame?—Where did this sample come from?—Can you actually draw a random sample from your target population?—How well are you covering your target population? There are issues of response rates, in general—again, when they are used as a sole mode, web surveys tend to have lower response rates than other modes. That said, they are more often used as a mixed mode. Respondents have to be computer-literate and have access. There are hardware and software differences that can affect your presentation. Finally, there are some respondent concerns about confidentiality when giving information over a web site.

In terms of follow-up analysis or testing, I want to make two points. One is that pre-testing is just as important [as follow-up]—have the questions been tested to determine whether they are functioning as intended? When you're putting this on a web instrument, you need to do the usability testing as well. As far as follow-up analyses to assess a potential non-response bias—we all recognize that non-response rates *don't* indicate non-response error—they're an indicator for the *potential* for non-response bias. We expect that surveys collecting influential information should achieve high response rates, and agencies need to consider how what they are doing is going to give them data of the quality that they need. Our guidance, as many of you are probably aware, says that if an agency is getting a response rate of less than 80 percent, they need to plan a non-response bias analysis. There's a variety of ways of doing that—I think some of the people [fellow panelists] are going to talk about some specific examples here. Bob Groves and Mike Brick have taught a course now several times at a few federal agencies as well as to the general public—this is in the joint program in survey methodology—on Practical Tools for Non-Response Bias Analyses.

Shelby Gerking, University of Central Florida

I want to report on some joint work that Mark Dickie and I have done using web-based surveys in valuation studies. We have some experience, at least, working with internet panels. We've worked with the CentERpanel at Tilberg University in the Netherlands and looked at willingness to pay for greater protection of seals there. That was back in the early part of this decade. Also using CentERpanel, we've looked at willingness to

pay for reduced risk of pancreatic cancer. Using Knowledge Networks, we've looked at blue-collar workers' willingness to pay for on-the-job-safety improvements. That was another study done earlier in this decade. More recently we've looked at parents' willingness to pay for reduced skin cancer risk to themselves and to young children ages 3 – 12. This last study is the one that I want to base my remarks on now, because it serves as a side-by-side comparison to the computer-assisted study that Mark Dickie reported on earlier.

The Knowledge Networks, or KN, Skin Cancer Survey in 2005 was transmitted to about 1200 panelists, and we, in one way or another, got down to 644 panelists with a child between the ages of 3 and 12 years who actually did complete the survey that was provided. The panelists completed the survey at home—there was about a 3-month period for Knowledge Networks to design, pre-test, and field the survey and for respondents to return a usable data set. It was a very smooth process, with very good, helpful people to work with.

The comparison is with the Hattiesburg Skin Cancer Survey from 2002 that Mark Dickie reported on. The survey was virtually identical, though not exactly identical, to the Knowledge Networks survey. It consisted of a sample of 612 parents with at least one child between the ages of 3 and 12 years. As Mark indicated, that survey was obtained by random-digit dialing of Hattiesburg area residents, and it took about a hundred calls from the poor students there to generate one completed survey. There were lots of hang-ups and lots of reasons why people might say "No," but there was also an eligibility problem, of course, because people had to have at least one biological child living at home between the ages of 3 and 12 years—that accounts for a lot of the extra phone calls. Respondents came to the University of Southern Mississippi campus and took this survey in a computer lab there, so rather than just being able to off-load the survey to the good folks at Knowledge Networks, we needed a lot of students and oversight to make sure that we at least knew what was going on in this computer lab.

As to the cost, using Knowledge Networks cost us $82 per completed survey. This excludes the investigator time needed to develop the survey—in other words, the clock starts running when you hand the survey to Knowledge Networks. It includes all pre-test costs and all of Knowledge Networks costs and all university indirect costs. With the Hattiesburg survey, it cost about $123 per completed survey. Again, that excludes the cost of investigator time used for survey development. Although it's not exactly the same, I tried to make the comparison as much apples to apples as I could. Anyway, the oversight that you need with one of these computer-assisted surveys is significant, and I valued Mark's time and my time on that job at about 9 cents per hour. The Hattiesburg cost includes the $25 participation fees provided to those who came to the computer lab and took the survey, and it includes all pre-test expenses, programming costs, labor and telephone charges, and university indirect costs. So, the Hattiesburg survey came at about a 50 percent cost premium.

Data Quality:
As far as sample composition, the Hattiesburg sample was more highly educated than the KN sample, and this is what you would expect, given that random-digit dialing was used to recruit the survey. The sample was more highly educated than you would have expected, given the census data for the Hattiesburg area. The Knowledge Networks survey was more representative of the United States population, but I would call attention to the fact that we're not really sure who completed all the surveys. When we were debriefing pre-test participants, out of eight such persons that we spoke with (Knowledge Networks had arranged the calls and was on the line also), we found out that one of them was not the person who had completed the survey—it was that person's spouse, instead. How widespread this problem is I have no idea—I'm not trying to condemn the Knowledge Networks survey on the basis of one observation.

The average survey completion time for the Hattiesburg survey was 26 minutes, and we had projected a completion time of 25 – 30 minutes, based on our own experience taking the survey and the time it took pre-test respondents. Twenty-three percent completed the survey in 20 minutes or less—you also want to know how many people just ripped right through it and probably didn't pay too much attention to what they were doing. In the KN survey, it took 1178 minutes for those respondents to complete the survey. One interpretation is that these people obviously work much more carefully than they do in Hattiesburg, but there are other interpretations as well that could be offered. One is that if you're at home and you're doing this on the internet, you're free to look at the survey. That's when the clock starts running, and then you say, "Yes, I see what this is—it looks very interesting—I think I'll do it in three days." That's possible. Another possibility is that you look at the survey and begin to do it but you decide to come back later to finish it. Then when you return, you have to pick up where you left off and reconstruct your train of thought. Seventeen percent of the surveys returned were "resumed interviews"—this is how Knowledge Networks refers to a survey that exceeds 100 minutes. Actually, I would classify a resumed interview as any that took from 30 minutes on, but this is how Knowledge Networks furnishes the data. Thirty-nine percent of KN respondents completed the survey in 20 minutes or less.

Another issue is the level of respondent engagement—the question distractions and interruptions come in. Looking at the KN survey, you begin to look at that average completion time, and you begin to think a lot about distractions and interruptions. Imagine someone trying to complete the survey and the cat is climbing up the drapes, the dog is barking, and the kids are playing with matches, someone's at the door, the telephone's ringing—all these things could be happening at once, who knows? Or, none of those things could be happening and someone just decided on their own that they would rather complete the survey later. Again, who knows? Anyway, the possibility of distractions and interruptions is certainly there.

With the Hattiesburg survey, where the respondents were completing the survey in the university computer lab, about the only possible distraction would be someone teaching calculus across the hall and a respondent might decide that they would rather go learn about the quotient rule. I don't think this happened, though.

A number of people in the KN sample had taken a lot of surveys—presumably they were experienced—that could be good, it could be bad—Reed referred to this sort of thing as the trained seal effect. Who knows? With the Hattiesburg study, it was a fresh sample—they hadn't participated in any previous surveys, at least none that we had done. There was also more item non-response in the KN survey than in the Hattiesburg survey. In the computer-assisted survey we had *practically no* item non-response, whereas in the KN survey there was *a lot*.

Did changes in features of the hypothetical sun lotion that Mark described alter willingness to purchase it in a predictable way? Well, with a change of price, yes. As the price went up, willingness to buy the stuff went down. How about extent of risk reduction? In the Hattiesburg survey, in a between-respondent comparison, we got higher willingness to pay for larger risk reduction, so there's an external scope test there. With the KN survey, again in a between-respondent comparison, we got *significantly lower* willingness to pay for larger risk reductions. What are possible explanations for the difference in outcome? I mentioned the greater education level of the parents in Hattiesburg. Maybe better-educated people are just in a better position to do these surveys than less-educated people. We did a variety of tests to try to detect whether education level had any bearing on the outcome of the extent of risk questions that we asked, and the answer was "no." It was just that the KN respondents, in general, were poorer at this than the Hattiesburg sample.

As far as resumed interviews in the KN sample, if you just took out all the people who took 100 minutes or more to complete the survey, would the basic results change? The answer is "no." Was there a greater level of engagement on the part of the respondents in the Hattiesburg survey? Maybe—I don't know—but it is a concern. One thing I wish we could have generated was some within-respondent evidence as to how people respond to changes in risk.

Alan Krupnick, Resources for the Future

Wow—those are quite problematic responses to that survey of Knowledge Networks, and I don't want this panel to become a referendum or a judgment of Knowledge Networks, but it's probably worth saying why we mention Knowledge Networks so much. There may be people here who don't understand that. The reason that Knowledge Networks is so attractive is because they made an attempt through random-digit dialing to convert people to their panel who were not internet users. They gave them this special technology, webTV. You don't need a computer to take these surveys when you have this technology, so it deals with the problem of non-internet-users.

We (Maureen [Cropper], Nathalie [Simon], Anna [Alberini], and myself) did a national U.S. mortality-based survey in the year 2000 or so for our mortality risk valuation work, which has been reported in a couple of different journals. I wanted to talk a little bit about our experiences, particularly in regard to some of the responses I have after

listening to Shelby's presentation. Then I want to give a little advertisement for what's going to happen at Resources for the Future in October.

So, we've had experience with both Knowledge Networks and Ipsos Reed, which is a Canadian firm that does probability-based internet sampling but doesn't have the webTV technology. First, going through the work on mortality risk valuation, we basically had exactly the same setup, although different locations, as Shelby. In our Canada sample, it was a random-digit-dialed sample of people in Hamilton, Ontario that came to a central location to take the survey on a computer. Then later we did a national sample using Knowledge Networks on webTV or the computer. We got extremely close results on both of those surveys. Many of you in the audience have seen our bar graphs—almost equal responsiveness to the bids, which were basically PPP-corrected, so they were equivalent bids across the two countries. We had significant external scope effects. We had very little item non-response. Maybe this can be explained partly by the fact that we were using the panel in its early days—by the time Shelby got to it, it was rather old.

The one benefit that we saw from Knowledge Networks that you can't get easily from these in-person, self-administered surveys at centralized locations is that you can pick up infirm or immobile people—if they're in your panel or however you get them. That's important to many health surveys, so we thought that was a benefit from our work although I can't prove it. We also looked at the timing issues—these people who take 100 minutes or more, and so on. As Shelby mentions, we didn't find any effects on timing.

So, let me go to our Adirondack survey. This was done by Knowledge Networks in New York, so our sample of people was panelists from New York state, where we estimated the willingness to pay for improvements in the Adirondacks, and it was set up with an external scope test framework. What we did here is we used two different modes—an RDD mail survey and a panel internet survey. We had Knowledge Networks do both of these for us. The survey for the two was as identical as we could make it, given the difference in mode.

So, we did a few things. The first is that we looked at the demographics comparing the two modes to each other and comparing them to the census. We did pretty well. There were some observable differences across various samples, which we corrected using weighted regression. Differences in observables really don't cause any major problems. Then we used a Heckman selection analysis on the panel internet survey using KN's panel data, so we know from the panel who was exposed to the survey and had an opportunity to take the survey but chose not to. We did the analysis with that group and with the group that did take them, and we did find some groups less likely to respond to our survey—women, minorities, and the lower-educated were less likely to respond—but we didn't' find any statistical effect of the unobservable component of response on willingness to pay. Of course, the limitation of this kind of analysis is that we did not look further back in the chain to compare our results to people who chose not to be on the panel. So, that's going all the way back to the beginning of the RDD effort, and we weren't able to do that.

Finally, we compared the frame mode, the RDD mail results for willingness to pay to the panel internet results for willingness to pay, and we found that they were quite similar—there was no statistical difference between those two. For what it's worth, that's what we found.

Finally, I just want to mention what we're going to be doing in October. We'll be hosting an OPEI-funded workshop on the general topic of sampling bias. It's called "Sampling Representativeness: Implications for Administering and Testing Stated-Preference Surveys." We're going to bring in experts—some of the people on the pane here—survey researchers, statisticians, cognitive psychologists, and government officials, including Brian [Harris-Kotejin] and others to help better define the problems and work toward a solution. Our motivation here is this linkage that OMB makes between low response rates and therefore unreliability of the surveys. Our view is that you could have an 80 percent response rate that doesn't guarantee representativeness, or you could have a 10 percent response rate that does. What we need to do is decide what our performance measures are going to be and then what protocols we need to follow—and I know OMB is interested in defining those kinds of protocols—to permit us to take advantage of internet technologies that are out there to get these surveys done at low cost, quickly, and flexibly to give all the advantages that Nathalie mentioned and not give that up on what may be a false goal of lowering non-response rates. What we want to lower is sampling bias, and that's a different thing.

———————

Jon Krosnick, Stanford University

I'm a professor of communication, political science, and psychology at Stanford University, and I'm delighted to have the opportunity to speak with you this afternoon. I make my life, among other things, focusing on survey methodology. Increasingly lately I've found myself obsessed with mode—doing mode studies for a variety of reasons and trying to answer the general question of: What impact does mode choice have on survey outcomes?

As some of you no doubt know, there are lots of different sources of error in surveys. One is coverage error. That is, if we're doing a telephone survey, we'll fail to reach households that have no telephone access at the moment that we call. There is non-response error. That is, people of particular types choose not to participate and therefore bias the sample composition. Interviewers make errors in reading questions and in hearing and recording answers. Respondents make errors in interpreting questions and in doing inadequate memory searches for relevant information—integrating or reporting, as well. When you put all of this together, if produces what we think of these days as "total survey error"—that's sort of the sum of all of these errors. In order to provide the most accurate measurements from a survey, we want to minimize all of these various sorts of error. My focus during my few moments today is on how mode can impact the sum total.

There are various ways to think about how mode choice does have impact. As I've said already, if you decide to do a telephone interview, you have coverage error—period. That doesn't mean your results will be different from the results you would get if you had overcome that coverage error, but it does mean that if you ask people a question like "Do you have working telephone service in your house?" you will not get the right answer because of the method you used to contact people.

But, there are some other cases in which mode differences are less predictable, less expected, and less anticipatable. Let me say from the start here, my discussion is going to focus on probability samples only. As you've heard already, there are internet survey firms offering, at fabulous prices, internet surveys provided from non-probability samples. We *have* done work on non-probability samples, and we find consistently that those samples are less accurate in the data that they produce, sometimes *dramatically* inaccurate. I personally don't take them seriously for the kinds of work that requires generalization to populations, so I'm not going to spend any time talking about that today. What I am going to talk about very briefly [referring to slide] are the four primary "contender" modes these days and the considerations or variables associated with these modes that can help differentiate between them. I'm not going to go into great detail, but we could think through how face-to-face interviews, versus telephone interviews, versus paper-and-pencil questionnaires could differ in the rapport and trust that the respondents feel they have in researchers, in the confidentiality they feel their responses can be assured, the modeling of commitment that a researcher or an interviewer might provide and become contagious with respondents, and so on. There are lots of these different factors and 10 minutes is not adequate time to go through this theoretical analysis.

What I do want to do, though, is very quickly skate you across a set of mode comparisons leading to the ones we care about most on the internet. First of all, comparing face-to-face with telephone interviews, you know that in the late 60's to early 70's when telephone penetration in households became essentially universal, the appeal of the many practicalities of the telephone attracted researchers to that mode, especially the reduced cost. The question that arises is: Was there any price paid by saving that money and moving to the telephone and not having to ship interviewers around the country, being able to supervise them closely, being able to complete surveys much more quickly, and so on. [Dr. Krosnick then showed a slide that listed "all the studies that had been done comparing face-to-face to telephone interviewing before we did our work, and showing all the design flaws that they suffer from that prevent you from being able to make any inferences, unfortunately, from them about the question we care about."]

So, we did a study that used three different national experiments—a data set collected in 1976, another one in 1982, and another one in 2000—conducting the same survey side-by-side, random-digit-dialed telephone nationally as well as face-to-face with area probability samples. I want to just show you, without going into great detail, that for the full samples [again, referring to slide] there was more reporting error in the telephone data than in the face-to-face data across the board. The data show that the real cost of moving to the phone is for the least-educated respondents—they get hit the hardest by the added cognitive burdens of a telephone conversation. In addition, the telephone

respondents complained more often about how long the interview was lasting, they expressed more dissatisfaction with the length of the interview, they said that the interview was "too long" more often, and, amazingly, their interviews were *shorter* than those of the face-to-face respondents. Is it surprising that people feel rushed on the phone?—maybe not.

Interviewers also rated the respondents on the phone as "less interested" in the interview process and "less cooperative" with the response process, and we found that the telephone respondents were more likely to distort answers in socially desirable directions than were the face-to-face respondents, who presumably developed a sense of rapport and trust with their interviewers more effectively. In addition, the telephone respondents said they were more uncomfortable discussing sensitive topics, and the interviewers rated the phone respondents as being "more suspicious" than the face-to-face respondents.

Okay, that was very, very quick, but you get the bottom line, which is that in this contest face-to-face wins.

What about a competition between telephone and paper-and-pencil, as we move closer to the internet case? In this case, this is a study that we did for NASA, funded by the FAA—a study of airline pilots who fly you and me around on commercial airplanes. This was using a survey project called the National Aviation Operations Monitoring System. A field experiment was involved—licensed pilots were interviewed and they were randomly assigned either to be interviewed by telephone or self-administered questionnaires, and they were asked factual questions. We built into the experiment a measure of the accuracy of answers, and what we found was that the telephone provided substantially more accurate responses than the paper-and-pencil questionnaires did. So, in this case when you take the interviewer out and leave respondents on their own, the quality goes down. In general, the respondents forgot events they should have reported more on paper than they did when they were walked through the questionnaire by an interviewer on the telephone.

The respondents answering the paper-and-pencil questionnaire actually realized that their answers were less accurate. When we asked them to rate how accurate the answers were as descriptions of their experiences, they reported significantly lower confidence in the accuracy of their answers. The real story here is this one: Whereas it took 27 minutes on average for the respondents to complete the interview by telephone, it took only 16 minutes for the paper-and-pencil respondents to complete that very same questionnaire. They rushed through the questionnaire; they overlooked events and by failing to report them, compromised the accuracy of the data they provided. As a result, the winner in this little "race" is the telephone.

Now we move, finally, to your favorite topic: telephone versus internet. So, paper-and-pencil and computer modes seem pretty similar—no interviewer involved, just answering questions on your own—maybe we should be worried about this competition, maybe we should be pessimistic. What do the data say? Well, we have two kinds of data [again, referring to slide]. One is a lab experiment, where we brought a group of respondents

into our lab and randomly assigned them either to complete a questionnaire on a computer by themselves in a cubicle or to complete the very same questionnaire over an intercom system, being interviewed orally by an interviewer down the hall. What we found is, depending on which measure of validity we looked at, . . . large majorities of comparisons showed statistically significantly higher validity for the computer than for the oral interview and *no* statistically significant differences suggesting the oral interview was superior to the computer. So, interestingly, we find here that the computer yields more-accurate reports than the oral administration. Furthermore, in the computer case, manipulating the order in which response choices were presented to people had no meaningful impact on those answers—54 percent versus 51 percent. However, on the intercom we found a very pronounced order effect, where we manipulated the order of choices and it produced a big difference in the answers people gave.

Lastly [again, referring to a slide], the pressures toward social desirability were more powerful on the telephone than on the computer. On the computer, White respondents were quite willing to say they were in favor of decreased government help for Black Americans, whereas being interviewed on the intercom the plurality of respondents said they supported *increased* help for Black Americans instead.

So, what are our conclusions? Well, face-to-face beats telephone. Computer beats telephone. Telephone beats paper-and-pencil. So, one possibility is that face-to-face produces better data quality than computer, which produces better data quality than telephone, which produces better data quality than paper-and-pencil. If this were true, it would sort of be the case that you get what you pay for—the more expensive the method, the higher quality the data. . . . We shouldn't over-generalize here, but I guess what I would say is I think there's a lot of promise in the data I've shown you for the potential of the internet mode to produce valid data. The question is: Can it be accomplished effectively?

J.R. DeShazo, UCLA

Given all the discussion about the benefits, I don't think I'm going to cover the benefits. Let me briefly tell you what my experience has been in the context of four surveys and then talk about sample selection correction, because following up on Alan's point, I think what we do want to reduce is sample selection bias. We, entirely through the efforts of Trudy [Cameron], did go back to the random-digit-dial stage and evaluate sample selection bias for both opinions that were expressed and the propensity of being our final samples for the first three surveys that we did through Knowledge Networks.

[goes through a series of slide that describe the surveys they did]

We were very much concerned that our estimates of willingness to pay would not be representative of the U.S. population, and so Trudy began thinking about how to go about correcting for that. . . . One of the problems in random digit dialing is to figure out who chooses not to be recruited by Knowledge Networks, but the problem doesn't stop there.

Here's a summary of the process so you can get an idea of the magnitude of the problem: There's the initial random-digit-dialed contact, at which time individuals can select out of the sample if they're not recruited. They could be recruited by Knowledge Networks and not profile—that is to say, not enter their panel at time "t." Assuming they enter their profile at time "t," they may at time "t + 1" select out of the panel and not be active and thus not be available to us when we draw our sample. Then, of course, the final selection stage occurs if they are not drawn randomly or otherwise by Knowledge Networks as part of our estimated sample. What we wanted to do is explore differences and describe the systematic selection out of our estimated sample as a function of a set of individual characteristics. . . .

One of our surveys gathered data on public opinion with respect to whether the government ought to intervene in environmental health and safety programs. One concern of ours was "did the panel have a liberal bias?" and we thought we could get at this question by focusing on this question about the appropriateness of government intervention. A more fundamental question, given that we are interested in estimating demand and peoples' willingness to pay is: Does this selection process lead to a non-representative sample that is going to express a biased willingness to pay? The second approach goes about estimating marginal selection probabilities, conditional selection probabilities, and then allowing the marginal utility associated with the attributes of the programs we're interested in the peoples' willingness to pay for to depend on the propensity to respond to the survey.

Approximately half a million individuals were contacted by Knowledge Networks or one of their subcontractors. We placed a restriction on our sample—we wanted adults over 24 years of age . . .there were 1600 individuals that were recruited for the sample. The nice thing about the random-digit-dial information that we were able to obtain is that we could match it with census data. This was not easy and it took a huge amount of time. Basically, the way we did it is we used individuals' addresses and their telephone exchange and Trudy developed an algorithm to associate the probability that that individual in either that address or telephone exchange would be associated with a particular census tract. Then she very cleverly developed a set of 15 orthogonal factors plus using data on voting behavior—basically, these propensities to participate or to persist in the sample. This was extremely laborious, so much so that it justified a paper by itself (Cameron and Crawford). Let me say that there are three papers that are available on our attempts at sample selection correction.

These 15 orthogonal factors explain 88 percent of the variation [unintelligible words] characteristics across tracks, so this is a very robust selection model.

Given the limited time, let me just get to the conclusions. For the first analysis on the question of liberal bias, whether or not we were obtaining an average representation of peoples' opinions as to whether or not the government should intervene via environmental health and safety programs—we found that there was basically an insignificant point estimate of bias in the distribution of attitudes toward regulation. So,

there was no appreciable effect that resulted from selection on the response item of interest.

In the second analysis, we did find statistically significant but very, very, very tiny effects on the key parameters across respondents' propensities to persist in the panel, so much so that they were, in the context of our willingness to pay estimates, insignificant—and I'll stop there.

END OF SESSION III

Summary of the Q&A Discussion Following Session III

Mary Evans (University of Tennessee)
"It's my understanding of these panels, such as Knowledge Networks and Harris Interactive, that if you submit a fairly small number of questions they may sort of piggyback your questions onto a larger survey. I'm wondering, first, if that may explain some of the differences in Shelby Gerking's experience and Alan Krupnick's experience with Knowledge Networks in particular. Secondly, I'm wondering if anyone is aware of any studies that look at the effect this kind of piggybacking has on results, whether there's a systematic bias."

1st responder (Gerking or Krupnick)
He responded, "The Knowledge Networks study that we did was not piggybacked on any other" and added that, in fact, it was sufficiently long that Knowledge Networks determined that a time constraint should be imposed on it—they wouldn't piggyback it with another one.

2nd responder (the other one)
He added "and that's the same with ours. Ours was about 30-32 minutes on average, as well, and there was no piggybacking, so that won't explain it."

3rd responder
This person clarified that there are two kinds of piggybacking. One is when your questions go first before other people's questions, in which case there's no impact so who cares? The other possibility is that your questions get added to the end of somebody else's, and this creates two issues. He explained, "One is that your questions are now appearing when respondents are more fatigued. Secondly, prior questions have been on particular topics and have activated thinking in particular directions. There's plenty of literature suggesting that fatigue and the content of prior questions can indeed influence answers to later questions, so there's every reason to believe that that's problematic." He continued on saying, "On the other hand, there's absolutely nothing unique to Knowledge Networks or Harris Interactive in piggybacking, because if you take Alan's survey or any survey that I've done, all of the questions at the end of the questionnaire are sort of piggybacked on all the questions at the beginning of the questionnaire. So, anything that comes late in the questionnaire could be influenced by what came earlier, just as in any other case." He concluded by saying that although it's not unreasonable to ask if there's impact of early questions on late questions, but it's not unique to those firms.

Here, the questioner made an unintelligible follow-on comment.

3rd responder
This person replied, "Absolutely," and he said he would repeat the comment so that everyone could hear it. He summarized the comment, saying "that it would make a difference on the results of willingness to pay for asthma if the prior questions were about cell phones versus whether they were about asthma medications." He went on to say that he doesn't think there's any doubt about that, "and it could very well be true that your

early questions in your questionnaire can influence the later ones regarding asthma, too. In particular, there's one very well documented danger: If you ask early questions on willingness to pay for cleaning up pollution in the ocean, people will feel as if they have less disposable money available by the time they get to the asthma questions. We know of that problem, and that will occur in any questionnaire as a result."

4th responder
"There is a related issue, as well, which is the expectations of respondents when they first begin to take the survey. If they're seasoned panelists, they may be used to taking surveys where the questions are similar to: Would you open an account with thus-and-so bank if the account had these features and we threw in a free pizza? That's one kind of question. Or, to go along with Jon Krosnick's presentation: Would you vote for President Bush if he stood for election today?—Yes, No, Don't know. When you follow such a question with one such as: Now, assume you're an asthmatic—would you pay for this or that type of medication to control these or those kinds of symptoms?—then you're just increasing the level of difficulty for these questions. If somebody was not expecting to see something that difficult, maybe that would be a flag."

Trudy Cameron, (University of Oregon)
Dr. Cameron said she just wanted to acknowledge "the remarkable cooperation" that she and Dr. DeShazo received from Knowledge Networks in doing the non-response study, "going all the way back to the original RDD contacts." She specifically acknowledged the hard work of Mike Dennis and Rick Lee as well as a consultant, Dale Culp. She added, "If I had been them, I would have been very much more nervous about the downside of this enterprise. All of us, collectively, heaved a sigh of relief when things turned out pretty well, . . . but we put them way out on a limb, and we're very grateful they did cooperate in providing that data." Adding that the exercise has been done as much "at arm's length" as possible, she closed by saying that she is "comfortable that what we're finding is the right stuff."

J.R. DeShazo (University of California, Los Angeles)
Dr. DeShazo added, "These are firms—and they'll respond if we tell them what we need and they have enough lead time and planning time. One of the challenges Knowledge Networks had was that they hadn't thought to keep track of all of their random-digit-dialed contacts. They had to go back and recover that and were uncertain as to whether or not they could. So, whether we're expressing professional standards for data quality or responding to OMB, I think that there's a market out there for data collection. If we communicate our needs clearly, we're large enough demanders of the product that they are going to be responsive."

Unidentified speaker
"We use Harris Interactive to get access to their chronic disease panel for surveying patients. I've never tried to do a general population survey with them—I've done a couple with Knowledge Networks. One of the marketing strategies that Harris uses, that I believe they have implemented subsequent to Jon's study six years ago, is a fairly sophisticated propensity weighting scheme, in which every other month they conduct a

random-digit-dial survey and an internet survey from their panel and then attempt to devise a weighting scheme to match not only the demographics but the responses to certain attitude questions, particularly attitude questions that screen well for people who take internet surveys. Jon, are you aware of this sheme? Does it make sense to you? Do you think it's fixing some problems? Knowledge Networks' argument is that we can match the demographics but it doesn't really necessarily match people who are going to join a panel and answer survey questions every week."

Jon Krosnick, (Stanford University)
Saying he was happy to comment on this, Dr. Krosnick responded, "The Harris Interactive propensity weighting scheme is proprietary—they will not describe how they do it to anybody—and they *did* have it in place at the time that we did our 2000 study, which I showed you. We were provided with the proprietary propensity weights, and when we analyzed the Harris data, both with the weights and without the weights, we found that it did not change the substantive results at all—it didn't change the means or the distributions of variables. What it *did* do was increase the standard errors of the estimates. The reason for this is because when we looked at the weights, there were some as large as 20 or 30 and some as small as 0.1 or 0.2. So, the weights are dramatic and they didn't have any real impact on the results that we looked at. As it turns out, Harris will not normally reveal the questions that they use in those parallel surveys to develop the weights, but they actually accidentally sent us the questions. So, having seen the questions, I can tell you that I'm not even slightly surprised that they don't do anything helpful."

Dr. Krosnick continued, "The more recent study we've done, which I haven't mentioned to you, is one in which we compared the same questionnaire administered by random-digit-dial telephone, Knowledge Networks, and six other firms that use volunteer samples, some of which do weighting by quotas on demographics and one of which provided proprietary propensity weights. We found the same thing—the propensity weights didn't change anything, and the volunteer samples were substantially less accurate. So, my results that I showed you earlier and these new results are not focused on demographics. The vast majority of our results comparing the reliability and validity have to do with substantive measures of attitudes, beliefs, behaviors, and so on. In cases where you can compare factual matters—like whether people have a driver's license or not, whether they have a passport or not—and other figures where there are *official* numbers to compare to, the probability samples from telephone and from Knowledge Networks were equivalently accurate and the volunteer samples were notably less accurate."

In clarifying how the panels process a shorter survey, Dr. Krosnick stated, "Their panelists are answering questions every week, so they'll add your question to a survey that's already going to go out anyway. How much does this cost them to add one more question?—nothing—get your $500—fabulous.

Another responder

"I just wanted to mention with respect to the cost figures that were presented before—we might have received a bulk discount. Our experience, just for the benefit of future negotiations, was that the total cost for Knowledge Networks was less than $45 an observation for a 30-minute survey."

Jon Krosnick
"Definitely a bulk discount."

Unidentified questioner
"Does that include university overhead?" When a responder replied, "No, it doesn't," the questioner said, "Okay, that's part of the difference."

Jon Krosnick
Dr. Krosnick added, "There's also a very subtle but interesting issue on overhead for those who care about this. The universities make a distinction between subcontracts and service purchases. If it's a subcontract, you only pay indirects on the first—let's say $20,000; if it's a service purchase, you're paying indirects on everything. You definitely want to negotiate with your university to make it a subcontract so you don't pay more indirects than necessary."

James Hammitt, (Harvard University)
Dr. Hammitt said he wanted to get some of the panelists' perspectives on a question related to the cost issue. He continued, "When I first got involved in internet surveying, it seemed to me that compared with phone surveys the fixed cost of setting it up might be high but the marginal cost per respondent would be very much lower because you don't need the live interviewer. With something like a Knowledge Networks panel, there's obviously a cost to maintain the panel and an opportunity cost to use it up. Is it right that the marginal cost per respondent will tend to be much lower with internet than with phones, for example? It seems to me that that would have implications for how we design surveys, because, as Jon has commented, there's a concern that if you ask people a lot of questions they get tired out and the responses toward the end may not be very good. However, if the marginal cost per respondent is low, we should just have very short surveys of a very large sample, whereas with phone surveys there's so much cost involved in getting somebody on the line who is willing to answer your questions that we tend to go for a longer interview with them."

Jon Krosnick
Dr. Krosnick replied, "I think that's definitely misleading. Basically, when you think about fixed costs of telephone interviewing—you have to hire a staff, you have to train the staff, you have to have supervisors, you have to have facilities and machines and all that—then once you get them in there, if they keep making more phone calls obviously making one extra phone call doesn't require all that much more staff time. Similarly, Knowledge Networks has to invest a bunch of money in recruiting a panel and then equipping the panel and paying them incentives and keeping them all going every week. My guess is that adding another respondent to the panel is actually considerably expensive—you have to make recruiting phone calls and get them signed up and send

them the equipment and all that—and you have only so many people in your internet panel. So, when you say that adding one extra respondent doesn't increase the cost very much, that's sort of true, but the whole fixed cost scheme is pretty burdensome, I think. You might say that you don't have to make a new phone call. Adding that marginal respondent on the internet case isn't that expensive if you weren't going to use them anyway that week, but it's not clear that Knowledge Networks doesn't want to use them anyway."

Kelly Maguire, (U.S. EPA)
Addressing Brian Harris-Kojetin, Dr. Maguire stated that he had mentioned that "many federal governments are moving toward using mixed modes," and she said she was wondering whether any of the panelists have experience with using mixed modes. She added that one of her concerns is that "when you start to use multiple modes within one research study, you introduce other biases that become more problematic than say the non-response bias that you're trying to correct in the first place."

Alan Krupnick (RFF)
Dr. Krupnick responded, "I mentioned in my remarks that we *did* use mixed mode—we used a mail survey and the Knowledge Networks internet survey." He acknowledged that the two surveys were "not exactly the same" due to the "issues you have to confront in switching these modes"—but they were pretty close. He added, "Maybe we were fortunate to have our willingness to pay estimates not be any different across these two modes. If they had been different, then we would have faced the issue of trying to explain why, but we didn't have to do that."

_____

END OF SESSION III Q & A