Web Survey Methodology: Interface Design, Sampling and Statistical Inference

Mick Couper

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

53

Web Inkesten Metodologia:
Interfazeen diseinua, laginketa eta inferentzia estatistikoa

Web Survey Methodology:
Interface Design, Sampling and Statistical Inference

Metodología de Encuestas Web:
Diseño de interfaces, muestreo e inferencia estadística

Mick Couper

University of Michigan
Institute for Social Research
e-mail: MCouper@umich.edu

11/11/2011

Lanketa / Elaboración:

Euskal Estatistika Erakundea
Instituto Vasco de Estadística (EUSTAT)

Argitalpena / Edición:

Euskal Estatistika Erakundea
Instituto Vasco de Estadística
Donostia – San Sebastián, 1 – 01010 Vitoria – Gasteiz

Euskal AEko Administrazioa
Administración de la C.A. de Euskadi

Ale-kopurua / Tirada:
100 **ale** / ejemplares

XI-2011

Inprimaketa eta Koadernaketa:
Impresión y Encuadernacion:
Composiciones RALI, S.A.
Costa, 10-12 - 7ª - 48010 Bilbao

I.S.B.N.: 978-84-7749-468-3
Lege-gordailua / Depósito Legal: BI 2993-2011

AURKEZPENA

Nazioarteko Estatistika Mintegia antolatzean, hainbat helburu bete nahi ditu EUSTAT- Euskal Estatistika Erakundeak:

– Unibertsitatearekiko eta, batez ere, Estatistika-Sailekiko lankidetza bultzatzea.
– Funtzionarioen, irakasleen, ikasleen eta estatistikaren alorrean interesatuta egon daitezkeen guztien lanbide-hobekuntza erraztea.
– Estatistika alorrean mundu mailan abangoardian dauden irakasle eta ikertzaile ospetsuak Euskadira ekartzea, horrek eragin ona izango baitu, zuzeneko harremanei eta esperientziak ezagutzeari dagokienez.

Jarduera osagarri gisa, eta interesatuta egon litezkeen ahalik eta pertsona eta erakunde gehienetara iristearren, ikastaro horietako txostenak argitaratzea erabaki dugu, beti ere txostengilearen jatorrizko hizkuntza errespetatuz; horrela, gai horri buruzko ezagutza gure herrian zabaltzen laguntzeko.

Vitoria-Gasteiz, 2011ko Azaroa

JAVIER FORCADA SAINZ
EUSTATeko Zuzendari Nagusia

PRESENTATION

In promoting the International Statistical Seminars, EUSTAT-The Basque Statistics Institute wishes to achieve several aims:

– Encourage the collaboration with the universities, especially with their statistical departments.
– Facilitate the professional recycling of civil servants, university teachers, students and whoever else may be interested in the statistical field.
– Bring to the Basque Country illustrious professors and investigators in the vanguard of statistical subjects, on a worldwide level, with the subsequent positive effect of encouraging direct relationships and sharing knowledge of experiences.

As a complementary activity and in order to reach as many interested people and institutions as possible, it has been decided to publish the papers of these courses, always respecting the original language of the author, to contribute in this way towards the growth of knowledge concerning this subject in our country.

Vitoria-Gasteiz, November 2011

JAVIER FORCADA SAINZ
General Director of EUSTAT

PRESENTACIÓN

Al promover los Seminarios Internacionales de Estadística, el EUSTAT-Instituto Vasco de Estadística pretende cubrir varios objetivos:

– Fomentar la colaboración con la Universidad y en especial con los Departamentos de Estadística.
– Facilitar el reciclaje profesional de funcionarios, profesores, alumnos y cuantos puedan estar interesados en el campo estadístico.
– Traer a Euskadi a ilustres profesores e investigadores de vanguardia en materia estadística, a nivel mundial, con el consiguiente efecto positivo en cuanto a la relación directa y conocimiento de experiencias.

Como actuación complementaria y para llegar al mayor número posible de personas e Instituciones interesadas, se ha decidido publicar las ponencias de estos cursos, respetando en todo caso la lengua original del ponente, para contribuir así a acrecentar el conocimiento sobre esta materia en nuestro País.

Vitoria-Gasteiz, noviembre 2011

JAVIER FORCADA SAINZ
Director General de EUSTAT

BIOGRAFI OHARRAK

Mick P. Couper irakasle ikertzailea da Frankfurteko Gizarte Ikerketarako Institutuko Inkesten Ikerketako Zentroan eta Marylandeko Unibertsitateko Inkesten Metodologiari buruzko Baterako Programan. Soziologiako doktorea da Rhodesko Unibertsitatean, gizarte ikerketa aplikatuko lizentziatua Michigango Unibertsitatean, eta gainera, Gizarte Zientzietako lizentziatua Mendebaldeko Lurmutur Hiriko Unibertsitatean. *Nonresponse in Household Intervieuw Surveys* [Erantzun eza etxebizitza partikularretan egindako inkestetan] lanaren egilekidea da, *Computer Assisted Survey Information Collection* [Inkesten informazioaren bilketa informatizatua] bildumaren editore burua, *Survey Methodology* [Inkesten Metodologia] lanaren egilekidea (horiek guztiak Wileyk argitaratuak) eta *Designing Effective Web Surveys* (Cambridge) [Web inkesten diseinu eraginkorra] lanaren egilea. Gaur egun darabilen ikerketa-ildoaren ardatza, inkestatzaileek zein inkestatuek inkestak egiteko teknologia erabiltzea da. Azken 10 urteotan inkesten web diseinuari eta horren aplikazioari buruzko ikerketa zabala egin du.

BIOGRAPHICAL SKETCH

Mick P. Couper is a Research Professor in the Survey Research Center at the Institute for Social Research and in the Joint Program in Survey Methodology at the University of Maryland. He has a Ph.D. in sociology from Rhodes University, an M.A. in applied social research from the University of Michigan and an M.Soc.Sc. from the University of Cape Town. He is co-author of *Nonresponse in Household Interview Surveys*, chief editor of *Computer Assisted Survey Information Collection*, co-author of *Survey Methodology* (all published by Wiley), and author of *Designing Effective Web Surveys* (Cambridge). His current research interests focus on aspects of technology use in surveys, whether by interviewers or respondents. He has conducted extensive research on Web survey design and implementation over the past 10 years.

NOTAS BIOGRÁFICAS

Mick P. Couper es profesor investigador en el Centro de Investigación de Encuestas del Instituto de Investigación Social así como en el Programa Conjunto sobre Metodología de Encuestas de la Universidad de Maryland. Posee un doctorado en Sociología por la Universidad de Rhodes, una licenciatura en investigación social aplicada de la Universidad de Michigan, además de una licenciatura en Ciencias sociales en la Universidad del Cabo Occidental. Es coautor de *Nonresponse in Household Intervieuw Surveys* [No respuesta en encuestas realizadas en domicilios particulares], editor jefe de *Computer Assisted Survey Information Collection* [Recopilación informatizada de Información de Encuestas], coautor de *Survey Methodology* [Metodología de Encuestas] (todos publicados por Wiley) y autor de *Designing Effective Web Surveys* (Cambridge) [Diseño efectivo de encuestas web]. Su línea de investigación actual se centra en la utilización de la tecnología para la realización de encuestas, tanto por los encuestadores como por los encuestados. Ha llevado a cabo una amplia investigación sobre el diseño web de encuestas y su aplicación durante los últimos 10 años.

Index

Tables & Figures

Introduction

In the last two decades, Web or Internet surveys have had a profound impact on the survey world. The change has been felt mostly strongly in the market research sector, with many companies switching from telephone surveys or other modes of data collection to online surveys. The academic and public policy/social attitude sectors were a little slower to adopt, being more careful about evaluating the effect of the change on key surveys and trends, and conducting research on how best to design and implement Web surveys. The public sector (i.e., government statistical offices) has been the slowest to embrace Web surveys, in part because the stakes are much higher, both in terms of the precision requirements of the estimates and in terms of the public scrutiny of such data. However, National Statistical Offices (NSOs) are heavily engaged in research and development with regard to Web surveys, mostly notably as part of a mixed-mode data collection strategy, or in the establishment survey world, where repeated measurement and quick turnaround are the norm. Along with the uneven progress in the adoption of Web surveys has come a number of concerns about the method, particularly with regard to the representational or inferential aspects of Web surveys. At the same time, a great deal of research has been conducted on the measurement side of Web surveys, developing ways to improve the quality of data collected using this medium.

This seminar focuses on these two key elements of Web surveys — inferential issues and measurement issues. Each of these broad areas will be covered in turn in the following sections. The inferential section is largely concerned with methods of sampling for Web surveys, and the associated coverage and nonresponse issues. Different ways in which samples are drawn, using both non-probability and probability-based approaches, are discussed. The assumptions behind the different approaches to inference in Web surveys, the benefits and risks inherent in the different approaches, and the appropriate use of particular approaches to sample selection in Web surveys, are reviewed. The following section then addresses a variety of issues related to the design of Web survey instruments, with a review of the empirical literature and practical recommendations for design to minimize measurement error.

A total survey error framework (see Deming, 1944; Kish, 1965; Groves, 1989) is useful for evaluating the quality or value of a method of data collection such as Web or Internet surveys. In this framework, there are several different sources of error in surveys, and these can be divided into two main groups: errors of non-observation and errors of observation. Errors of non-observation refer to failures to observe or measure eligible members of the population of interest, and can include coverage errors, sampling errors, and nonresponse errors. Errors of non-observation are primarily concerned about issues of selection bias. Errors of observation are also called measurement errors (see Biemer et al., 1991; Lessler and Kalsbeeck, 1992). Sources of measurement error include the respondent, the instrument, the mode of data collection and (in interviewer-administered surveys) the interviewer. In addition, processing errors can affect all types of surveys. Errors can also be classified according to whether they affect the variance or bias of survey estimates, both contributing to overall mean square error (MSE) of a survey statistic. A total survey error perspective aims to minimize mean square error for a set of survey statistics, given a set of resources. Thus, cost and time are also important elements in evaluating the quality of a survey. While Web surveys generally are significantly less expensive than other modes of data collection, and are quicker to conduct, there are serious concerns raised about errors of non-observation or selection bias. On the other hand, there is growing evidence that using Web surveys can improve the quality of the data collected (i.e., reduce measurement errors) relative to other modes, depending on how the instruments are designed.

Given this framework, we first discuss errors of non-observation or selection bias that may raise concerns about the inferential value of Web surveys, particularly those targeted at the general population. Then in the second part we discuss ways that the design of the Web survey instrument can affect measurement errors.

Part 1: Inference in Web Surveys

Inference in Web surveys involves three key aspects: sampling, coverage, and nonresponse. Sampling methods are not unique to the Web, although identifying a suitable sampling frame presents challenges for Web surveys. I'll address each of these sources of survey error in turn.

1.1 Sampling

The key challenge for sampling for Web surveys is that the mode does not have an associated sampling method. For example, telephone surveys are often based on random-digit dial (RDD) sampling, which generates a sample of telephone numbers without the necessity of a complete frame. But similar strategies are not possible with the Web. While e-mail addresses are relatively fixed (like telephone numbers or street addresses), Internet use is a behavior (rather than status) that does not require an e-mail address. Thus, the population of "Internet users" is dynamic and difficult to define. Furthermore, the goal is often to make inference to the full population, not just Internet users.

Given this, there are many different ways in which samples can be drawn for Web surveys. These vary in the quality of the inferential claims that they can support. Dismissing all Web surveys as bad, or praising all types of Web surveys as equally good, is too simple a characterization. Web surveys should be evaluated in terms of their "fitness for the [intended] use" of the data they produce (Juran and Gryna, 1980; see also O'Muircheartaigh, 1997). The comparison to other methods should also be explicit. For example, compared to mall intercept surveys, Web surveys may have broader reach and be cheaper and faster. Compared to laboratory experiments among college students, opt-in panels offer larger samples with more diversity. However, compared to face-to-face surveys of the general population, Web surveys may have serious coverage and nonresponse concerns. Further, accuracy or reliability needs to be traded off against cost, speed of implementation, practical feasibility, and so on. Understanding the inferential limits of different approaches to Web survey sample selection can help guide producers on when a Web survey may be appropriate and when not, and guide users in the extent to which they give credibility to the results of such surveys.

In an early paper (Couper, 2000), I identified a number of different ways to recruit respondents for Internet surveys. I broadly classified these into probability-based and non-probability approaches. This dichotomy may be too strong, and one could better think about the methods arrayed along a continuum, with one end represented by surveys based on volunteers with no attempt to correct for any biases associated with self selection. At the other end of the continuum are surveys based on probability samples of the general population, where those without Internet access (the non-covered population) are provided with access, and high response rates are achieved (reducing the risk of nonresponse bias). In practice, most Web surveys lie somewhere between these two end-points.

Table 1. Types of Web Survey Samples

Type of Survey	Definition
Non-Probability Samples	
0) Polls for entertainment	Polls that make no claims regarding representativeness; respondents are typically volunteers to the Web site hosting the survey
1) Unrestricted self-selected surveys	Respondents are recruited via open invitations on portals or frequently visited Web sites; these are similar to the entertainment polls, but often make claims of representativeness
2) Volunteer opt-in or access panels	Respondents take part in many surveys as members of a Web panel; panel members are usually recruited via invitations on popular Web sites
Probability Samples	
3) Intercept surveys	Sample members are randomly or systematically selected visitors to a specific Web site, often recruited via pop-up invitations or other means
4) List-based samples	Sample members are selected from a list of some well-defined population (e.g., students or staff at a university), with recruitment via e-mail
5) Web option in mixed-mode surveys	A Web option is offered to the members of a sample selected through traditional methods; initial contact often through some other medium (e.g., mail)
6) Pre-recruited panels of Internet users	A probability sample, selected and screened to identify Internet users, is recruited to participate in an online panel
7) Pre-recruited panels of the full population	A probability sample is recruited to take part in a Web panel; those without Internet access are provided with access

Without going into the details of each type of Web survey, I'll offer a few observations about selected types, and discuss two more recent approaches to selecting or recruiting respondents for Web surveys. First, entertainment polls are not really surveys at all, but are just ways to engage an (online) audience and get feedback. However, they often look like surveys, and have been used by policy-makers as if the data are real. So, they can be viewed by lay persons (as opposed to survey professionals) as real surveys with inferential value. Second, although data to support this contention are scarce, the vast majority of surveys that people are invited to or participate in online are non-probability surveys. This is important because the target population might not be able to distinguish the difference between different types of surveys, and may treat all such surveys as of equal quality and importance.

A third observation about this typology is that intercept surveys (Type 3 in Table 1) have become increasingly popular in recent years. Almost every online transaction these days (from a purchase to a hotel stay or flight) are followed up by a satisfaction questionnaire asking about the experience. While technically a probability-based approach, the low response rates that are likely in this type of survey (few organizations report such response rates) raises questions about their inferential value. Another increasingly popular version of intercept surveys is so-called "river sampling". This approach has been offered as an alternative to opt-in or access panels, which are suffering from high nonresponse and attrition rates (see Section 1.3). The idea of river sampling is to "intercept" visitors to selected Web sites, ask a few screener questions, and then direct them to appropriate Web surveys, without having them join a panel. In practice, river samples suffer from the same low recruitment rates and self-selection biases as do opt-in panels (see Baker-Prewitt, 2010). In other words, while the approach is technically a probability sample of visitors to a web site, the nonresponse problem may lead to inferential error.

Another approach that is gaining attention recently is respondent-driven sampling (RDS). While RDS was developed as a method to recruit members of rare populations (e.g., drug users, sex workers, the homeless), efforts are being made to apply these methods to Web surveys, given the recruitment difficulties of the medium. If the assumptions of RDS are met (see Heckathorn 1997, 2002; Wejnert and Heckathorn, 2008), it could be viewed as a probability sample. In practice, however, the assumptions are rarely met, and the recruitment process can produce serious biases (see Lee, 2009; Mavletova, 2011; Schonlau and Kapteyn, 2011; Toepoel, 2011). The method relies on initial recruits ("seeds") identifying and recruiting other members of the population of interest to participate in the survey. If the chains are of sufficient length (i.e., each recruit identifies and recruits the same number of additional recruits, and this continues until equilibrium is reached), the method could yield a representative sample of that population. In practice, recruitment is rarely that successful, and the depth and breadth of social ties not as large as expected, raising questions about the method.

More recently, a lot of attention — especially in market research — has turned to social media as a way to recruit participants for Web surveys (see Poynter, 2010). The argument is made that the use of these media is so widespread that they make ideal recruiting platforms. For example, as of September 2011, Facebook reported having over 750 million subscribers worldwide — it is popular to point out that if it was a country, it would be the third largest country in the world. In practice, however, researchers cannot get access to the frame of registered Facebook users from which to draw a sample. They are therefore forced to use snowball or respondent-driven sampling methods or to post advertisements on Facebook to recruit subjects. Thus far, these

efforts have not proved very successful (see e.g., Toepoel, 2011; Bhutta, 2010). Further, although a very large group, the set of registered Facebook users represent only that population — while of interest in their own right, the set of registered Facebook users do not represent any other known population of interest.

What difference does it make if a sample consists of self-selected volunteers rather than a probability sample from the target population? The key statistical consequence is bias — unadjusted means or proportions from non-probability samples are likely to be biased estimates of the corresponding population means or proportions. The size and direction of the bias depend on two factors — one reflecting the proportion of the population with no chance of inclusion in the sample (for example, people without Web access or people who would never join a Web panel) and one reflecting differences in the inclusion probabilities among the different members of the sample who could in principle complete the survey:

$$\begin{aligned} Bias &= E(\bar{y} - \bar{Y}) \\ &= P_0(\bar{Y}_1 - \bar{Y}_0) + \frac{Cov(p,y)}{\bar{p}} \end{aligned} \qquad (1)$$

Where $\bar{y}$ represents a sample statistic (e.g., mean or proportion) based on those who complete the web survey; $\bar{Y}$ represents the corresponding population statistic; P_0, the proportion of the population of interest with no chance at all of participating in the survey (e.g., those without Web access); $\bar{Y}_1$, the mean among those with a non-zero chance of taking part; $\bar{Y}_0$, the mean among those with zero probability of taking part; $Cov(p,Y)$, the covariance between the probabilities of inclusion (p) and the survey variable of interest (y) among those with some chance of taking part; and $\bar{p}$, the mean probability of inclusion among those with a non-zero probability of taking part.

According to the equation, the bias due to the use of samples of volunteers rather than probability samples has two components. The first term in the second line of Equation 1 reflects the impact of the complete omission of some portion of the population of interest; it is the product of the proportion of the target population that is excluded from the sample entirely and the difference between the mean for this group and the mean for the remainder of the population. The second term in the second line of the equation reflects the impact of differences in the inclusion probabilities (among those with non-zero probabilities); to the extent that these probabilities covary with the survey variable of interest (y), then the second bias component will be nonzero. Although Equation 1 applies to the unweighted sample mean, $\bar{y}$, it provides some useful distinctions for understanding how more complex estimators affect the bias. In non-probability samples, p and $\bar{p}$ are generally unknown or cannot be estimated. Furthermore, in both probability and non-probability samples, $\bar{Y}$ is not known — if it was, there would be little or no need to do the survey. Thus, selection bias cannot be estimated in practice for most survey variables of interest.

1.2 Coverage

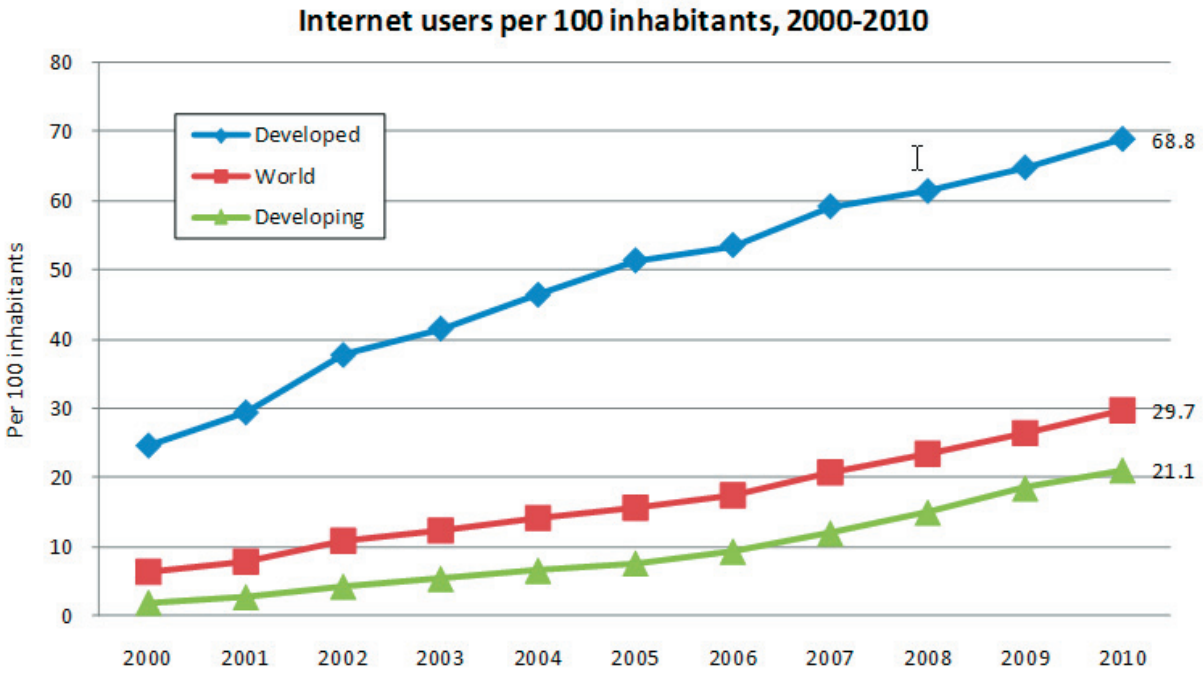
If one has access to a sampling frame, the sampling process is itself quite straightforward, using standard sampling methods (e.g., systematic sampling with or without stratification). The big issue is with regard to the frame, particularly the exclusion of certain groups. This is a problem of coverage. There are two factors contributing to coverage bias: the proportion without Internet access and the difference between those with and without access on the variable or statistic of interest. The proportion without Internet access corresponds to P_0 in Equation 1 above; the differences between those with Internet access and those without it correspond to the $(\overline{Y}_1 - \overline{Y}_0)$ term in that equation.

We should first clarify what we mean by *access*. Unlike early telephone surveys, where a landline telephone was a fixed attribute of a household, Internet access or use can be thought of in many different ways. Some surveys ask if the *household* has Internet access. Others capture whether the *person* has access to the Internet, whether at work, home or somewhere else. Still others define Internet access in terms of frequency of *use*. There are parallels to defining the mobile phone user. Having a mobile phone does not mean one can always be reached on it. Similarly, not all users equipped with smart phones (Web-enabled devices) use that capability. So, Internet access and use are dynamic terms, with implications not only for estimating coverage error, but also for sampling and nonresponse.

Regardless of how it is defined, Internet access appears to be increasing in most countries, although it appears that the *rate* of increase might be slowing (consistent with the standard S curve of adoption). The key question is whether the level of Internet penetration will reach 100% and if not, at what level will it stop. While Internet penetration is higher in some countries than others, it is still not universal. Further, the nature of Internet access and use is rapidly changing, with many new users skipping the standard browser-based approach and instead using Internet-enabled smart phones or other mobile devices (like tablet computers) to communicate. Indeed, the use of e-mail as a communication method (and potentially as a sampling and recruitment method) is rapidly being overtaken by text-messaging (whether SMS, Twitter, or other means), and social media such as Facebook are dominating the use of the Internet. So, the very nature of the Internet is changing as survey researchers try to figure out how best to use the medium for conducting surveys. In similar fashion to cell-phone only users versus traditional landline telephone users, we cannot assume that the method we have developed for standard browser-based Web surveys will apply in the same way to identifying and recruiting the new types of Internet users to our surveys.

Rates of Internet use (defined as accessing the Internet from any place at least once a week in the past 3 months) across the 27 countries of the European Union have increased from an average of 36% in 2004 to an average of 65% in 2010 (Eurostat, 2011). There is considerable variation within the European Union, with reported usage rates of 90% or over for Iceland and Norway, followed by several countries with rates over 85% (e.g., Denmark, Luxembourg, Netherlands, and Sweden) in 2010. At the other end of the distribution, several countries (e.g., Bulgaria, Greece, Italy, Portugal, and Romania) had Internet usage rates below 50% in 2010. With 58% of adults reporting regular Internet use in 2010, Spain is slightly below the European average.

Figure 1. World Internet Use over Time



The developed/developing country classifications are based on the UN M49, see:
http://www.itu.int/ITU-D/ict/definitions/regions/index.html
Source: ITU World Telecommunication /ICT Indicators database

Of somewhat more importance than the rate of penetration or use is whether and how those with Internet access differ from those without. This is referred to as the "digital divide" and initially referred to the clear demographic differences between users and non-users that were found in the early days of the Internet (e.g., NTIA, 1998, 1999). While some demographic differences (e.g., gender and race) appear to be disappearing, at least in the US, other differences (especially with regard to age and education) appear to be persisting. For example, data from a March 2011 survey by the Pew Internet and American Life Project (see www.pewinternet.org) shows that while 95% of those 18-29 use the Internet, only 42% of those 65 and older do so; similarly, 94% of those with at least a college degree use the Internet, compared to 42% of those who have not completed high school; 96% of those making $75,000 or more a year are online, while only 63% of those making less than $30,000 a year are[1].

Furthermore, it is not just the demographic differences that are important — it is the differences on all of the key variables of interest in our surveys, controlling for these demographic differences. While the demographic differences can potentially be adjusted for (given unbiased population estimates on these characteristics), it is proving to be much harder to reduce biases on key attitudinal, behavioral, and lifestyle variables.

[1] Note that these estimates may themselves be subject to error, as they come from a telephone survey which is itself subject to coverage, nonresponse, and measurement errors.

The research evidence suggests that the digital divide is not restricted to demographic characteristics, but extends to a wide range of health variables and attitude measures, for example (see Couper et al., 2007; Dever, Rafferty, and Valliant, 2008; Lee, 2006; Schonlau et al., 2009, for further evidence on this point). Those with Internet access seem to differ on a variety of characteristics from those who have not yet gotten online. Adjusting for demographic differences between those online and those not online does not make these other differences disappear. Coverage thus remains a serious concern for inference to the general population. Without alternative modes of data collection or other ways to include the non-Internet population, serious biases are likely.

1.3 Nonresponse

Another source of potential inferential error in Web surveys relates to nonresponse. Even if we could create a sampling frame of the population of interest and invite a sample of the frame to participate in our survey, not everyone will be reached (contacted) and agree to participate. Again, as with coverage, nonresponse bias is a function of the rate of nonresponse and the differences between the respondents and the nonrespondents on the variables of interest.

Nonresponse has different implications for probability and non-probability Web surveys. Nonresponse error can be expressed as follows:

$$\bar{y}_r = \bar{y}_n + \left(\frac{m}{n}\right)(\bar{y}_r - \bar{y}_m) \quad \text{or} \quad \bar{y}_r = \bar{y}_m + \frac{\sigma_{yp}}{\bar{p}} \qquad (2)$$

Where $\bar{y}_r$ is the respondent mean for the statistic of interest, $\bar{y}_m$ is the nonrespondents mean, $\bar{y}_n$ is the mean for the full sample, and m/n is the proportion of the population that is nonrespondent. Nonresponse error ($\bar{y}_r - \bar{y}_n$) increases as a function of the nonresponse rate (m/n) and the difference between respondents and nonrespondents ($\bar{y}_r - \bar{y}_m$). The second expression in Equation 2 is equivalent to the first, where σ_{yp} is the covariance between y (the variable of interest) and *p* (the propensity to respond), and $\bar{p}$ is the average response propensity in the sample, equivalent to the response rate. This expression focuses attention on the association between the propensity to respond and the variable of interest, rather than on the nonresponse rate (see Groves, 2006). In order to estimate nonresponse bias, one needs the value of the survey statistic for both respondents ($\bar{y}_r$) and nonrespondents ($\bar{y}_m$), or the covariance between the variable of interest and the response propensity (σ_{yp}).

There is relatively little research on nonresponse *bias* in Web surveys, in part because the population parameters for the variables of interest are rarely known. What little there is has focused primarily of demographic variables or examined relatively homogenous populations (e.g., college students). Instead, most of the research has focused on response *rates* in Web surveys. Further, in non-probability surveys, nonresponse error reflects the differences between the survey respondents and the pool of volunteers from which the respondents came (e.g.,

members of an access panel), but the inference of interest is not to the access panel but to the population at large. In that sense, calculating response rates as indicators of nonresponse error makes little sense, and the term is misleading. Callegaro and DiSogra (2008) suggest using "completion rate" for the response to a specific survey sent to members of an opt-in or access panel, while the AAPOR Task Force (2010) recommend using the term "participation rate". I shall use the latter term here.

Two recent meta-analyses have examined response rates to Web surveys relative to comparable modes of data collection. Lozar Manfreda and colleagues (2008) conducted a meta-analysis of 45 experimental mode comparisons between Web and other survey modes (mostly mail), with random assignment to mode. They found that, on average, response rates to the Web surveys were 11 percentage points lower than those in the alternative mode. When the analysis was restricted to the 27 studies where the other mode was mail, the average difference in response rates was 12 percentage points in favor of mail.

Shih and Fan (2008) restricted their meta-analysis to 39 studies directly comparing Web to mail. They found an average unweighted response rate of 34% for Web surveys and 45% for mail surveys, which yielded a weighted difference of 11 percentage points, very close to that obtained by Lozar Manfreda and colleagues. Shih and Fan further examined five different study features in an attempt to account for these differences. The type of population surveyed has a significant effect, accounting for about a quarter of the effect size. The smallest difference between Web and mail response rates (about 5 percentage points) was for college populations, while the largest (about 23 percentage points) was for surveys of professionals.

Both studies found considerable variation in the response rate differences, with response rates for some Web surveys exceeding those of the other mode. But the number of studies is not sufficiently large to tease out the source of these differences, or identify under what circumstances Web surveys may yield higher response rates than other modes.

Turning to probability-based panels, three examples can be provided. The FFRISP (or "Face-to-Face Recruited Internet Survey Platform"; see Krosnick et al., 2009) panel used an area probability sample and face-to-face recruitment, obtaining a response rate of 51% for the household screener (among eligible households), 90% for the recruitment interview (among screened households), and 40% for enrollment in the panel (among those who completed the recruitment interview), yielding a cumulative recruitment rate of 18% (Sakshaug et al., 2009). Participation rates to the individual surveys sent to panel members will further lower response rates (see below).

The Dutch LISS (Longitudinal Internet Studies for the Social Sciences) panel used an address frame and telephone and face-to-face recruitment. Scherpenzeel and Das (2011) report that in 75% of eligible households a contact person completed the short recruitment interview or answered a subset of central questions. Among these, 84% expressed willingness to participate in the panel and 76% of those registered for panel membership, yielding a cumulative recruitment rate of 48%.

The Knowledge Networks (KN) Panel used RDD telephone methods for recruitment until 2009 when it switched to a mix of RDD and address-based sampling (ABS). Using a specific example from 2006, Callegaro and DiSogra (2008) report a mean household recruitment rate of 33%, and a household profile rate (panelists who completed the set of profile questionnaires after joining the panel) of 57%, yielding a cumulative recruitment rate of 18.5%.

In all these cases, the panels also suffer from attrition over the life of the panel, along with nonresponse to specific surveys sent to panelists. For example, Callegaro and DiSogra (2008) report an 85% response rate to one survey in the KN panel. Scherpenzeel and Das (2011) report response rates in the 60-70% range for individual surveys sent to LISS panel members. These examples show the challenges of recruiting and retaining panel members using probability-based methods. The response rates and nonresponse bias at the recruitment stage may be similar to that for other modes of data collection. But, given that this is followed by additional sample loss following recruitment, the nonresponse problem is compounded. However, once panelists have completed the screening interview or profile survey additional information is available to assess (and potentially reduce) nonresponse bias at the individual survey level and attrition across the life of the panel.

In summary, response rates across various all type of Web surveys appear to be lower than for other modes and — as is true of all modes of data collection — appear to be declining. One hypothesis for the lower response rates to online surveys relative to other modes of data collection may be that Web surveys are still relatively new, and methods for optimizing response rates are still under development. I turn next to a discussion of strategies to increase response and participation rates in Web surveys. There is a growing body of research on ways to increase response rates in Web surveys. Again, I offer a brief review of some key findings here. For more research on nonresponse in Web survey, the interested reader is directed to www.websm.org, which has an extensive bibliography on Web survey methods.

One factor affecting response rates is the number and type of contact attempts. Both the Lozar Manfreda et al. (2008) and Shih and Fan (2008) meta-analyses find significant effects of the number of contacts on the differences in response rates between Web surveys and other modes. When the number of contacts was small in each mode, the response rate differences were closer than when a large number of contacts were used, suggesting that additional contact attempts may be of greater benefit in other modes than they are in Web surveys. There is evidence suggesting that while e-mail reminders are virtually costless and that additional e-mail reminders continue to bring in more respondents (see, e.g., Muñoz-Leiva et al., 2010), there is a sense of diminishing returns, with each additional contact yielding fewer additional respondents. It also suggests that the value of an e-mail contact to a respondent may not be as great as, say, a mail contact.

A related factor that has received research attention is that of prenotification. A prenotice is a contact prior to the actual survey invitation, informing sample members of the upcoming request. Prenotification may be thought of another contact, much like reminders. The research evidence suggests that the mode of prenotification may be more important than the additional contact it represents. Several types of prenotification have been studied in addition to e-mail, including letters (Crawford et al, 2004; Harmon, Westin, and Levin, 2005), postcards (Kaplowitz, Hadlock, and Levine, 2004; Kaplowitz et al., in press), and SMS (Bosnjak et al., 2008). The findings

suggest that an e-mail prenotice may not offer many advantages over no prenotice, but a prenotice in another mode (letter, postcard, or SMS) may be effective in increasing Web survey response rates.

Another well-studied topic in Web surveys relates to incentives. Much of this work is summarized in a meta-analysis by Göritz (2006a; see also Göritz, 2010). Across 32 experimental studies, she found that incentives significantly increased the proportion of invitees starting the survey (odds ratio = 1.19; 95% confidence interval: 1.13-1.25). The general finding in the survey literature is that prepaid incentives are more effective than promised or conditional ones, and cash incentives are more effective than alternatives such as in-kind incentives, prize draws or lotteries, loyalty points, and the like.

Despite this research evidence, lotteries or loyalty-point incentives, conditional on completion, are popular in Web surveys, especially among opt-in or access panels. A key reason for this is that is not possible to deliver prepaid cash incentives electronically. To do so by mail is expensive, and requires a mailing address. If the response rate is likely to be very low (as we have seen above), the increase in response may not justify the investment for a prepaid incentive (but see Alexander et al., 2008). Further, the cost of lotteries is usually capped, i.e., a fixe amount of money is allocated for the prizes regardless of the number of participants. This makes it easier for panel vendors to manage costs.

Given the popularity of lotteries or loyalty points among vendors, are they effective in encouraging response from sample persons? Göritz (2006a) found that lottery incentives produce higher response rates than no incentives in her meta-analysis of 27 experimental studies involving lotteries, most based on commercial panels. In her meta-analysis of 6 incentive experiments in a non-profit (academic) panel, she found no significant benefit of a cash lottery (OR = 1.03) over offering no incentive (Göritz, 2006b). Thus, lotteries may be better than no incentive for some types of samples, but it is not clear whether they are more effective than alternative incentive strategies.

Bosnjak and Tuten (2003) tested four incentive types in a survey among 1332 real estate agents and brokers. A $2 prepaid incentive via PayPal achieved a 14.3% response rate, while a $2 promised incentive via PayPal obtained 15.9%, a prize draw after completion obtained 23.4%, and a control group with no incentive obtained 12.9%. One explanation for the relative success of the prize draw is the cash was not used for the prepaid or promised incentives — for the PayPal incentive to be of value, one must have a PayPal account and have an expectation of additional money added to that account.

Birnholtz and colleagues (2004) conducted an experiment among earthquake engineering faculty and students. A mailed invitation with a $5 prepaid cash incentive obtained a response rate of 56.9%, followed by a mailed invitation with a $5 Amazon.com gift certificate (40.05 response rate) and a n e-mailed invitation with a $5 Amazon.com e-certificate (32.4% response rate). This study suggests that cash outperforms a gift certificate (consistent with the general incentives literature), and also points to the potential advantage of mail over e-mail invitations.

Alexander and colleagues (2008) conducted an incentive experiment for recruitment to an online health intervention. They tested a variety of different incentives in mailed invitations to potential

participants. Further, they found that a small prepaid incentive ($2) was cost-effective relative to larger promised incentives, even with enrollment rates in the single digits.

This brief review suggests that incentives seem to work for Web surveys in similar fashion to other modes of data collection, and for the same reasons. While it is impractical for access panels to send mail invitations with prepaid incentives when they are sending tens of thousands of invitations a day, the combination of an advance letter containing a small prepaid cash incentive, along with an e-mail invitation, may be most effective for list-based samples.

Again, there isn't (as yet) as much research on nonresponse in Web surveys as there has been in other modes of data collection. It may be that, because non-probability surveys dominate the Web survey world, nonresponse is of less concern. The market research world is focused on respondent engagement, which is more concerned with keeping respondents engaged in the survey once started (i.e., preventing breakoffs) than with getting them to start in the first place.

1.4 Correcting for Selection Biases

There are a number of different ways researchers attempt to correct for selection biases, both for probability-based and non-probability online surveys. In probability-based surveys, separate corrections can sometimes be made for coverage and nonresponse error, using different auxiliary variables. In non-probability surveys, this is often done in a single step, attempting to correct also for selection error, i.e., differences between the survey population (Internet users) and those who are recruited into the panel and selected to take part in the specific survey.

There are four key approaches to correcting for selection biases (see Kalton and Flores-Cervantes, 2003). These include:

1) Poststratification or weighting class adjustments
2) Raking or rim weighting
3) Generalized regression (GREG) modeling
4) Propensity score adjustment (PSA)

Several of these methods are closely related to each other. Both GREG and raking are special cases of calibration weighting. Post-stratification, in turn, is a special case of GREG weighting. All of the methods involve adjusting the weights assigned for the survey participants to make the sample line up more closely with population figures. I will not review these methods in detail, but rather provide brief commentary of the underlying assumptions and the challenges faced by non-probability surveys.

The first method that has been used to adjust for the sampling and coverage problems in Web surveys is known variously as ratio adjustment, post-stratification, or cell weighting. The procedure is quite simple — the weight for each respondent (typically, the inverse of the case's selection probability) in a weighting cell (or post-stratum) is multiplied by an adjustment factor:

$$w_{2ij} = \frac{N_i}{\sum_{n_i} w_{1ij}} w_{1ij} \ , \qquad (3)$$

in which w_{2ij} is the adjusted or post-stratified weight, w_{1ij} is the unadjusted weight, and the adjustment factor is the ratio between the population total for cell j (N_i) and the sum of the unadjusted weights for the respondents in that cell. For many Web surveys, the initial weights are all one, reflecting equal probabilities of selection. After adjustment, the weighted sample totals for each cell exactly match the population totals.

Post-stratification will eliminate the bias due to selection or coverage problems, provided that, within each adjustment cell, the probability that each case completes the survey is unrelated to that case's value on the survey variable of interest. This condition is sometimes referred to as the *missing at random* (MAR) assumption (Little and Rubin, 2002). In terms of Equation 1, a post-stratification adjustment will eliminate the bias if the within-cell covariance between the participation probabilities (p) and the survey variables (y) goes to zero:

$$Cov(p, y \mid \underline{X}) = 0$$

where $\underline{X}$ is the vector of categorical variables that are cross-classified to form the adjustment cells. This condition of zero covariance can be met is several ways: The participation probabilities can be identical within each cell; the values of the survey variable can be identical within each cell; or values for the two can vary independently within the cells. As a practical matter, post-stratification will reduce the magnitude of the bias whenever the absolute value of the within-cell covariance term is less than overall covariance term:

$$\left|Cov(p, y \mid \underline{X})\right| < \left|Cov(p, y)\right| \qquad (4)$$

Most survey statisticians use post-stratification in the belief that the inequality in Equation 4 holds, not that the bias disappears entirely.

Raking (or rim weighting) also adjusts the sample weights so that sample totals line up with external population figures, but the adjustment aligns the sample to the *marginal* totals for the auxiliary variables, not to the cell totals. Raking is preferred when population figures may not be available for every adjustment cell formed by crossing the auxiliary variables; or, there may be very few participants in a given cell so that the adjustment factors become extreme and highly variable across cells; or, the researchers may want to incorporate a large number of variables in the weighting scheme, too many for a cell-by-cell adjustment to be practical. Raking is carried out using iterative proportional fitting. Raking reduces or eliminates bias under the same conditions as post-stratification — that is, when the covariance between the probability of participation and the survey variable is reduced after the auxiliary variables are taken into account — but assumes a more stringent model, in which the interactions between the auxiliary variables can be ignored or bring only small additional reductions in bias.

Generalized regression (GREG) weighting is an alternative method of benchmarking sample estimates to the corresponding population figures. This approach assumes a "linear relationship between an analysis variable y and a set of covariates" (Dever, Rafferty, and Valliant, 2008). As with post-stratification and raking, GREG weighting eliminates the bias when the covariates remove any relationship between the likelihood of a respondent completing the survey and the survey variables of interest.

Another popular adjustment method — especially in non-probability settings — is propensity score adjustment (PSA) or propensity weighting. A number of papers have examined the use of propensity score adjustment to improve web survey estimates by reducing biases due to non-coverage or selection or both (Berrens et al., 2003; Dever, Rafferty, and Valliant, 2008; Lee, 2006; Lee and Valliant, 2009; Schonlau, van Soest, and Kapteyn, 2007; Schonlau et al., 2004; and Schonlau et al., 2009). A propensity score is the predicted probability that a case will end up in one group rather than another — for example, the probability that someone will be among those that have Internet access (versus not having access). The technique was originally introduced as a way of coping with confounds in observational studies between cases who got a given treatment and similar cases who did not (Rosenbaum and Rubin, 1984). Such confounds are likely to arise whenever there is non-random assignment of cases to groups as in non-experimental studies. Propensity score adjustment simultaneously corrects for the effects of multiple confounding variables on which the members of the two groups differ.

With Web surveys, the two groups are typically defined as the respondents to a Web survey (for example, the Web panel members who completed a specific Web questionnaire) and the respondents to a reference survey (for example, the respondents to an RDD survey conducted in parallel with the Web survey). The reference survey is assumed to have little or no coverage or selection bias so that it provides a useful benchmark to which the Web survey results can be adjusted (see Lee and Valliant, 2008, for a useful discussion of propensity weighting).

The first step in propensity weighting is to fit a model predicting the probability of membership in one of the groups. The usual procedure is to fit a logistic regression model:

$$\log(p(\underline{x})/(1-p(\underline{x})) = \alpha + \sum_{j}^{p} \beta_j x_j, \qquad (5)$$

in which $p(\underline{x})$ is the probability that the case will be in the group of interest (e.g., will complete the web survey), the x's are the covariates, α is an intercept term, and the β's are logistic regression coefficients. Next, cases are grouped (typically into quintiles) based on their predicted propensities, that is, their value for $\hat{p}(\underline{x})$. Finally, the existing weight (if any) for the case is adjusted by dividing by the predicted propensity of the case:

$$w_{2i} = \frac{w_{1i}}{\hat{p}_i(\underline{x})} \qquad (6)$$

If the cases have been grouped in propensity strata, then the mean (or harmonic mean) of the propensities in the stratum would be used in place of $\hat{p}_i(\underline{x})$ in the denominator of Equation 6. As Lee and Valliant (2008) point out, propensity adjustments work best when the logistic regression model includes predictors that are related to both the propensities and to the substantive variables (Little and Vartivarian, 2004, make the same point about post-stratification adjustments). Simulations by Lee and Valliant (2009) show that even when the reference sample completely covers the target population, propensity adjustments alone do not completely remove the coverage bias (see also Bethlehem, 2010).

While there are some variations in the variables used and how the models fit, all adjustment methods rely on some key assumptions. Key among these is the MAR assumption, that is, within the cells formed by cross-classifying the covariates (in the case of poststratification) or

conditional on the auxiliary variables included in the model (in the case of GREG and PSA), there is no relationship between the probability a given case will be in the respondent pool (i.e., is covered, selected, and respondents) and that case's value on the survey variable *y*. Clearly, the same adjustment may eliminate this bias for estimates based on some survey variables but not those based on others.

Propensity scoring goes further in that it assumes that all the information in the covariates is captured by the propensity score. This condition is often referred to as strong ignorability. For the bias to be eliminated by a propensity weighting model, then, conditional on the fitted propensities, a) the distribution of values of the survey variable must unrelated to what group the case came from (for example, the pool of web respondents versus the pool of respondents to the calibration survey) and b) the survey outcomes must be unrelated to the covariates. These conditions imply that

$Cov(p, y \mid \hat{p}(\underline{x})) = 0$.

A further drawback of PSA is that the variables used in the model must be measured in both the web survey sample and the calibration sample. I return to this issue below in a discussion of opt-in panels.

How effective are the adjustments at removing the bias? Regardless of which method of adjustment is used, the following general conclusions can be reached:

1) The adjustments remove only part of the bias.
2) The adjustments sometimes increase the biases relative to unadjusted estimates.
3) The relative biases that are left after adjustment are often substantial.
4) There are large differences across variables, with the adjustments sometimes removing the biases and other times making them much worse.

Overall, then, the adjustments seem to be useful but fallible corrections for the coverage and selection biases inherent in web samples, offering only a partial remedy for these problems.

Most of the focus on adjustment methods had been on the reduction of bias. When a relatively small reference survey (for example, a parallel RDD survey) is used to adjust the estimates from a large Web survey, the variance of the estimates is likely to be sharply increased (Bethlehem, 2010; Lee, 2006). This variance inflation is not just the byproduct of the increased variability of the weights, but reflects the inherent instability of the estimates from the reference survey.

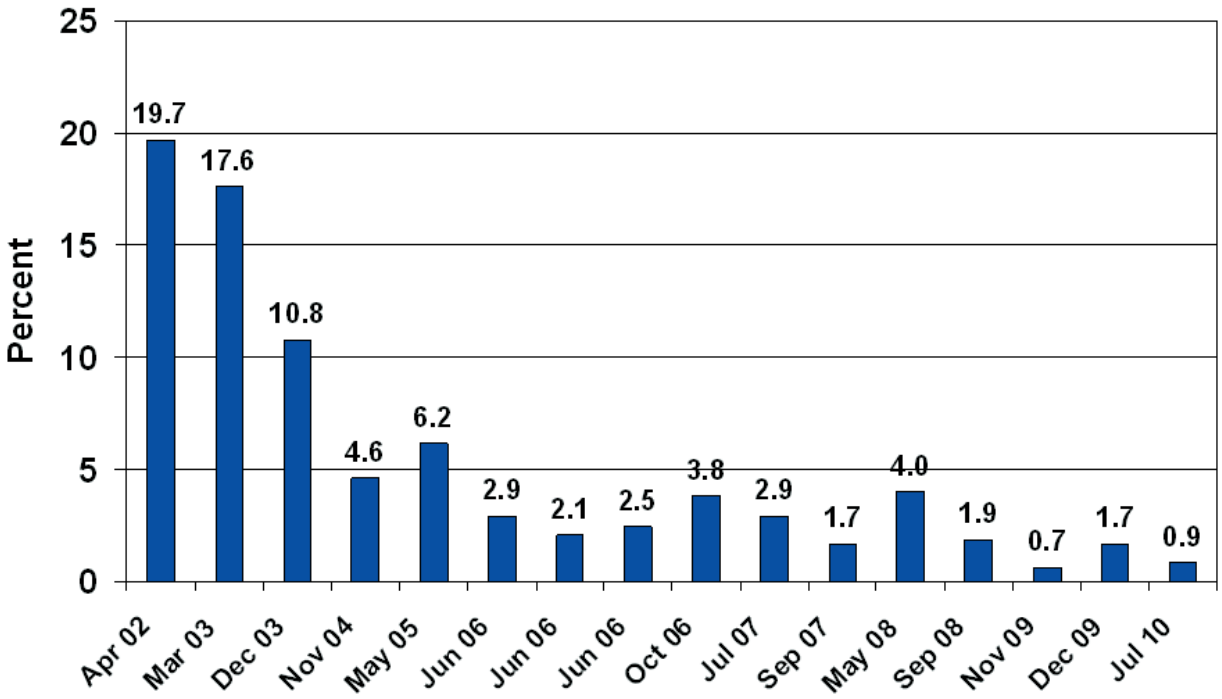
The general conclusion is that when the Web survey is based on a probability sample, nonresponse bias and, to a lesser extent, coverage bias, can be reduced through judicious use of postsurvey adjustment using appropriate auxiliary variables. However, when the estimates are based on a set of self-selected respondents, where the selection mechanism is unknown, and unlikely to be captured by a set of key demographic variables, the adjustments are likely to be more risky.

1.5 Online Access Panels

With the above discussion on issues of representation in mind, let's focus a little more attention on opt-in or volunteer access panels. These have been enormously popular in North America and Europe over the last decade, with scores of different panels competing for market share and for panelists in each country. The promise that these panels offer is a ready pool of potential respondents, many of whom have been pre-screened on key characteristics. For researchers who need a large number of respondents quickly and cheaply, but are less concerned about inference, these panels have provided a very valuable service. However, in recent years there have been increasing concerns about the quality of these panels. These concerns have been manifested in several different ways.

First, there is growing evidence of over-saturation of these panels, with the demand (both the number of surveys and the number of respondents per survey) outstripping supply (the number of panelists who complete surveys). This can be seen in the declining participation rates of panelists, and in the increasing number of invitations panelists receive. Data on this issue is not made available by the panel vendors, so this is hard to assess. But, we have been using the same vendor for our experiments on Web survey design (see Part 2) for several years. The participation rates (the number of registered panelists who complete a particular survey) have been steadily declining, as seen in Figure 2.

Figure 2. Participation Rates for Comparable Samples from Same Vendor



To provide one concrete example, for our survey experiment conducted in July 2010, over 138,000 panelists were invited to obtain a set of 1,200 respondents. This represents a significant fraction of the panel. These participation rates vary widely across different vendors, in part

because of different practices in maintaining the panel, especially with regard to inactive members. In a 2005 report, ComScore Networks claimed that 30% of all online surveys were completed by less than 0.25% of the population, and these panelists completed an average of 80 surveys in 90 days. This estimate is likely to be high as the source of the data was itself a volunteer panel in which members had agreed to have their Internet activities monitored. However, a 2006 study among 19 online panels in the Netherlands (Vonk, Willems, and van Ossenbruggen, 2006), found that 62% of respondents reported belonging to more than one panel, with the average being 2.73 panels per respondent. A small group (3% of respondents) reported belonging to 10 or more panels.

Another piece of evidence related to over-saturation comes from a US panel I've been a member of for several years. In response to ESOMAR's (2008) 26 questions, the vendor claimed an average response (participation) rate of 35%-40% (contrast this with the rates in Figure 2). The vendor also stated that panelists are contacted 3-4 times a month to participate in surveys. However over the past 5 years, I have received an average of 43.3 unique invitations (excluding reminders) per month, ranging from an average of 30 per month in 2007 to an average of 63 per month in 2010.

Related to the issue of over-saturation is the rising concern among panel vendors about "professional respondents" — those who do so many surveys that they may not be paying attention to the questions, instead speeding through the survey to get the rewards (incentives or points) for participation. One estimate is that about 7% of respondents are "deliberately doing a poor job" (Giacobbe and Pettit, 2010). This is manifested in problems such as over-qualifying (e.g., saying "yes" to all screener questions to qualify for a survey), duplication or "hyperactives" (e.g., belonging to the same panel under different guises, or belonging to multiple panels), speeding (answering too fast to have read the question), or inattention (e.g., straightlining in grids, inconsistency across repeated questions, failing a specific instruction to select a response). In one study reported by Downes-Le Guin (2005), 13% of respondents claimed to own a Segway and 34% failed a request to check a specific response (second from the left) in a grid. Given this, panel vendors are developing increasingly sophisticated methods of identifying and dealing with such professional respondents (e.g., Cardador, 2010). However, recent research done by the industry itself (see the ARF's Foundations of Quality Study; http://www.thearf.org/assets/orqc-initiative) suggests that the problem might not be as bad as claimed. Nonetheless, the issue continues to raise concern for users of online panels as well as for vendors.

In the last few years, several large buyers of online market research have raised questions about the replicability or reliability of the results from opt-in panels. For example, in 2005, Jeff Hunter, Consumer Insights Director at General Mills, delivered a keynote address at the ESOMAR Worldwide Panel Research Conference in which he described a concept test where the same survey was administered to different samples from the same panel, and produced substantially different results on whether to launch the product. Similarly, Kim Dedeker, VP of Global Consumer and Market Knowledge at Procter & Gamble (one of the largest buyers of market research in the world), gave a talk in which she described situations where online concept tests identified a strong concept. A large investment was then made to launch the products, but later concept tests got disappointing results. She noted that "Online research … is the primary driver

behind the lack of representation in online testing. Two of the biggest issues are the samples do not accurately represent the market, and professional respondents."

In response to these rising concerns, the Advertising Research Foundation (ARF) launched its own study in cooperation with panel vendors. The Foundations of Quality (FOQ) project was designed to address the following key questions: 1) Why do online results vary? 2) How much panel overlap is there, and how much does this affect results? 3) What are the effects of "bad" survey-taking behavior? Part of the design involved a 15-minute online survey administered to members of 17 different US panels. The full report is available at a high price, but ARF press releases claim that the problems of panel overlap are not as bad as others have argued.

The American Association for Public Opinion Research (AAPOR) created a task force to review online panels (see AAPOR, 2010). The task force made several recommendations regarding such panels, some of which are as follows:

1) Researchers should avoid nonprobability online panels when a key research objective is to accurately estimate population values … claims of "representativeness" should be avoided when using these sample sources
2) There are times when a nonprobability online panel is an appropriate choice
3) There are significant differences in the composition and practices of individual panels that can affect survey results
4) Panel vendors must disclose their methods

Despite their inferential limitations, opt-in panels have a number of uses. They can provide data quickly and cheaply. They are useful for identifying and surveying a set of subjects with known characteristics or behaviors, based on extensive screening data that are often available. Some other examples of the uses of such panels include: 1) pretesting of survey instruments, 2) testing new concepts, theories, or measurement, 3) methodological or substantive experiments (where volunteer bias is not a concern), 4) trend analysis (assuming a stable panel population), and possibly 5) correlational analysis (although selection bias is still a concern). It is recommended that opt-in or access panels should not be used as the sole source of data, but that they should be used in combination with some other methods.

While many online panel vendors make claims of comparability to national estimates, there are only a few independent studies examining this issue. For example, Yeager et al. (2009) compared an RDD telephone survey with a probability-based Internet survey and 7 non-probability Internet surveys (6 access panels and 1 river sample) in 2004, with an average sample size of around 1,200 respondents. They compared survey estimates to known benchmarks from large federal surveys in the US. They found that the probability sample surveys done by telephone or Web were consistently highly accurate across a set of demographic and non-demographic variables, especially after post-stratification with primary demographics. Further, non-probability sample surveys done via the Internet were always less accurate, on average, than the probability sample surveys, and were less consistent in their level of accuracy. Finally, post-stratification using demographic variables sometimes improved the accuracy of non-probability sample surveys and sometimes reduced their accuracy.

In a later analysis using other items from the 7 non-probability panels in the same study, Yeager et al. (2010) reported large differences in levels of reported consumption of a variety of consumer products (e.g., 43% to 74% reporting consumption of Coke) across the non-probability panels, but the rank orders of reported consumption were relatively stable. Generally, the associations between variables were consistent across panels, suggesting less volatility than others have suggested. However, some key conclusions did not replicate across panels, and sometimes the differences were very large (e.g., up to 45 percentage point differences between pairs of surveys on selected consumer categories. Furthermore, it is difficult to predict when these large differences will occur — that is, it is not always the same pair of panels that produced the same magnitude or direction of differences.

In their analysis of the aggregated results from 19 different online panels in the Netherlands, Vonk, Willems, and van Ossenbruggen (2006) found several large differences compared to official data from Statistics Netherlands (CBS). For example, 77% of panelists reported not belonging to a church (compared to 64% from CBS), 29% reported supporting the CDA party (compared to 16% from CBS), and 2% of panelists were identified as foreigners living in large cities (compared to the official estimate of 30%).

Self-selection may not only affect point estimates, but also correlations (and, by extension, coefficients from regression and other models). For example, Faas and Schoen (2006) compared three different surveys in Germany prior to the 2002 Federal election: a face-to-face survey, an online access panel and open-access Web survey. They concluded that "…open online surveys do not yield results representative for online users (either in terms of marginal distributions or in terms of associations)" (Faas and Schoen, 2006, p. 187). They further noted that weighting adjustments did not help to reduce the bias in the online polls.

Loosveldt and Sonck (2008) compared data from a Belgian access panel to data from the European Social Survey (ESS). They compared unweighted, demographically weighted, and propensity-weighted estimates from the panel. They found significant differences in responses on several different themes, including political attitudes, work satisfaction, and attitudes towards immigrants. They also found that post-stratification adjustment based on demographics had no substantial impact on the bias in the estimates. Further, propensity score adjustment had only a minimal effect, with some differences becoming larger rather than smaller.

How do opt-in or access panels deal with the inferential issues? Many panel vendors provide the data "as is," without any attempt at adjustment, leaving the user to draw inferences about the representativeness of the data. A few use some form of poststratification or raking adjustment to match the set of panel respondents to the broader population on key demographic variables. The use of propensity score adjustment or similar strategies (e.g., matching) is rare. One panel provider in the US, Harris Interactive, has promoted the use of PSA for general population inference.

The Harris Interactive approach to PSA is as follows (see Terhanian et al., 2000, 2001):

- Ask a small set of "Webographic" questions in all online panel surveys.
- Ask the same set of questions in occasional RDD telephone surveys.

- Use these common variables to predict the likelihood of being in the Web or RDD group, using a logistic regression model.
- Use the predicted values (propensities) to adjust the Web responses either directly (using propensity scores) or indirectly (by creating weighting classes based on the scores). Typically, respondents to both surveys are sorted into 5 bins (quintiles) based on propensity scores.
- Assign weights such that the Web survey's (weighted) proportion of respondents in each bin matches the reference (telephone) survey's proportion.

Several key assumptions need to be met for this approach to be successful at eliminating selection bias. The first is that the questions asked in both surveys capture the full range of differences in selection into the two samples (i.e., the selection mechanism is MAR or ignorable conditional on these variables). While Harris Interactive does not disclose the items used, examples of "Webographic" questions have included the frequency of watching the news on TV, frequency of vigorous physical activity, ownership of a non-retirement investment account, and whether a variety of items are considered invasions of privacy. A second assumption is that there is no measurement error, i.e., that the same answers would be obtained to these questions regardless of the mode (telephone or Web) in which they are asked. Third, in using the telephone survey as a population benchmark, the PSA ignores selection bias in the telephone survey. With response rates to RDD surveys as low as 20%, and concerns about coverage of cell phone only households, this is a serious concern.

In addition, as Bethlehem (2010) has noted, the variance of the resultant estimator should take into account the fact that the RDD benchmark survey is itself subject to a high level of variation, depending on sample size. Some users of PSA treat these as population values (in terms of both bias and variance), ignoring this uncertainty. According to Bethlehem (2010), the variance of the post-stratification estimator for an Internet (*I*) sample weighted using a reference sample (*RS*) is:

$$V(\bar{y}_I) = \frac{1}{m}\sum_{h=1}^{L} W_h(\bar{Y}_I^{(h)} - \tilde{Y})^2 + \frac{1}{m}\sum_{h=1}^{L} W_h(1-W_h)V(\bar{y}_I^{(h)}) + \sum_{h=1}^{L} W_h^2 V(\bar{y}_I^{(h)}) \qquad (7)$$

Where $\bar{y}_I^{(h)}$ is the Web survey estimate for the mean of stratum *h*, and m^h / m is the relative sample size in stratum *h* for the reference sample. Thus, the first term in Equation 7 will be of the order 1/*m*, the second term of order 1/*mn*, and the third of order 1/*n*, where *n* is the Web sample size and *m* the reference sample size. As Bethlehem (2010) notes, *n* will generally be much larger than *m* in most situations, so the first term in the variance will dominate; that is, the small size of the reference survey will have a big influence on the reliability of the estimates. While Duffy et al. (2005) and Börsch-Supan et al. (2004) both acknowledge that the design effects of propensity score weighting will significantly reduce the effective sample size, this issue appears to have been largely ignored by those using PSA in practice.

Another approach to the inferential challenges of volunteer online panels has been to use sample matching techniques, the approach advocated by YouGov Polimetrix (see Rivers, 2006, 2007; Rivers and Bailey, 2009). Here, a target sample is selected from the sampling frame representing the population to which one wants to make inference. However, instead of attempting to interview that sample, a matched sample is selected from a pool of available respondents (e.g., from an online panel) and those are interviewed. As with post-survey

adjustment approaches, the success of this method in eliminating or reducing bias relies on the matching variables used (i.e., it uses an MAR assumption, conditional on the matching variables). Given that the data available for the target population often only consists of demographic variables, the model assumes that controlling for such demographic differences will eliminate bias on other variables measured in the survey. As shown earlier, such assumptions do not eliminate bias in all circumstances. While model-assisted sampling (see Särndal, Swensson, and Wretman, 1992) is gaining in popularity, and all adjustment procedures rely on model assumptions (see Section 1.4), fully model-based approaches require a greater reliance on the model to be effective in reducing bias. To the extent that the model does not accurately reflect the effect of the selection process on the variables or statistics of interest, procedures like sample matching and propensity score adjustment are likely to have varying success in minimizing or eliminating bias. Design-based approaches (like those based on traditional sampling theory) can protect against failures in the model assumptions.

In general, most opt-in or access panels appear to be focusing more on the measurement concerns than on the inferential concerns. Attention is focused on reducing duplicate of fraudulent respondents, identifying and removing professional or inattentive respondents, and designing surveys to increase respondent engagement. It seems clear that the inferential issues are being largely ignored. This may be a result of the opt-in panels turning away from early attempts to make population projections, such as with regard to pre-election polls (e.g., Taylor et al., 2001), and focusing more on market research applications, where the pressure to produce accurate estimates may be less strong, and the prediction failures receive less media attention.

The general recommendation is that when using opt-in or access panels, one should avoid making inferential claims beyond what can be supported by the data. While there may be some situations where the estimates from such panels appear to reliable (e.g., in pre-election polls), this cannot be generalized to all situations. In other words, while these panels have a wide variety of uses, broad population representation on a wide range of topics is likely not one of them.

1.6 Web Surveys as Part of Mixed-Mode Data Collection

Given the inferential challenges facing Web surveys discussed above, National Statistical Offices (NSOs) and researchers concerned with broad population representation are increasingly turning to mixed-mode surveys involving Web data collection in combination with other modes. The hope is that by combining modes, the weakness of one mode (e.g., the coverage concerns and lack of a sampling frame for Web surveys) can be compensated for by using other modes.

The combination of Web surveys with mail surveys has received the most attention in recent years. These two modes share similar measurement error properties, and mail is a logical method for inviting people to Web surveys. There are two main approaches to this mode combination, concurrent mixed-mode designs and sequential mixed-mode designs. Concurrent designs send a paper questionnaire to sample persons or households, but provide them with the opportunity to complete the survey online. However, several early studies have found that providing

respondents such a choice does not increase response rates, and may in fact result in *lower* response rates than the mail-only approach. For example, Griffin, Fischer, and Morgan (2001) reported a 37.8% response rate for the American Community Survey (ACS) with a Web option, compared to 43.6% for mail only. This early result led the US Census Bureau to be cautious about offering an online option for the 2000 decennial census, and no such option was available in the 2010 census. In studies of students in Sweden, Werner (2005) reported lower response rates (62%-64%) for the mail+Web versions than for the mail-only control version (66%). Brennan (2005) used a sample from the New Zealand electoral register, and obtained lower response rates for the mail+Web option (25.4%) than for the mail-only design (40.0%), and Brøgger et al. (2007) obtained similar results (44.8% for mail+Web versus 46.7% for mail-only) in a survey among adults age 20-40 in Norway. Gentry and Good (2008) reported response rates of 56.4% for those offered an eDiary option for radio listening, compared to 60.6% for those offered only the paper diary. Other studies (e.g., Tourkin et al., 2005; Hoffer et al., 2006; Israel, 2009; Cantor et al., 2010; Smyth et al., 2010; Millar and Dillman, 2011) have also found disappointing results for the concurrent mixed-mode approach. A number of hypotheses are being advanced for these results, and research is continuing on ways to optimize designs to encourage Web response while not negatively affecting overall response rates.

More recent studies have focused on sequential mixed-mode designs, where samples members are directed to one mode initially, rather than being given a choice, and nonrespondents are followed up in another mode. One example is the study of adults in Stockholm by Holmberg, Lorenc, and Werner (2010). They compared several different sequential strategies involving mail and Web. While overall response rates did not differ significantly across the five experimental conditions, they found that the proportion of respondents completing the survey online increased as that option was pushed more heavily in a sequential design. For example, when the first two mail contacts (following the prenotification or advance letter) mentioned only the Web option, and the mail questionnaire was provided only at the third contact, the overall response rate was 73.3%, with 47.4% of the sample using the Web. In contrast, in the condition where the mail questionnaire was provided in the first contact, the possibility of a Web option was mentioned in the second (reminder) contact, and the login for the Web survey was not provided until the third contact (along with a replacement questionnaire), the overall response rate was 74.8%, but only 1.9% of the sample completed the Web version. Millar and Dillman (2011) report similar findings for a mail "push" versus Web "push" approach. While none of the sequential mixed-mode designs show substantial increases in overall response rates, the increased proportion of responses obtained via the Web represents a potential cost saving that could be directed at additional follow-up in other modes.

Despite these somewhat disappointing results, a growing number of NSOs are providing an Internet option for census returns with apparent success. For example, Singapore reported that about 15% of census forms were completed online in the 2000 population census. In the Norwegian census of 2001, about 9.9% of responses were reportedly obtained via the Web. Statistics Canada reported that 18.3% of Canadian households completed their census form online in 2006, and this increased to 54.4% in the recently-completed 2011 census. Preliminary estimates from the Australian census in August 2011 suggest a 27% uptake of the Internet option, while South Korea anticipates that 30% of forms will be completed online for their

census in November 2011. The United Kingdom also recently completed its census (in March 2011) and heavily promoted the online option, but the success of this effort is not yet known.

The success of these census efforts may suggest that the length of the form or questionnaire may be a factor in whether it is completed online or not. In addition, censuses tend to be heavily-promoted public events and this may play a role in the successful outcome. Much more research is needed into the conditions under which mixed-mode designs involving mail and Web will yield improvements in response rate — and reductions in nonresponse bias.

In addition, a key assumption underlying this strategy is that the measurement error differences between the modes are not large — or at least not large enough to negate the benefits of mixing modes. The primary focus thus far has been on response rates, with much less attention paid to measurement differences between the modes. But this suggests that the mail-with-Web-option strategy may be most effective when the survey is very short and measures demographic variables that are les likely to be affected by mode.

1.7 Summary on Inferential Issues

As has been seen in this section, inferential issues remain a challenge for Web surveys aimed at broad population representation. Sampling frames of e-mail addresses or lists of Internet users in the general population do not exist. While the proportion of the population without Internet access has been declining, there remain substantial differences between those with access and those without on a variety of topics. Nonresponse also remains a challenge for Web surveys relative to other (more expensive) modes of data collection. Statistical adjustments may reduce the bias of self-selection in some cases, but substantial biases may remain.

Nonetheless, there remain a number of areas where Web surveys are appropriate. For example, surveys of college students and members of professional associations are ideally suited to Web data collection. Establishment or business surveys may also benefit from online data collection, especially as part of a mixed-mode strategy. There are a number of creative ways to address these challenges (such as the development of probability-based access panels), but for now at least, Web surveys are likely to supplement rather than replace other modes of data collection for large-scale surveys of the general public where high levels of accuracy and reliability are required.

Part 2: Interface Design

This second part of the seminar focuses on the design of Web survey instruments and data collection procedures, with a view to minimizing measurement error or maximizing data quality. The particular focus is on those aspects unique to Web surveys. For example, question wording is an issue relevant for all modes of data collection, and is not a focus of this seminar. Further, I will not address technical issues of Web survey implementation, such as hardware, software or programming. Given the unique features of Web surveys, there are many challenges and opportunities for survey designers. The seminar is not intended to be an exhaustive review of the topic, but rather to provide empirical evidence on illustrative examples to emphasize the importance of careful design in developing and implementing Web surveys. Couper (2008a) goes into these — and other — deign issues in more depth.

2.1 Measurement Error

Measurement error involves a different type of inference to that discussed in Part 1 above, that is from a particular observation or measurement from the *i*-th respondent (y_i) to the "true value" for that measure for that respondent (μ_i), sometimes measured across several trials (*t*). The simplest expression of measurement error is as follows:

$$y_{it} = \mu_i - \varepsilon_{it} \qquad (8)$$

where ε_{it} is the error term for respondent *i* and trial *t*. In order to estimate measurement error using this expression, we need to know the true value. In practice, the true value is rarely known. Researchers tend to rely on alternative approaches to examine measurement error properties of a mode or a design. One common approach is to examine differences in responses to alternative presentations of the same questions. The measurement error model applicable to this approach is as follows:

$$y_{ij} = \mu_i + M_{ij} + \varepsilon_{ij} \qquad (9)$$

where y_{ij} is the response for the *i*-th person using the *j*-th form of the question or instrument, and M_{ij} is the effect on the response of the *i*-th person using the *j*-th method. The classic split-ballot experiments to examine question wording effects (e.g., Schuman and Presser, 1981) are examples of this approach.

One of the advantages of Web surveys lies in the ease with which randomization can be implemented, giving researchers a powerful tool to explore measurement effects. This has led to a large number of experimental comparisons of different design options. In such Web design studies, indirect measures of data quality or measurement error are often used, involving not only an examination of response distributions across versions, but also other indicators such as missing data rates, breakoff rates (potentially leading to increased nonresponse error), speed of completion, and subjective reactions by respondents. Together these all point to potential comparative advantages of one particular design approach relative to another, without directly assessing the measurement error. So, in this part, the focus is more on the measurement process than on measurement error.

2.2 Measurement Features of Web Surveys

Web surveys have several features or characteristics that have implications for the design of survey instruments, and hence for measurement error. By themselves, each of these characteristics is not unique to Web surveys, but in combination they present both opportunities and challenges for the survey designer.

First, Web surveys are self-administered. In this they share attributes of paper self-administered questionnaires (e.g., mail surveys) and computerized self-administered questionnaires (e.g., computer-assisted self-interviewing [CASI] or interactive voice response [IVR]). While this attribute also has implications for sampling, coverage, and nonresponse error, our focus here is on measurement error. Self-administration has long been shown to be advantageous in terms of reducing effects related to the presence of the interviewer, such as social desirability biases. At the same time, the benefits of interviewer presence — such as in motivating respondents, probing, or clarifying — are also absent. From an instrument design perspective, this means that the instrument itself must serve these functions. It must also be easy enough for untrained or inexperienced survey-takers to complete.

Second, Web surveys are computerized. Like computer-assisted personal interviewing (CAPI) and computer assisted telephone interviewing (CATI), but unlike paper surveys, computerization brings a full range of advanced features to bear on the design of the instrument. Randomization (of question order, response order, question wording or format, etc.) is relatively easy to implement in Web surveys. Other aspects of computer-assisted interviewing (CAI) that are easy to include in Web surveys but relatively hard in paper surveys include automated routing (conditional questions), edit checks, fills (inserting information from prior answers in the current question), and so on. Web surveys can be highly customized to each individual respondent, based on information available on the sampling frame, information collected in a prior wave

(known as dependent interviewing), or information collected in the course of the survey. This permits the use of complex instruments and approaches such as computerized adaptive testing and conjoint methods, among others. However, adding such complexity increases the need for testing, increases the chances of errors, and makes careful specification and testing of the instrument all the more important.

A third feature of Web surveys, and related to their computerized nature, is that they can be designed with varying degrees of interactivity. Conditional routing is one form of interactivity. But in this respect, Web surveys can be designed to behave more like interviewer-administered surveys, for example, prompting for missing data, seeking clarification of unclear responses, providing feedback, and the like.

A fourth characteristic of Web surveys is that they are distributed. In CAI surveys, the technology is in the hands of the interviewer, using hardware and software controlled by the survey organization. This means that the designer has control over the look and feel of the survey instruments. In contrast, the Web survey designer has little control over the browser used by the respondent to access and complete the survey, or the hardware used. Increasingly, people are accessing the Web using a variety of mobile devices (such as smart phones or tablets), and these present new challenges for the designer. But Web surveys can vary in many other ways too — from the user's control over the size of the browser, to the security settings that may affect whether and how JavaScript, Flash, or other enhancements work, to the connection type that determines the speed with which respondents can download or upload information online. Hypertext markup language (HTML) is standard across browsers and platforms, but JavaScript (for example) does not always behave in an identical manner across different operating systems. While these variations give the respondent great flexibility in terms of how, when, and where they access the survey instrument, it presents design challenges in term of ensuring a consistent look and feel for all respondents in the survey.

Finally, a feature of Web surveys that has already been widely exploited by Web survey designers is that it is a visually rich medium. It is true that other modes are visual too — for example, pictures or images have been used in paper surveys. But it is the ease with which visual elements can be introduced in Web surveys that makes them distinctive as a data collection mode. The visual nature of web surveys means much more than just pictures. Visual elements include colors, shapes, symbols, drawings, images, photographs, and videos. The cost of adding a full-color image to a Web survey is trivial. Visual in this sense extends beyond the words appearing on the Web page, and can extend to full multimedia presentation, using both sound and video. The visual richness of the medium brings many opportunities to enhance and extend survey measurement, and is one of the most exciting features of Web surveys. On the other hand, the broad array of visual enhancements also brings the risk of affecting measurement in ways not yet fully understood.

Together these characteristics make the design of Web surveys more important than in many other modes of data collection. As already noted, a great deal of research has already been conducted on alternative designs for Web surveys, and such research is continuing. It is not possible to summarize this vast literature here. Rather, I will present a few selected examples of key design issues to illustrate the importance of design for optimizing data quality and

respondent engagement in Web surveys. The interested reader is encouraged to read Couper (2008a) for more information.

2.3 Paging versus Scrolling Design

One of the early choices a Web survey designer needs to make is whether to present all questions in a single scrolling Web page, or present each question on its own page. There are many other alternatives along this broad continuum. Sometimes this decision is constrained by the software used. But when the designer has a choice, the decision should be determined by the objectives of the survey and the content of the questions.

Despite the importance of this decision for a variety of other design elements I discuss later, there is relatively little research on the topic. A few small-scale comparisons (e.g., Burris et al., 2001; Clark and Nyiri, 2001; Nyiri and Clark, 2003; Vehovar, Lozar Manfreda, and Batagelj, 2001) yielded inconclusive results, mostly because of small sample sizes. Only one study has explored the design choice with sufficient sample size to detect differences. Peytchev et al. (2006) compared a scrolling design (with the survey divided into 5 sections) with a paging design (with one or more questions on a page) for a survey on drug and alcohol use among college students. They found no differences in unit nonresponse, partial nonresponse (breakoffs), and nonsubstantive responses (explicit refusals or "don't know" responses), and the response distributions and key associations did not differ between versions. However, they did find significantly more item missing data and longer completion times in the scrolling version, which they attribute to the enhanced control over navigation and routing offered by the paging design. Examples of the scrolling and paging designs from Peytchev et al. (2006) are presented in Figures 3 and 4.

Despite the lack of empirical evidence to date, scrolling designs are preferred in practice for short surveys with few skips or edits (such as customer satisfaction surveys) or in mixed-mode designs where the goal is to replicate the paper questionnaire. Paging designs appear to be much more common, and give the designer control over the order in which questions are presented, automating skips and routing, permitting edit checks at the point of occurrence, and so on. In other words, the more complex and customized the instrument is, and the more the designer wants to control the interaction, the more likely a paging design will be employed. This is just one example of many design issues where the decisions one makes about a particular design approach should be informed by research.

Figure 3. Scrolling Survey Example

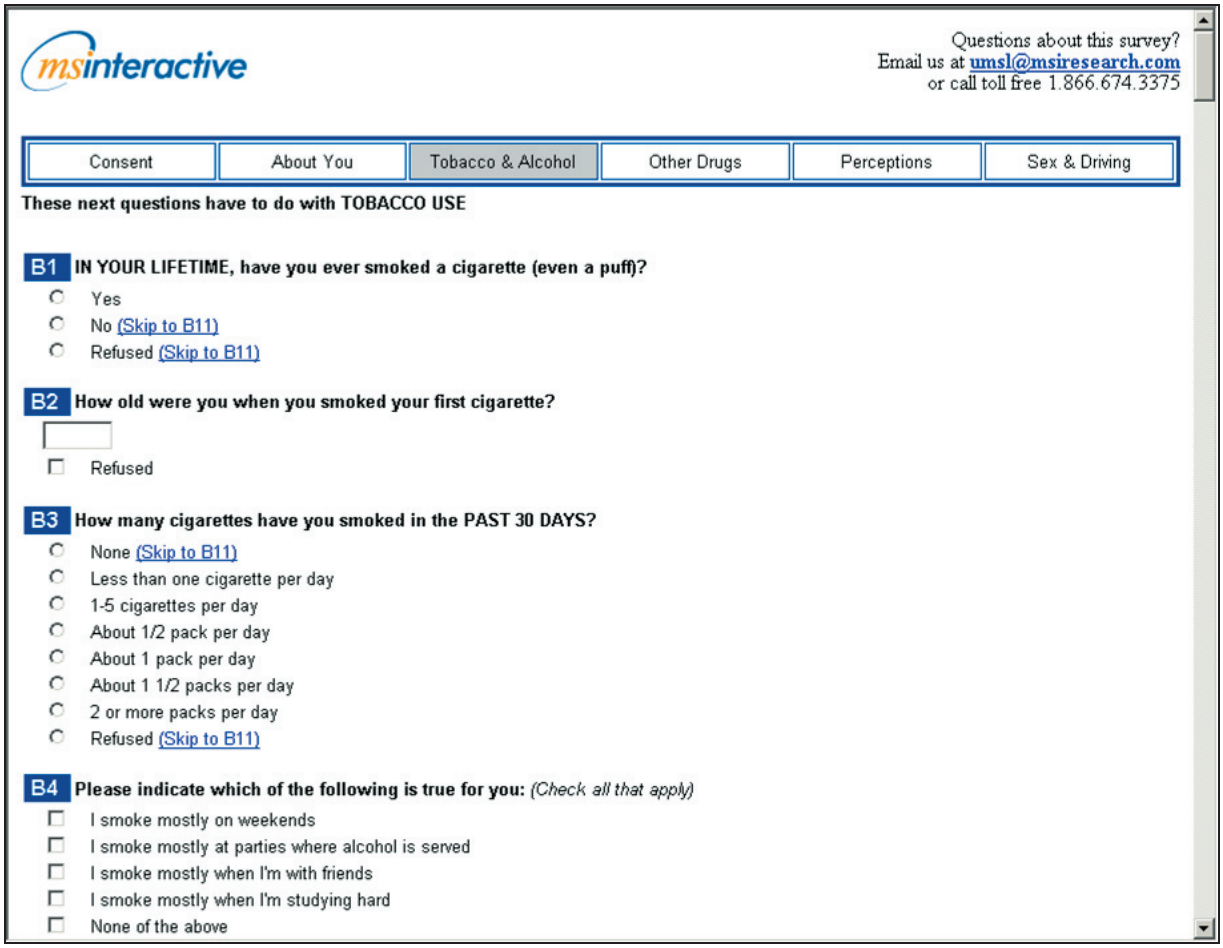
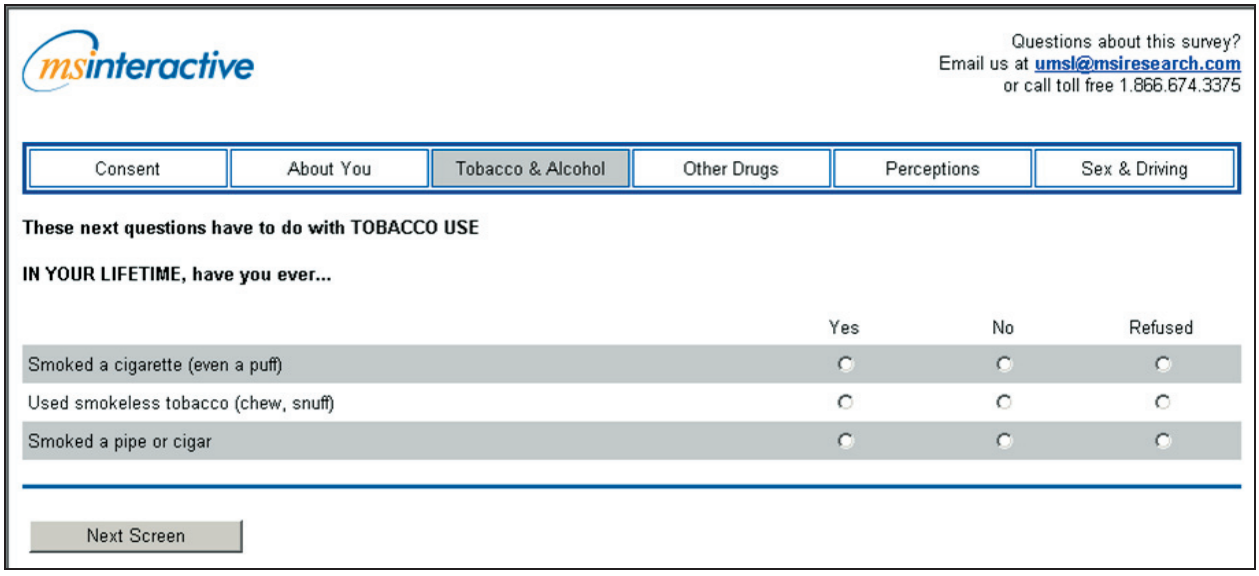


Figure 4. Paging Survey Example



2.4 Choice of Response or Input Formats

Another important design decision, but one that affects individual items rather than the entire questionnaire, relates to the choice of input tools to capture respondents' answers. In contrast to paper where the response options place no constraint on the respondent, the choice of input tool in Web surveys may serve as both a visual guide as to how to respond and as a forcing function (see Couper, 2008b). For example, radio buttons permit the selection of one and only one response. Drop boxes or select lists similarly constrain the selection to one of the available options (unless programmed to permit multiple selections). This places more responsibility on the designer to ensure that all possible options are available to the respondent and that the design does not unnecessarily limit the options a respondent can choose.

The set of input formats available to the designer may be limited to those available in HTML and those that can be created using active scripting such as JavaScript, Java, or Flash. Within HTML, the following major choices are available:

- Radio buttons: commonly used for single-response questions.
- Check boxes: used for multiple-response or check-all-that-apply items.
- Drop boxes or select lists: allow respondents to make single or multiple selections from a long scrolling list.
- Text fields and text areas: allow respondents to enter unformatted text or numeric responses.

Using active scripting, the type of input tools one can create is almost limitless, including visual analog scales (VAS) or slider bars, drag-and-drop tools, card sort methods, map-based input, and so on. The extent to which these interactive features add value — either for the survey researcher in terms of the quality of measurement, or for the respondent in terms of enjoyment or flow — is still a matter of debate, and in need of more empirical evidence.

There are two key decisions the designer must make with rest to input formats. The first is the choice of the appropriate tool for a particular question; the second is how best to design that tool to facilitate the task of providing a response. I often see examples of Web surveys where the input tool does not match the intended task. There are a few studies examining the effects of the type of selection or input tool on the number, quality, and completeness of responses, and these could guide practical design decisions.

Couper, Traugott, and Lamias (2001) contrasted radio buttons versus text boxes for a series of questions related to the race/ethnicity of one's friends, classmates, etc., where the task was to enter a number from 1 to 10 for each of 5 groups. They found that text boxes permitted respondents to give invalid (out-of-range) responses, while the radio button version prevented respondents from doing so, yielding significant differences in both the proportion of invalid responses and missing data (including "don't know" or "not applicable" responses). Further, the size of the text box also affected responses, with a longer box eliciting more invalid responses than the short box. However, the text boxes made it easier for respondents to make the five numbers sum to 10 (as instructed). Thus, while the text entry box made it easier for respondent to avoid answering the question or providing uncodable or out-of-range responses, the radio button

format made it harder for respondents to make the numbers add up as requested. In part, the text boxes revealed the difficulty of the underlying task, which was not as apparent in the radio button version.

Heerwegh and Loosveldt (2002) compared radio buttons versus drop boxes for a series of questions. They found no effect of format on completion rates, on nonsubstantive responses, or on missing data. They did not examine the substantive distributions. However, they did find that radio buttons required more time to download (affecting those with slower Internet connections) while drop boxes were harder to use. The download speed is likely to be less of an issue today.

Healey (2007) similarly tested radio buttons versus drop boxes on a series of questions across several pages of the survey. He found that format choice did not significantly affect survey completions (breakoff rates), the number of nonsubstantive answers, or overall completion time. Again, Healey did not examine substantive distributions between versions. However, drop boxes led to slightly higher item nonresponse and longer response times per item. Furthermore, those who used the scroll mouse to complete the survey (about 76% of respondents did so), were prone to accidentally changing an answer in the drop box condition (see also Lebrasseur et al., 2010).

Couper and colleagues (2004) examined response order effects using three different formats: a series of radio buttons, a drop box, and a scroll box (with a partial list of responses visible). They found that the magnitude of the response order effects depended on the format used to display the items, with significantly bigger order effects in the scroll box version. This suggests that while serial reading order produces primacy effects in visually-presented measures, the effect is exacerbated when the respondent is required to use the scroll bar to see the items that are not initially displayed.

Couper et al. (2006) compared radio buttons and text boxes (along with visual analog scales) for a series of eight attitude items with a 21-point response scale. They found more missing data for the text box version (an average of 2.7% missing across the 8 items) than for the radio button version (a missing rate of 1.0%). Average completion time was 125 seconds for the radio button version and 158 for the text box version.

Finally, Hennessy (2002) conducted a series of usability evaluations of alternative input formats. She found that respondents preferred radio buttons over drop boxes because they required fewer mouse clicks and all options were visible from the start. Together these studies suggest that the choice of input tool may affect the answers obtained in Web surveys. More research has focused on the design of different input tools (particularly text fields), and we explore this in more detail below.

2.5 The Design of Input Fields

Radio button and check box inputs can be designed using the standard HTML form elements or graphical equivalents. While the latter are popular in market research, I know of no research on these alternatives. Text input gives the designer much more flexibility, and there are a number of

studies examining the effect of text box design on the quality and completeness of the answers obtained. There are two main types of text input. Text areas are large open boxes that expand up to 32,000 characters as information is typed. Text fields are more constrained, where limits can be placed both on the size of the box displayed, and on the number of characters accepted.

Text boxes are often used to elicit open-ended responses. There are a variety of types of open-ended responses, and they may each require different types of design. Some of the types are as follows (see Couper et al., 2010):

- Narrative responses, e.g., "What is the biggest problem facing the country?" HTML text areas are most appropriate for such questions.
- Short verbal responses with length constraint but no formatting constraint, e.g., "What kind of work were you doing last week? (For example: registered nurse, personnel manager, supervisor of order department, secretary, or accountant)." HTML text fields are most appropriate.
- Single word/phrase verbal responses, e.g., country of origin, medical diagnosis, etc. An HTML text field or HTML or JavaScript drop box (select list) could be used.
- Frequency or numeric responses, e.g., "During the past 12 month, how many times have you seen or talked with a doctor about your health?" Other examples include probability ratings, feeling thermometers, etc. Several alternative inputs are possible, including text fields, radio buttons, drop boxes, or visual analog scales.
- Formatted numeric or verbal response, e.g., date of birth, telephone number, currency amounts. Input alternatives include one or more text fields and/or drop boxes.

The research focuses both on the choice of input tool (where acceptable alternatives exist), but also on the design of the fields, especially in the use of masks or templates to guide input in the correct format (see below).

A couple of studies have examined the effect of the length of the input field on narrative responses. Dennis, deRouvray, and Couper (2000) found that longer fields encouraged longer responses. Smyth et al. (2009) found that increasing the size of input fields had no effect on early respondents, but did increase the length of responses and the number of themes mentioned among late respondents. There has been relatively little work on the next three types of open-ended responses.

The results for frequency or numeric responses are mixed. Couper, Traugott, and Lamias (2001) found that respondents answering frequency questions were significantly more likely to provide ranges, qualified responses, and other non-numeric input when the input field was larger (about 16 spaces) than when it was smaller (2 spaces). Fuchs (2009a) found significant effects of the length of the input field on responses in the paper version (with longer fields producing more ill-formed answers), but not for the Web version. In a second experiment, Fuchs (2009b) varied both the length of the input field and the use of labels in a Web survey, and again found no differences between the long and short input fields in whether the desired numeric format was entered.

Several studies have found that the design of text fields for formatted numeric or verbal responses affects how well respondents provide the data in the desired format. Christian, Dillman, and Smyth (2007) experimentally varied a number of design features in questions eliciting dates (month and year) in a survey among college students. They found that the size of the answer spaces, the style of labels used (words versus symbols), and the alignment of the labels and answer spaces "independently and jointly increase the percentage of respondents reporting their answer in the desired format," where the desired format was a 2-digit month and 4-digit year.

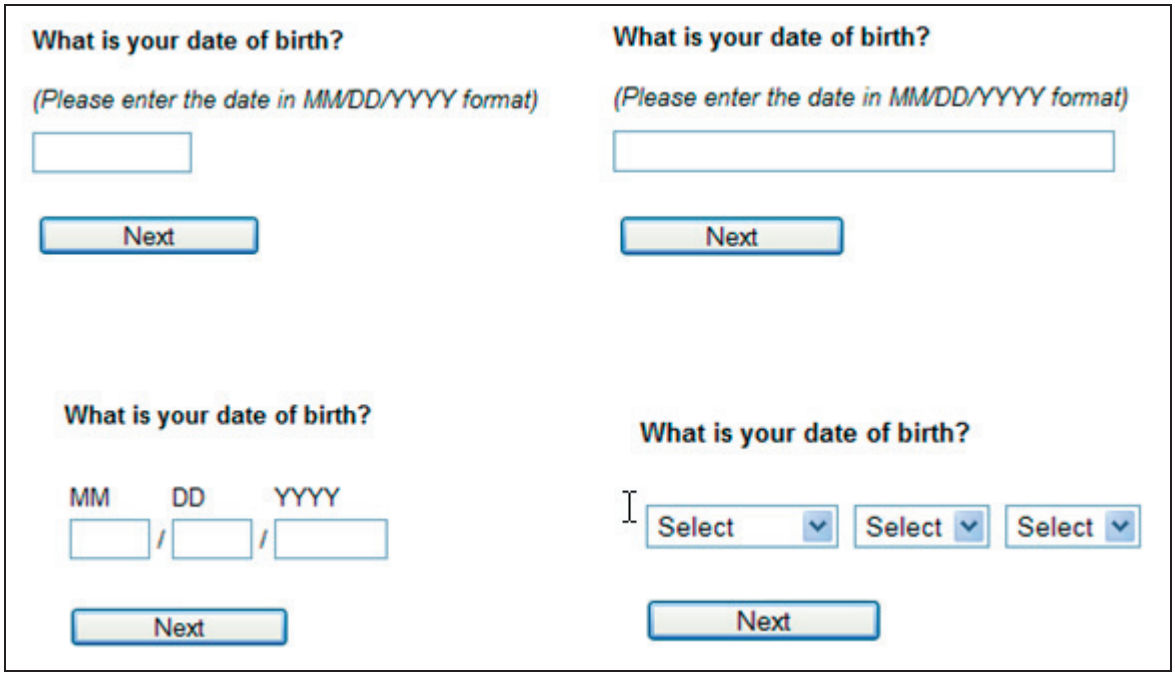
In a third Web experiment, Fuchs (2009b) experimented with placing a correctly formatted default value ("0,000.00") inside the input field for a question on the amount spent on alcoholic beverages. Those who got the default value were significantly more likely than those who didn't get the default to enter integers (34% versus 17%) and correspondingly less likely to enter alphanumeric values (6% versus 22%), suggesting that such defaults guide the respondent to the desired form of input.

Couper et al. (2010) conducted a series of experiments on questions eliciting formatted responses such as dates and currency amounts. They found that the proportion of "ill-formed" answers (i.e., those that do not match the desired format) can be significantly reduced by proving the respondent with masks or templates to guide input. An example of such a format appears in Figure 5.

Figure 5. Example of Question with Template

Couper et al. (2010) found that while 81.5% and 80.1% of respondents provided a correctly-formatted response (an integer) when not provided with a template, for prescription and non-prescription drugs respectively, 96.0% and 96.8% did so when the template was displayed. Couper and colleagues also experimented with alternative input tools and formats for a question on data of birth (as illustrated in Figure 6).

Figure 6. Alternative Versions of Date of Birth Question



They found significant differences in the proportion of answers that were well-formed (i.e., MM/DD/YYYY in the US context), with 83.4% providing such answers in the short box version (top left of Figure 6), 91.1% in the long box version (top right), 96.4% in the version with 3 separate text boxes (bottom left), and 97.6% in the drop box version (bottom right). These results again point to the importance of both the choice of input tool and, given such a choice, the appropriate design of the tool, in affecting the quality of data obtained in Web surveys.

2.6 The Use of and Design of Grid or Matrix Questions

Whether one uses scrolling or paging designs, a common design decision in many surveys is whether to group related items that share a common set of response options into a grid or matrix. The use of grid questions is common practice in Web surveys, yet the research suggests that while grids may reduce completion time, the format may increase breakoffs, missing data, and straightlining (i.e., giving the same answer to all items). The research on grids is of two main types, the first focusing on comparing questions in grids to questions on separate pages or item-by-item approaches on the same page, and the second exploring the effects of alternative designs on respondent performance on grid questions. I briefly review the evidence from both streams of research here.

A fairly consistent finding in the literature is that grids take less time to complete than item-by-item approaches (e.g., Couper, Traugott, and Lamias, 2001; Bell, Mangione, and Kahn, 2001; Tourangeau, Couper, and Conrad, 2004; Callegaro, Shand-Lubbers, and Dennis, 2009). But speed is not the only consideration, and may be an indication of suboptimal response behavior.

There is less agreement on whether grids increase inter-item correlations and whether this is good or bad. For example, Couper, Traugott, and Lamias (2001), Bell, Mangione and Kahn (2001), and Toepoel, Das, and van Soest (2009) all found no significant differences in inter-item correlations. However, Tourangeau, Couper, and Conrad (2004) compared 8 agree-disagree items presented in 3 formats: 1) all items on a single page in a grid, 2) 4 items on one page and 4 on the next, also in grids, and 3) each item on a separate page, and found a significant trend in the alpha coefficients across the three conditions, with inter-item correlations increasing as the grouping of items increased. They also found that respondents who got the items in a grid showed less differentiation (i.e., were more likely to choose the same response option for all items). Further, the part-whole correlations for two reverse-worded items were weaker in the grid version, suggesting that respondents were less likely to notice the reverse wording in a grid. A reanalysis of these data by Peytchev (2005) using structural equation modeling suggests that the increased correlations may mean higher measurement error rather than improved reliability of measurement.

The evidence on missing data rates is also somewhat mixed. Couper, Traugott, and Lamias (2001) found significantly lower rates of item missing data ("don't know" [DK] or "not applicable" [NA] responses) in the grid version. Toepoel, Das, and van Soest (2009b) compared four versions of a 40-item arousal scale, with 1, 4, 10, or 40 items per page. They found that item missing data increased monotonically with the number of items on a page.

These results are quite mixed, suggesting that there may be variations in the question topics, the populations being surveyed, the number of items and response options in the grid, whether the items asked about attitudes or behavior, and so on. This suggests there may be something in the way grids are designed and implemented, rather than something inherently negative in grids.

More recent research has focused on how the design of grids may affect the data quality outcomes of interest. Kaczmirek (2009) contrasted a grid with shading of alternative rows (called preselection feedback) versus no shading to explore whether such shading helped respondents with orientation in a grid. He found a slightly lower missing data rate (13.9% with any of the 16 items missing) for the shaded version than the non-shaded one (17% with any missing), but the difference was not significant. Completion times did also not differ significantly between versions.

Galesic et al. (2007) tested dynamic shading where each row of the grid was grayed out after an answer had been selected (called postselection feedback by Kaczmirek). They found that dynamic shading (whether by graying out labels in the row, or changing the background to gray) significantly reduced item missing data relative to the standard version, for each of the three grids tested. For example, for an 18-item grid, 8.3% of respondents missed one or more items in the standard version, compared to 1.6% for the label shading and 3.4% for the background shading versions. Breakoff rates were also slightly (but not significantly lower) in the dynamic versions.

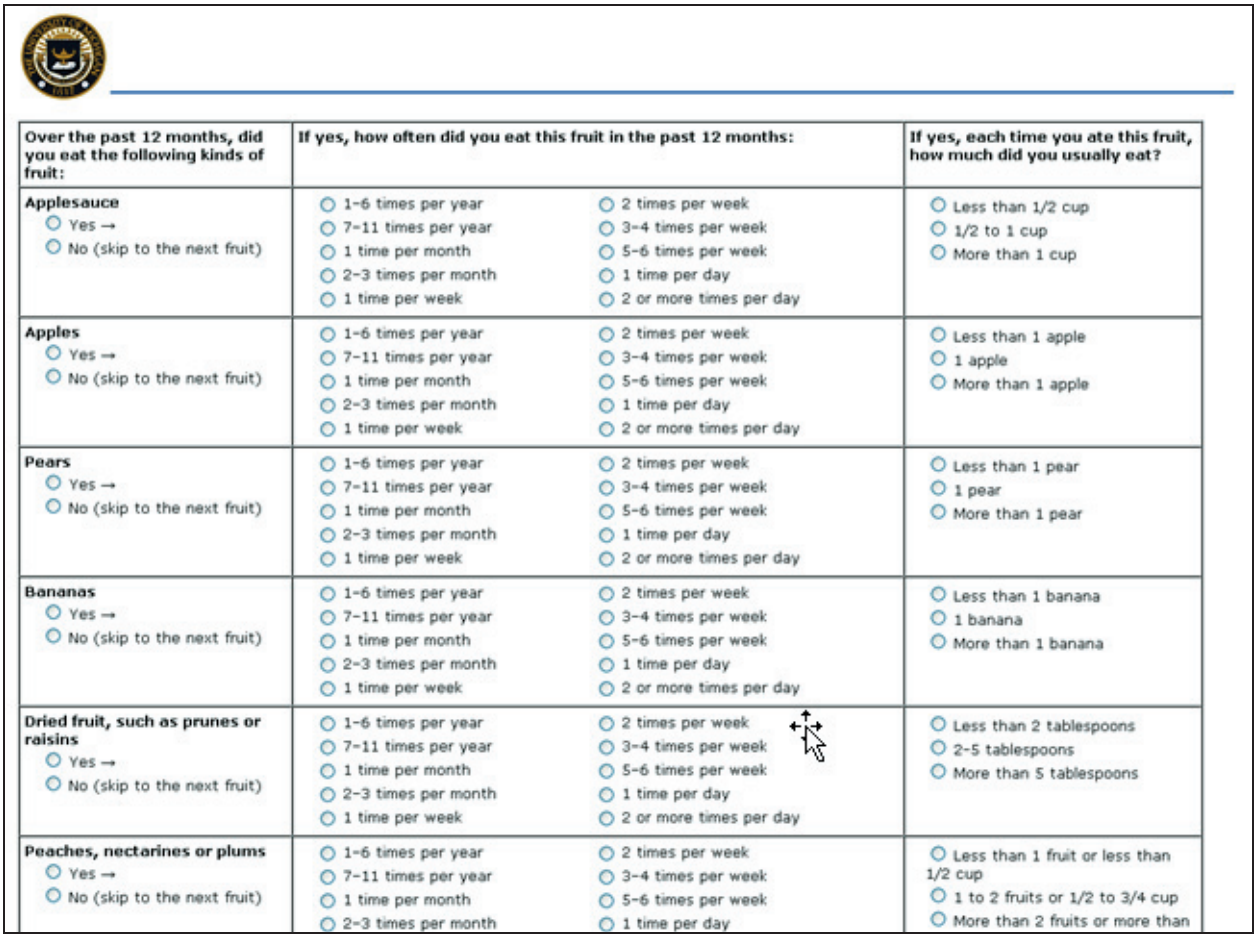
Kaczmirek (2009) also tested alternative postselection feedback design in grids, and found that graying out the row (in similar fashion to the Galesic et al. design) significantly reduced item

missing data relative to the control. However, more dynamic preselection feedback (i.e., a mouseover effect which highlights the row and column before selection) increased item missing data relative to the control, suggesting that this feature distracted respondents from completing the task. In a follow-up experiment (Kaczmirek, 2011), he found that highlighting the row both before and after selection helps with task completion, but that highlighting both rows and columns (cells in the table or grid) does not.

Couper and colleagues (Couper, Tourangeau, and Conrad, 2009; Couper et al., 2011) have examined design alternatives for a complex grid involving 13 items each of which has two or three possible questions. They found that dynamic postselection shading significantly reduced both errors of omission (item missing data) and errors of commission (answering inapplicable questions) over the control version with no shading. Further, they found that splitting the grids — that is, first asking frequency of consumption for each of the 13 fruits, then asking about amount of consumption for those fruits consumed — further reduced errors of omission and eliminated (by design) errors of commission on the follow-up question. Neither manipulation (dynamic shading or split grids) resulted in significant differences in completion time relative to the control.

In a second study, Couper and his colleagues (2011) turned the 2-item question into a 3-item version by first asking an explicit yes/no question about consumption, then asking about frequency of consumption and finally about amount of consumption. An example of the single-grid version of this question is shown in Figure 7. In the single-grid version where all three questions were visible to respondents, significantly fewer fruits were endorsed than in the version where the yes/no questions were asked in a separate grid from the follow-up questions. This suggests that by exposing respondents to the amount of information desired, reported fruit consumption is reduced, a phenomenon known as motivated underreporting (Kreuter et al., 2011). While the overall completion times increased slightly by splitting the questions into three successive grids, given the higher number of fruits endorsed in this version, the time per fruit endorsed was actually lower. Thus, the research suggests that splitting grids into their component questions may reduce overall missing data, reduce the number of "no" responses to avoid answering the follow-up questions, and reduce completion time per response.

Figure 7. Extract of Grid from Couper et al. (2011)



Over the past 12 months, did you eat the following kinds of fruit:	If yes, how often did you eat this fruit in the past 12 months:		If yes, each time you ate this fruit, how much did you usually eat?
Applesauce ○ Yes → ○ No (skip to the next fruit)	○ 1-6 times per year ○ 7-11 times per year ○ 1 time per month ○ 2-3 times per month ○ 1 time per week	○ 2 times per week ○ 3-4 times per week ○ 5-6 times per week ○ 1 time per day ○ 2 or more times per day	○ Less than 1/2 cup ○ 1/2 to 1 cup ○ More than 1 cup
Apples ○ Yes → ○ No (skip to the next fruit)	○ 1-6 times per year ○ 7-11 times per year ○ 1 time per month ○ 2-3 times per month ○ 1 time per week	○ 2 times per week ○ 3-4 times per week ○ 5-6 times per week ○ 1 time per day ○ 2 or more times per day	○ Less than 1 apple ○ 1 apple ○ More than 1 apple
Pears ○ Yes → ○ No (skip to the next fruit)	○ 1-6 times per year ○ 7-11 times per year ○ 1 time per month ○ 2-3 times per month ○ 1 time per week	○ 2 times per week ○ 3-4 times per week ○ 5-6 times per week ○ 1 time per day ○ 2 or more times per day	○ Less than 1 pear ○ 1 pear ○ More than 1 pear
Bananas ○ Yes → ○ No (skip to the next fruit)	○ 1-6 times per year ○ 7-11 times per year ○ 1 time per month ○ 2-3 times per month ○ 1 time per week	○ 2 times per week ○ 3-4 times per week ○ 5-6 times per week ○ 1 time per day ○ 2 or more times per day	○ Less than 1 banana ○ 1 banana ○ More than 1 banana
Dried fruit, such as prunes or raisins ○ Yes → ○ No (skip to the next fruit)	○ 1-6 times per year ○ 7-11 times per year ○ 1 time per month ○ 2-3 times per month ○ 1 time per week	○ 2 times per week ○ 3-4 times per week ○ 5-6 times per week ○ 1 time per day ○ 2 or more times per day	○ Less than 2 tablespoons ○ 2-5 tablespoons ○ More than 5 tablespoons
Peaches, nectarines or plums ○ Yes → ○ No (skip to the next fruit)	○ 1-6 times per year ○ 7-11 times per year ○ 1 time per month ○ 2-3 times per month	○ 2 times per week ○ 3-4 times per week ○ 5-6 times per week ○ 1 time per day	○ Less than 1 fruit or less than 1/2 cup ○ 1 to 2 fruits or 1/2 to 3/4 cup ○ More than 2 fruits or more than

Together, these studies suggest that the complexity of grids may be contributing to their negative effect on survey breakoffs, item missing data and respondent satisfaction. In other words, if grids are to be used, reducing their complexity — whether by reducing the number of items (rows) per grid, by splitting the questions (columns), or by using feedback to guide respondents in the completion task — may reduce some of the negative effects associated with grids.

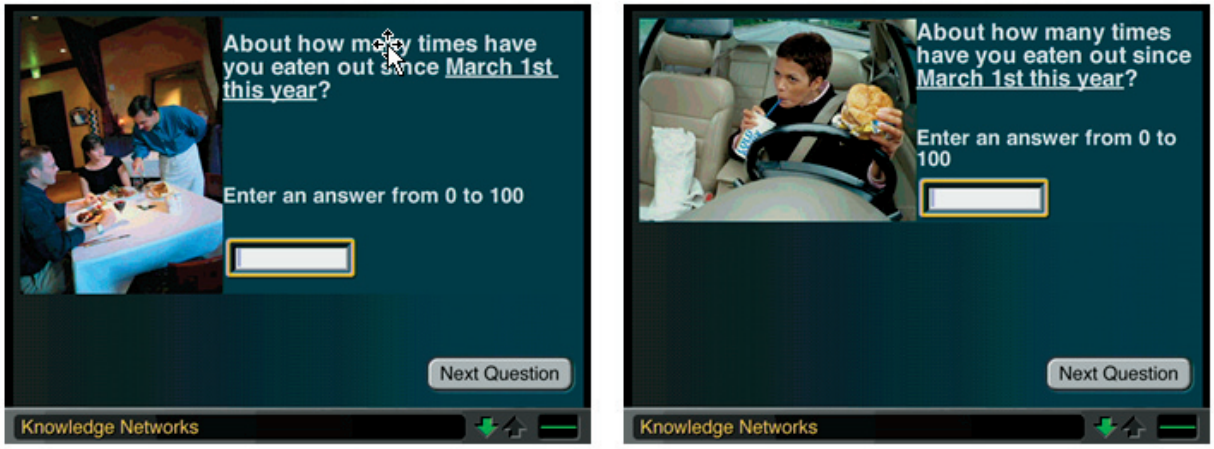
2.7 Images in Web Surveys

I noted earlier than one of the unique features of Web surveys is the ease with which graphical elements (lines, colors, drawings, images, etc.) can be added to the Web page. But the survey designer's question should be if such enhancements are useful in improving data quality or promoting respondent engagement, and if so, when and how should they be used. While the inclusion of images is likely to be beneficial in many different instances, several studies point to the inherent risks of including such images without careful consideration. Again, I briefly review a few of these to illustrate the issue, but point the reader to the individual papers or Couper (2008a) for a fuller discussion of the role of images in Web surveys.

In an early experiment, Couper, Kenyon, and Tourangeau (2004) looked at the effect of images on behavioral frequency reports, using images as exemplars. The pictures accompanying the survey question represented either low-frequency or high-frequency instances of the behavior in question (shopping, travel, eating out, etc.). Examples of a low and high frequency instance of a behavior (eating out) are shown in Figure 8. In this example, when shown a picture of a couple dining in a fine restaurant, respondents reported significantly fewer events (eating out in the past month) on average than when they were shown a high-frequency instance of the behavior (a person eating fast food in a car). The average number of episodes per month was 11.8 in the first instance and 16.2 in the second. Furthermore, those exposed to the fine restaurant image reported significantly higher levels of enjoyment, and a significantly higher price paid for the last meal eaten out. Similarly, respondents exposed to a picture of grocery shopping reported more shopping trips in the past month than those exposed to a picture of clothes shopping. Similar effects were found for several other behaviors. The images served as cues for the retrieval of relevant incidents from memory, hence affecting the frequency reports. In a partial replication in Germany, Hemsing and Hellwig (2006) found similar assimilation effects.

Toepoel and Couper (2011) replicated and extended this work in a probability-based online panel in the Netherlands. Using the same pictures, they found similar effects as did Couper, Kenyon, and Tourangeau (2004) among adult Dutch respondents. However, they also added a condition with verbal instructions, which in some cases contradicted the images (e.g., by instructing respondents to include only dining in a restaurant, not fast food). They found that the verbal instructions had a larger effect on the behavioral frequency reports than did the pictures, but both words and images affected the answers obtained. Their study provides further evidence of a hierarchy of cues suggested by Tourangeau, Couper and Conrad (2007), with verbal cues being stronger than visual ones (see also Toepoel and Dillman, 2011).

Figure 8. Low and High Frequency Examples of Eating Out Behavior



Similar assimilation effects were found by Tourangeau and colleagues (2009) in two studies looking at examples of types of food consumed. In these studies, respondents got photographs depicting examples of different categories of food; respondents who got pictures of foods that are generally eaten frequently (such as cheese and butter) reported eating more servings of foods in

the target category (dairy products) in a typical week than those who got pictures of foods eaten less often (frozen yogurt and sour cream). Tourangeau and his colleagues argue that respondents base their food frequency estimates on the small set of category members that come to mind and that the pictures change which examples respondents consider as they formulate their estimates. When the pictures showed frequently-consumed examples, the frequency estimates went up; when they show infrequently-consumed examples, the frequency estimates went down.

Another type of context effect that images can have is by serving as a standard of comparison affecting the judgments that respondents make. A set of experiments by Couper, Conrad, and Tourangeau (2007) displayed photographs either of a woman in a hospital bed or a woman jogging (or no picture, in some cases) to Web survey respondents while answering a standard self-rated health question (see Figure 9). In addition, they varied the position of the picture (putting it on the screen preceding the target question, on the same page just to the left and above the target question, or in the header at the top of the same screen with the target question) and size of the picture. Consistent with the context effects literature, they found a significant contrast effect, with those seeing the sick woman reporting significant higher levels of health than those seeing the fit woman. These effects were substantively meaningful: for example 43% of respondents reported themselves to be in very good health (8-10 on a 10-point scale) when exposed to the picture of the fit woman, but 58% did so when shown the picture of sick woman.

Figure 9. Images from Self-Reported Health Experiments

In their first experiment, Couper, Conrad, and Tourangeau (2007) found that when the pictures were put in the banner of the Web page, the effect was smaller than when the pictures appeared either alongside the question or the prior introductory screen. This produced some initial support for the "banner blindness" hypothesis that material in banners is likely to be ignored (Benway, 1998; Benway and Lane, 1998). But in subsequent replications of the experiment, the contrast effect was found, although not as strong as the effect when the picture appeared along side the question.

A follow-up study by Galesic and colleagues (2008) used eye tracking to examine how the placement of the picture affected the time respondents spent looking at it. When the picture appeared in the header respondents often ignored it, as suggested by the banner blindness hypothesis. Overall, respondents fixated less often on the picture and fixated for less overall time on average when the picture was presented in the header than when it was presented next to the question. Galesic et al. (2008) found some evidence that when respondents did fixate on the picture, the expected contrast effects were found; however, among respondents who never fixated on the picture in the header, there was some evidence of an assimilation effect.

These and other studies suggest that images can and do affect the responses obtained in Web surveys. However, there has been little or no research published in the marketing and advertising arenas to demonstrate the value of images (e.g., for brand recognition). Some have argued that images are useful in keeping respondents engaged in the survey, but so far we have found no evidence that the inclusion of images reduced breakoffs or increased subjective satisfaction. For now, the evidence suggests caution when including images in Web surveys. The decision to use images — and which images to use — should be taken with care as part of a deliberate design process.

2.8 Running Tallies

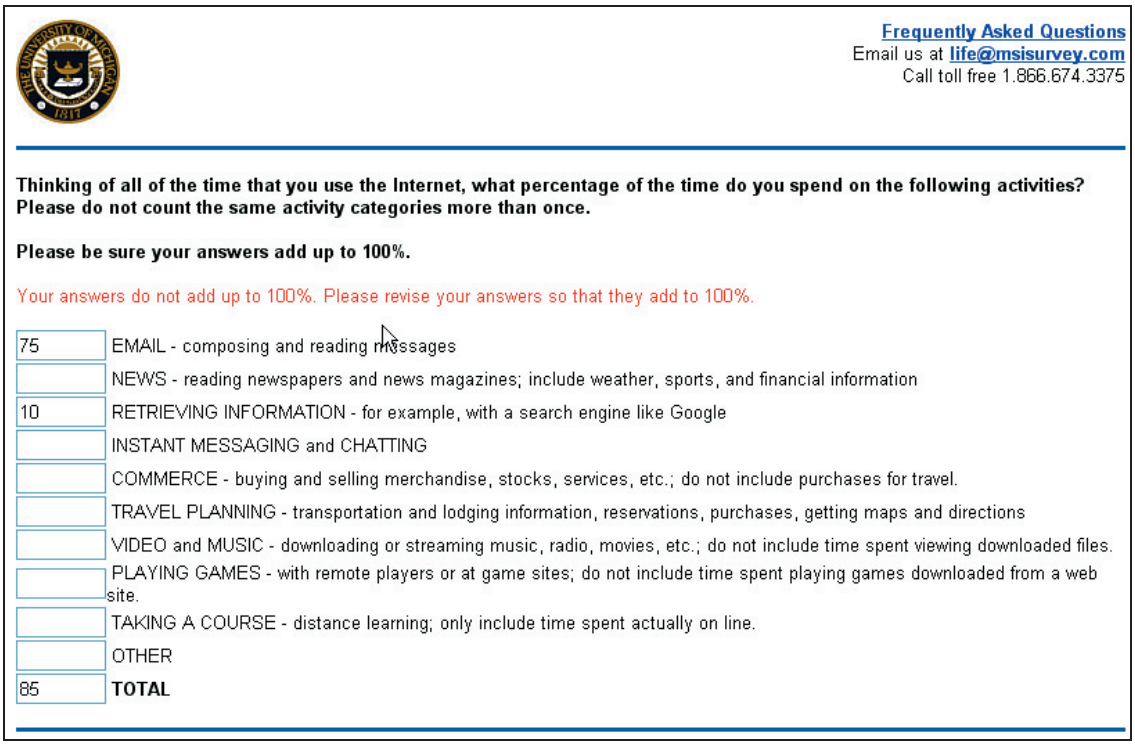
The previous sections have focused on the visual aspects of Web surveys, noting how important it is to pay attention to these design elements in Web surveys. In the next two sections, I turn to design features related to the interactive aspects of Web surveys. There are many interactive elements that can be added to Web surveys, including visual analog scales or graphic rating scales (i.e., slider bars), drag-and-drop card-sort tasks, map-based activities, and a variety of other ways of asking survey questions. Some of these examples come from the market research world, but academic researchers are also developing new ways to exploit the interactive capabilities of Web surveys to develop new ways of eliciting information from respondents. In general, these interactive elements require some form of active scripting, such as JavaScript or Flash. In this section, I focus on one example, the provision of a type of dynamic feedback known as running totals or tallies.

There are two main types of running tallies. The first type, referred to as fixed summation validations by Peytchev and Crawford (2005) and as constant sums by Conrad et al. (2005), are where the set of items should add up to some known total amount (e.g., 100%, 24 hours, etc.). An example of this type is shown in Figure 10. The total for the second type of tally is based on

information collected earlier in the survey (e.g., "Of the 12 trips you reported taking last year, how many were trips abroad?"). Peytchev and Crawford (2005) refer to this type as respondent behavior validations in their discussion of real-time validations in Web surveys.

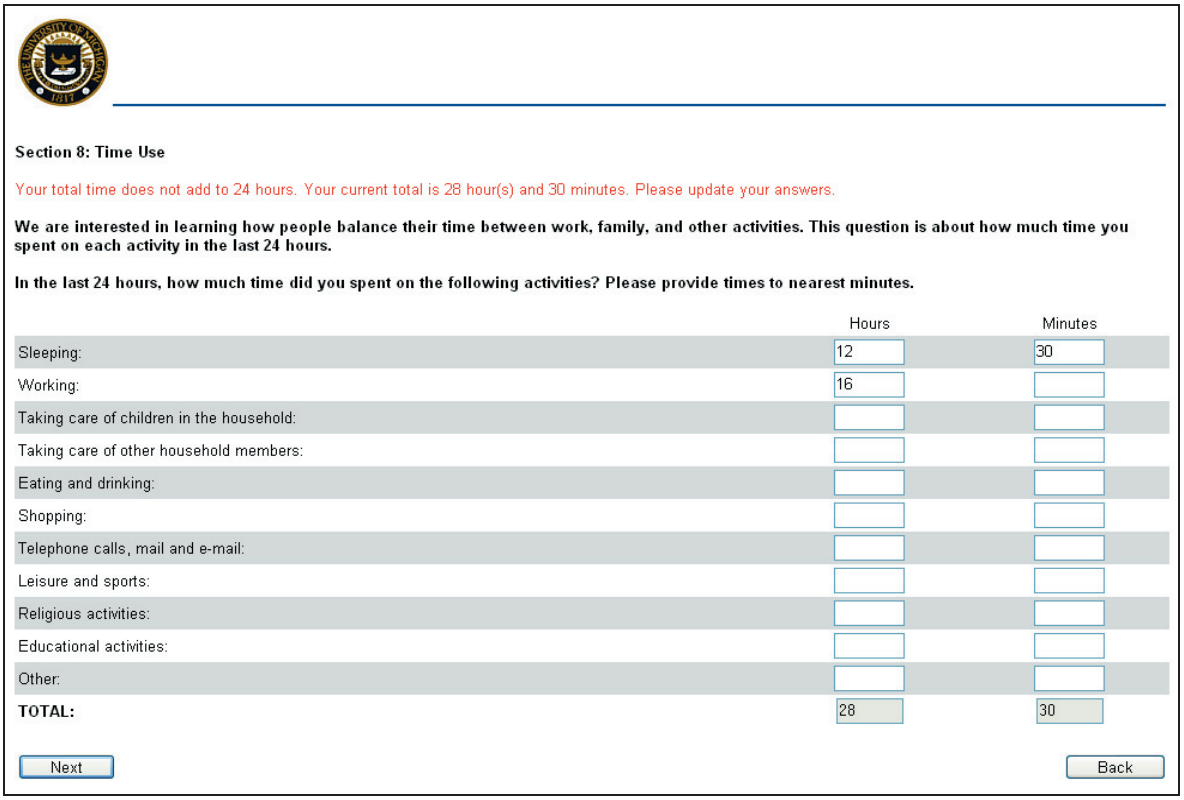
Conrad and colleagues (2005) compared a running tally with no feedback to one with only server-side (delayed) feedback, and one with both client-side (concurrent) and server side feedback. The example in Figure 10 has both client-side feedback (the 85 in the total box) and server-side feedback (the error message generated if the respondent pressed "Continue" with the total not equal to 100). With accuracy or "well-formedness" measured as the number of tallies that equaled the constant sum (100% in this case), accuracy was 84.8% for the version without any feedback, compared to 93.1% for the version with delayed feedback only, and 96.5% for the version with both concurrent and delayed feedback. There was also an advantage of concurrent over delayed feedback in terms of the proportion of respondents entering an accurate response at the outset — that is, on the first try. Response times were also faster for the version with the running tally.

Figure 10. Example of a Running Tally



In a subsequent study, Conrad and colleagues (2009) added a more complex question that would likely require greater use of the running tally. The item was a 24-hour time diary, with columns for entries in both hours and minutes (see Figure 11). The expectation was that the more complex the task (i.e., the more arithmetic is required to provide a well-formed answer), the more helpful a running tally would be.

Figure 11. Example of Complex Running Tally



Section 8: Time Use

Your total time does not add to 24 hours. Your current total is 28 hour(s) and 30 minutes. Please update your answers.

We are interested in learning how people balance their time between work, family, and other activities. This question is about how much time you spent on each activity in the last 24 hours.

In the last 24 hours, how much time did you spent on the following activities? Please provide times to nearest minutes.

	Hours	Minutes
Sleeping:	12	30
Working:	16	
Taking care of children in the household:		
Taking care of other household members:		
Eating and drinking:		
Shopping:		
Telephone calls, mail and e-mail:		
Leisure and sports:		
Religious activities:		
Educational activities:		
Other:		
TOTAL:	28	30

Next Back

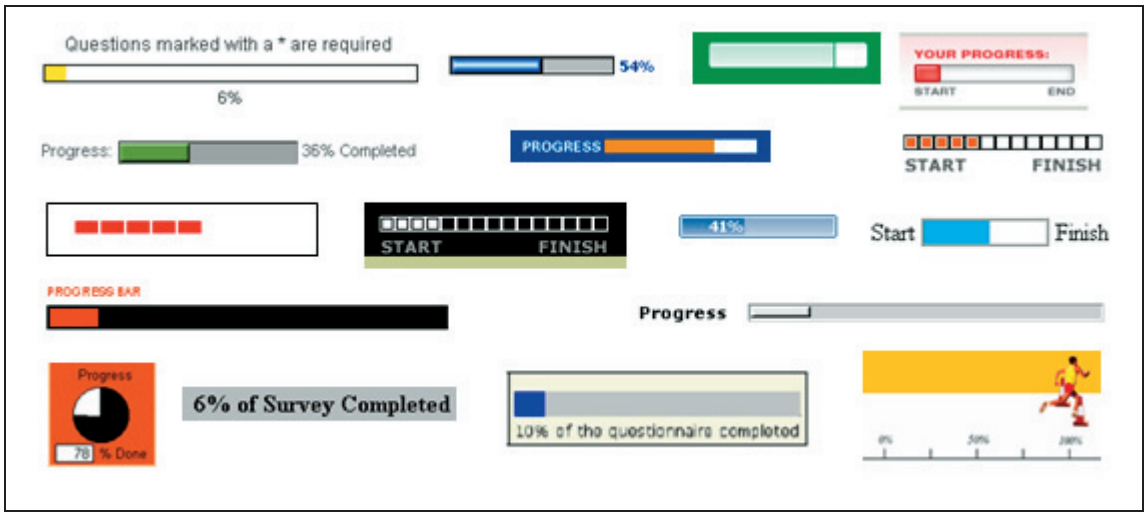
In this case, only 36.7% of respondents in the condition with no feedback provided well-formed answers (i.e., activities that added up to 24 hours and 0 minutes). This increased to 71.4% of respondents in the delayed feedback version, and 89.4% in the concurrent feedback (running tally) version. Conrad and colleagues compared the distribution of responses to data from the American Time Use Survey (ATUS) and found that discrepancies between the Web survey data and ATUS data were smaller when feedback was provided. This suggests that feedback helps both data completeness and validity. They also found similar benefits from a running tally when the question was presented at the end of the questionnaire and asked about time spent on each section of the questionnaire; the estimates for each section were intended to add to the total time spent on the questionnaire. The running tally (there was no delayed feedback) led to more well-formed and more accurate responses. So, it appears that the use of interactive feedback improves the quantity of usable data and the validity of the individual reports.

2.9 Progress Indicators

Another example of an interactive element in Web surveys that has received a lot of research attention is progress indicators. Progress indicators are popular tools in paging Web surveys, to convey to respondents how much there is left in the survey (or how much they have already completed). The stated goal of progress indicators is to prevent breakoffs. Progress indicators are not needed in scrolling surveys as the respondent can simply scroll down to the bottom of the

survey to see how many questions remain. Some examples of progress indicators are shown in Figure 12.

Figure 12. Examples of Progress Indicators



There have been a large number of studies on progress indicators (PIs) in Web surveys. I review some selected findings here, and provide an overall summary of what is known about the value of PIs.

In one of the first studies on PIs, Couper, Traugott, and Lamias (2001) used a graphical PI that showed progress as a pie chart, along with the percent completed in text. They found a slightly (but not significantly) lower breakoff rate (10.1%) for the PI version than the no-PI version (13.6%), and speculated that downloading the graphical image may have attenuated the positive effect of the PI. This was supported by the fact that the PI version took significantly longer to complete (22.7 versus 19.8 minutes).

In a subsequent study, Crawford, Couper, and Lamias (2001), used a text-based PI (e.g., "17% of Survey Completed") to eliminate any differences in download time between versions. They found a significant difference in breakoff rates, but *higher* for the PI group (31.5%) than the no-PI group (25.3%). The explanation for this negative effect seemed to lie in a series of time-consuming open-ended questions at the beginning of the survey. On average, when respondents were about halfway through the survey (in terms of elapsed time), the PI based on the number of questions was suggesting that they were only about one fifth of the way through the survey. A follow-up study among the same respondents six months later removed the set of open-ended questions, and found a lower but again insignificant (p=.07) breakoff rate for those getting the PI (8.2% versus 13.8% for no PI).

Conrad and colleagues (2010; see also Yan et al., 2011) have conducted a series of studies to explore the PI issue further, varying the "speed" of the PI, contrasting the standard "constant speed" approach with a "slow-to-fast" version in which the speed of progress increased over the course of the survey, and a "fast-to-slow" version in which speed decreased. As found in the earlier studies, the constant speed PI performed a little worse (p>.05) than no PI (breakoffs of

14.4% and 12.7% respectively). The version with the good news (faster progress) early on performed significantly better (11.3% breakoffs), while the version with bad news (slow progress) at the beginning did significantly worse (21.8% breakoffs), as expected from the Crawford, Couper, and Lamias (2001) result.

In a second study (Conrad et al., 2010), they crossed the same three PI speeds with whether the PI was always on, displayed intermittently (about every eighth page), or displayed only upon request (by clicking on a hyperlink on each page). The results for breakoff rates by PI speed matched those from the first study, with the fast-to-slow being best (11.3% breakoffs), the slow-to-fast worst (19.9%) with the constant speed PI (14.4%) and no PI (14.3%) in between. The type of PI (always on, intermittent, or on demand) had no apparent effect on breakoffs. Interestingly, in the on-demand group, only 37% of respondents ever requested the PI, and it was requested approximately 1% of the time (or 1 in every 100 pages) on average.

Other studies have similarly found mixed support for progress indicators. In a survey among students in Germany, Glauer and Schneider (2004) found a small but significant positive effect on breakoff rates for a PI (16.2%) than a no-PI (20.6%) version. Van der Horst, Snijders, and Matzat (2006) tested several different PIs in an opt-in panel. Breakoff rates were lowest for the no-PI group (8.3%) with the constant progress indicator (12.1%), a progressive PI (similar to Conrad et al.'s fast-to-slow PI, 10.3%) a degressive PI (like Conrad et al.'s slow-to-fast, 13.4%) all being worse in terms of breakoffs. They also tested a descending page indicator (e.g., "3 pages to go"), which performed worst of all, with 16% breakoffs. Kaczmirek and colleagues (2005) tested a PI that recalculated the remaining number of pages following a skip; this dynamic PI showed slightly lower breakoff rates than the control, but the difference was not significant.

Heerwegh and Loosveldt (2006) experimented with giving respondents a choice at the start of the survey of displaying the PI throughout or not (77.4% chose to do so). Those who were not given a choice were assigned to a PI, the university logo, or nothing, while those who chose not to see the PI were assigned to the logo or nothing. The PI was set to behave like the fast-to-slow version used in Conrad et al. (2010). Breakoff rates for those in the PI group (11.3%) were not significantly different than for those not shown a PI (12.6%). Those who were given a choice and chose the PI had a similar breakoff rate (9.5%) to those not given a choice but shown the PI (11.3%). The group with the highest breakoff rate was those offered a PI who did not choose to display it (20.7%).

A recent meta-analysis by Callegaro, Villar, and Yang (2011) examined 32 experimental comparisons of progress indicators from 10 different papers. For the 18 studies comparing a constant PI to a no-PI condition, they found a small tendency (a log odds ratio of 0.072, p=0.37) for *higher* breakoff rates in the PI condition than in the control group. While they find a positive effect of fast-to-slow PIs across 7 studies, they note (as does Conrad et al., 2010) that using such misleading progress indicators raises ethical concerns.

Despite the popularity of progress indicators among survey designers, the research evidence suggests that the typical progress indicators used in Web surveys — always on, constant speed, progressive, not adapted to skips — do not help much in reducing breakoffs, and may even hurt

at times. This is another example of the risks of leaving Web survey design to those who are not survey researchers familiar with the literature. What may be technically feasible — and is often a promoted feature of many Web survey software systems — may not actually be helpful in practice.

2.10 Summary on Design Issues

In this part of the seminar, I have presented a selected set of examples of design choices that may affect the answers obtained in Web surveys. There are a large number of studies that suggest that these design issues matter for Web surveys. Careful attention to design of the online instruments can ensure good quality data.

The good news from a measurement perspective is that the flexibility and interactivity of the Web can improve the quality of data obtained, especially relative to paper-based surveys. Web surveys offer the advantages of self-administration on the one hand, and the power of computerized surveys on the other. There is growing evidence that Web surveys can reduce item-missing data and edit failures relative to mail surveys, that the open-ended responses obtained are at least as long and detailed as paper surveys, and that a number of complex question types can be administered that were previously only possible in interviewer-administered surveys. So, if the inferential challenges of Web surveys can be overcome, they offer the survey designer a great deal of opportunity in terms of improvement to the measurement process. Continued research focuses on ways to minimize the drawbacks of Web surveys so that their advantages may be put to better use.

References

AAPOR (2010), *AAPOR Report on Online Panels*. Deerfield, IL: American Association for Public Opinion Research.

Alexander, G.L., Divine, G.W., Couper, M.P., McClure, J.B., Stopponi, M.A., Fortman, K.K., Tolsma, D.D., Strecher, V.J., and Johnson, C.C. (2008), "Effect of Incentives and Mailing Features on Recruitment for an Online Health Program." *American Journal of Preventive Medicine*, 34 (5): 382-388.

Baker-Prewitt, J. (2010), "Looking beyond Quality Differences: How Do Consumer Buying Patterns Differ by Sample Source?" Paper presented at the CASRO Panel Conference, New Orleans, February.

Bell, D.S., Mangione, C.M., and Kahn, C.E. (2001), "Randomized Testing of Alternative Survey Formats Using Anonymous Volunteers on the World Wide Web." *Journal of the American Medical Informatics Association*, 8 (6): 616-620.

Benway, J.P. (1998), "Banner Blindness: The Irony of Attention Grabbing on the World Wide Web." *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*. Santa Monica: HFES, pp. 463-467.

Benway, J.P., and Lane, D.M. (1998), "Banner Blindness: Web Searchers Often Miss 'Obvious' Links." *Internetworking Newsletter*, ITG 1.3, December. Available at, http://www.internettg.org/newsletter/dec98/banner_blindness.html

Berrens, R.P., Bohara, A.K., Jenkins-Smith, H., Silva, C., and Weimer, D.L. (2003), "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Surveys." *Political Analysis,* 11 (1): 1-22.

Bethlehem, J. (2010), "Selection Bias in Web Surveys." *International Statistical Review*, 78 (2): 161-188.

Bhutta, C.B. (2010), "Not by the Book: Facebook as a Sampling Frame." Available at SSRN: http://ssrn.com/abstract=1721162

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (eds.) (1991), *Measurement Errors in Surveys*. New York: Wiley.

Birnholtz, J.P., Horn, D.B., Finholt, T.A., and Bae, S.J. (2004), "The Effects of Cash, Electronic, and Paper Gift Certificates as Respondent Incentives for a Web-based Survey of Technologically Sophisticated Respondents." *Social Science Computer Review*, 22 (3): 355-362.

Börsch-Supan, A., Elsner, D., Faßbender, H., Kiefer, R., McFadden, D., and Winter, J. (2007), "How to Make Internet Surveys Representative: A Case Study of a Two-Step Weighting Procedure." Paper presented at the MESS Workshop, 's-Hertogenbosch, Netherlands, August/September.

Bosnjak, M., Neubarth, W., Couper, M.P., Bandilla, W., and Kaczmirek, L. (2008), "Prenotification in Web-Based Access Panel Surveys: The Influence of Mobile Text Messaging versus E-Mail on Response Rates and Sample Composition." *Social Science Computer Review*, 26 (2): 213-223.

Bosnjak, M., and Tuten, T.L. (2002), "Prepaid and Promised Incentives in Web Surveys – An Experiment." *Social Science Computer Review*, 21 (2): 208-217.

Brennan, M. (2005), "The Effect of a Simultaneous Mixed-Mode (Mail and Web) Survey on Respondent Characteristics and Survey Responses." Paper presented at the ANZMAC 2005 Conference.

Brøgger, J., Nystad, W., Cappelen, I., and Bakke, P. (2007), "No Increase in Response Rate by Adding a Web Response Option to a Postal Population Survey: A Randomized Trial." *Journal of Medical Internet Research*, 9 (5): e40.

Burris, J., Chen, J., Graf, I., Johnson, T., and Owens, L. (2001), "An Experiment in Web Survey Design." Paper presented at the annual meeting of the American Association for Public Opinion Research, Montreal, Quebec, May.

Callegaro, M., and DiSogra, C. (2008), "Computing Response Metrics for Online Panels." *Public Opinion Quarterly*, 72 (5): 1008-1032.

Callegaro, M., Shand-Lubbers, J., and Dennis, J.M. (2009), "Presentation of a Single Item Versus a Grid: Effects on the Vitality and Mental Health Subscales of the SF-36v2 Health survey." Paper presented at the annual meeting of the American Association for Public Opinion Research, Hollywood, FL, May.

Callegaro, M., Villar, A., and Yang, Y. (2011), "A Meta-Analysis of Experiments Manipulating Progress Indicators in Web Surveys." Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, May.

Cantor, D., Brick, P.D., Han, D., and Aponte, M. (2010), "Incorporating a Web Option in a Two-Phase Mail Survey." Paper presented at the annual meeting of the American Association for Public Opinion Research, Chicago, May.

Cardador, J. (2010), "Using Overt and Covert Survey Traps to Maximize Data Quality." Paper presented at the annual meeting of the American Association for Public Opinion Research, Chicago, May.

Christian, L.M., Dillman, D.A., and Smyth, J.D. (2007), "Helping Respondents Get it Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys." *Public Opinion Quarterly*, 71 (1): 113-125.

Clark, R.L., and Nyiri, Z. (2001), "Web Survey Design: Comparing a Multi-Screen to a Single Screen Survey." Paper presented at the annual meeting of the American Association for Public Opinion Research, Montreal, Quebec, May.

Conrad, F.G., Couper, M.P., Tourangeau, R., and Galesic, M. (2005), "Interactive Feedback Can Improve Quality of Responses in Web Surveys." Paper presented at the annual meeting of the American Association for Public Opinion Research, Miami Beach, May.

Conrad, F.G., Couper, M.P., Tourangeau, R., Galesic, M., and Yan, T. (2009), "Interactive Feedback Can Improve the Quality of Responses in Web Surveys." Paper presented at the conference of the European Survey Research Association, Warsaw, Poland, July.

Conrad, F.G., Couper, M.P., Tourangeau, R., and Peytchev, A. (2010), "The Impact of Progress Indicators on Task Completion." *Interacting with Computers*, 22 (5): 417-427.

Couper, M.P. (2000), "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly*, 64 (4), 464-494.

Couper, M.P. (2008a), *Designing Effective Web Surveys*. New York: Cambridge University Press.

Couper, M.P. (2008b), "Technology and the Survey Interview/Questionnaire." In M.F. Schober and F.G. Conrad (eds.), *Envisioning the Survey Interview of the Future*. New York: Wiley, pp. 58-76.

Couper, M.P., Kapteyn, A., Schonlau, M., and Winter, J. (2007), "Noncoverage and Nonresponse in an Internet Survey." *Social Science Research*, 36 (1): 131-148.

Couper, M.P., Kennedy, C., Conrad, F.G., and Tourangeau, R. (2010), "Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys." *Journal of Official Statistics*, 27 (1): 1-22.

Couper, M.P., Singer, E., Tourangeau, R., and Conrad, F.G. (2006), "Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment." *Social Science Computer Review*, 24 (2): 227-245.

Couper, M.P., Tourangeau, R., and Conrad, F.G. (2009), "Improving the Design of Complex Grid Questions." Paper presented at the Internet Survey Methods Workshop, Bergamo, Italy, September.

Couper, M.P., Tourangeau, R., Conrad, F.G., and Zhang, C. (2011), "Further Research on the Design of Complex Grid Questions." Paper presented at the Internet Survey Methods Workshop, The Hague, Netherlands, September.

Couper, M.P., Tourangeau, R., and Kenyon, K. (2004), "Picture This! An Analysis of Visual Effects in Web Surveys." *Public Opinion Quarterly*, 68 (2): 255-266.

Couper, M.P., Traugott, M., and Lamias, M. (2001), "Web Survey Design and Administration." *Public Opinion Quarterly*, 65 (2): 230-253.

Crawford, S.D., Couper, M.P., and Lamias, M. (2001), "Web Surveys: Perceptions of Burden." *Social Science Computer Review*, 19 (2): 146-162.

Crawford, S.D., McCabe, S.E., Saltz, B., Boyd, C.J., Freisthler, B., and Paschall, M.J. (2004), "Gaining Respondent Cooperation in College Web-Based Alcohol Surveys: Findings from Experiments at Two Universities." Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, AZ, May.

Deming, W.E. (1944), "On Errors in Surveys." *American Sociological Review*, 9 (4): 359-369.

Dennis, M., deRouvray, C., and Couper, M.P. (2000), "Questionnaire Design for Probability-Based Web Surveys." Paper presented at the annual meeting of the American Association for Public Opinion Research, Portland, OR, May.

Dever, J.A., Rafferty, A., and Valliant, R. (2008), "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias." *Survey Research Methods,* 2 (2): 47-62.

Downes-Le Guin, T. (2005), "Satisficing Behavior in Online Panelists." Paper presented at the MRA Annual Conference & Symposium, June.

Duffy, B., Smith, K., Terhanian, G., and Bremer, J. (2005), "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research*, 47 (6): 615-639.

ESOMAR (2008), 26 Questions to Help Buyers of Online Samples. Amsterdam: ESOMAR, see http://www.esomar.org/index.php/26-questions.html

Eurostat (2011), *Information Society Statistics*, see http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/introduction

Faas, T., and Schoen, H. (2006), "Putting a Questionnaire on the Web is not Enough – A Comparison of Online and Offline Surveys Conducted in the Context of the German Federal Election 2002." *Journal of Official Statistics*, 22 (2): 177-190.

Fuchs, M. (2009a), "Asking for Numbers and Quantities: Visual Design Effects In Paper & Pencil Surveys." *International Journal of Public Opinion Research*, 21 (1): 65-84.

Fuchs, M. (2009b), "Differences in the Visual Design Language of Paper-and-Pencil Surveys Versus Web Surveys: A Field Experimental Study on the Length of Response Fields in Open-Ended Frequency Questions." *Social Science Computer Review*, 27 (2): 213-227.

Galesic, M., Tourangeau, R., Couper, M.P., and Conrad, F.G. (2007), "Using Change to Improve Navigation in Grid Questions." Paper presented at the General Online Research Conference (GOR'07), Leipzig, March.

Galesic, M., Tourangeau, R., Couper, M.P., and Conrad, F.G. (2008), "Eye-Tracking Data: New Insights on Response Order Effects and Other Cognitive Shortcuts in Survey Responding." *Public Opinion Quarterly*, 72 (5): 866-891.

Gentry, R., and Good, C. (2008), "Offering Respondents a Choice of Survey Mode: Use Patterns of an Internet Response Option in a Mail Survey." Paper presented at the annual meeting of the American Association for Public Opinion Research, New Orleans, May.

Giacobbe, J., and Pettit, A. (2010), "Looking Beyond Quality Differences: How do Consumer Buying Patterns Differ by Sample Source?" Paper presented at the CASRO Panel Conference, New Orleans, February.

Glauer, R. and Schneider, D. (2004), "Online-Surveys: Effects of Different Display Formats, Response Orders as Well as Progress Indicators in a Non-Experimental Environment." Paper presented at the 6th German Online Research Conference, Duisburg-Essen, Germany, March.

Göritz, A.S. (2006a), "Incentives in Web Studies: Methodological Issues and a Review." *International Journal of Internet Science*, 1 (1): 58-70.

Göritz, A.S. (2006b), "Cash Lotteries as Incentives in Online Panels." *Social Science Computer Review*, 24 (4): 445-459.

Göritz, A.S. (2010), "Using Lotteries, Loyalty Points, and Other Incentives to Increase Participant Response and Completion." In S.D. Gosling and J.A. Johnson (eds.), *Advanced Methods for Behavioral Research on the Internet*. Washington, DC: American Psychological Association, pp. 219-233.

Griffin, D.H., Fischer, D.P., and Morgan, M.T. (2001), "Testing an Internet Response Option for the American Community Survey." Paper presented at the annual meeting of the American Association for Public Opinion Research, Montreal, Quebec, May.

Groves, R.M. (1989), *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. (2006), "Nonresponse Rates and Nonresponse Error in Household Surveys." *Public Opinion Quarterly*, 70 (5): 646-675.

Groves, R.M., Fowler, F.J. Jr., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2009), *Survey Methodology* (2nd Edition). New York: Wiley.

Harmon, M.A., Westin, E.C., and Levin, K.Y. (2005), "Does Type of Pre-Notification Affect Web Survey Response Rates?" Paper presented at the annual conference of the American Association for Public Opinion Research, Miami Beach, May.

Healey, B. (2007), "Drop Downs and Scroll Mice: The Effect of Response Option Format and Input Mechanism Employed on Data Quality in Web Surveys." *Social Science Computer Review*, 25 (1): 111-128.

Heckathorn, D. (1997). "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems*, 44(2): 174-199.

Heckathorn, D. (2002). "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations." *Social Problems*, 49(1): 11-34.

Heerwegh, D., and Loosveldt, G. (2006), "An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey Sponsor Logos in Web Surveys." *Journal of Official Statistics*, 22 (2): 191-210.

Hennessy, D.G. (2002), "An Evaluation of Methods for Testing Internet Survey Questions." Wollongong, Australia: University of Wollongong, Department of Psychology, unpublished Honours thesis.

Hoffer, T.B., Grigorian, K., Sederstrom, S., and Selfa, L. (2006), *2003 Survey of Doctorate Recipients Mode Experiment Analysis Report*. Chicago: National Opinion Research Center, unpublished report.

Holmberg, A., Lorenc, B., and Werner, P. (2010), "Contact Strategies to Improve Participation via the Web in a Mixed-Mode Mail and Web Survey." *Journal of Official Statistics*, 26 (3): 465-480.

Israel, G. (2009), "Obtaining Responses by Mail or Web: Response Rates and Data Consequences." *Survey Practice*, June. Http://surveypractice.org/2009/06/29/mail-vs-web/

Juran, J.M., and Gryna, F.M. (1980), *Quality Planning and Analysis*. New York: McGraw-Hill.

Kaczmirek, L. (2009), *Human-Survey Interaction: Usability and Nonresponse in Online Surveys*. Cologne: Herbert von Halem Verlag.

Kaczmirek, L. (2011), "Attention and Usability in Internet Surveys: Effects of Visual Feedback in Grid Questions." In M. Das, P. Ester, and L. Kaczmirek (eds.), *Social Research and the Internet*. New York: Taylor and Francis, pp. 191-214.

Kaczmirek, L., Neubarth, W., Bosnjak, M., and Bandilla, W. (2005), "Progress Indicators in Filter Based Surveys: Individual and Dynamic Calculation Methods." Paper presented at the General Online Research Conference, Zurich, March.

Kalton, G., and Flores-Cervantes, I. (2003), "Weighting Methods." *Journal of Official Statistics*, 19 (2): 81-97.

Kaplowitz, M.D., Hadlock, T.D., and Levine, R. (2004), "A Comparison of Web and Mail Survey Response Rates." *Public Opinion Quarterly*, 68 (1): 94-101.

Kaplowitz, M.D., Lupi, F., Couper, M.P., and Thorp, L. (in press), "The Effect of Invitation Design on Web Survey Response Rates." *Social Science Computer Review*, in press.

Kish, L. (1965), *Survey Sampling*. New York: John Wiley.

Kreuter, F., McCulloch, S., Presser, S., and Tourangeau, R. (2011), "The Effects of Asking Filter Questions in Interleafed Versus Grouped Format." *Sociological Methods and Research*, 40 (1): 88-104.

Krosnick, J.A., Ackermann, A., Malka, A., Yeager, D., Sakshaug, J., Tourangeau, R., DeBell, M., and Turakhia, C. (2009), "Creating the Face-to-Face Recruited Internet Survey Platform (FFRISP)." Paper presented at the Third Annual Workshop on Measurement and Experimentation with Internet Panels, Santpoort, The Netherlands, August.

Lebrasseur, D., Morin, J.-P., Rodrigue, J.-F., and Taylor, J. (2010), "Evaluation of the Innovations Implemented in the 2009 Canadian Census Test." *Proceedings of the American Statistical Association Survey Research Methods Section*, pp. 4089-4097.

Lee, S. (2006), "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics*, 22 (2): 329-349.

Lee, S. (2009). "Understanding Respondent Driven Sampling from a Total Survey Error Perspective." Survey Practice, September: http://surveypractice.org/

Lee, S., and Valliant, R. (2009), "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods & Research*, 37 (3): 319-343.

Lensvelt-Mulders, G.J.L.M., Lugtig, P., and Hubregtse, M. (2009), "Separating Selection Bias and Non-Coverage in Internet Panels using Propensity Matching." *Survey Practice*, August: http://surveypractice.org/

Lessler, J.T., and Kalsbeek, W.D. (1992), *Nonsampling Error in Surveys.* New York: Wiley.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data* (2nd Edition). New York: Wiley.

Little, R.J.A., and Vartivarian, S.L. (2004), "Does Weighting for Nonresponse Increase the Variance of Survey Means?" University of Michigan Department of Biostatistics Working Paper Series, Working Paper 35.

Loosveldt, G., and Sonck, N. (2008), "An Evaluation of a Weighting Scheme for an Online Access Panel Survey." *Survey Research Methods*, 2 (2): 93-105.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., and Vehovar, V. (2008), "Web Surveys versus Other Survey Modes; A Meta-Analysis Comparing Response Rates." *International Journal of Market Research*, 50 (1): 79-104.

Mavletova, A. (2011), "Improving Validity in Web Surveys with Hard-to-Reach Targets: Online Respondent Drive Sampling Methodology." Paper presented at the Internet Survey Methodology Workshop, The Hague, August.

Millar, M.M., and Dillman, D.A. (2011), "Improving Response to Web and Mixed-Mode Surveys." *Public Opinion Quarterly*, 75 (2): 249-269.

Muñoz-Leiva, F., Sánchez-Fernández, J., Montoro-Ríos, F., and Ibáñez-Zapata, J.A. (2010), "Improving the Response Rate and Quality in Web-Based Surveys through the Personalization and Frequency of Reminder Mailings." *Quality and Quantity*, 44: 1037-1052.

National Telecommunications and Information Administration (NTIA) (1998), *Falling Through the Net II: New Data on the Digital Divide.* Washington, DC: U.S. Department of Commerce.

National Telecommunications and Information Administration (NTIA) (1999), *Falling Through the Net: Defining the Digital Divide.* Washington, DC: U.S. Department of Commerce.

Nyiri, Z., and Clark, R.L. (2003), "Web Survey Design: Comparing Static and Dynamic Survey Instruments." Paper presented at the annual conference of the American Association for Public Opinion Research, Nashville, May.

O'Muircheartaigh, C.A. (1997), "Measurement Error in Surveys: A Historical Perspective." In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: Wiley, pp. 1-25.

Peytchev, A. (2005), "How Questionnaire Layout Induces Measurement Error." Paper presented at the annual meeting of the American Association for Public Opinion Research, Miami Beach, FL, May.

Peytchev, A., Couper, M.P., McCabe, S.E., and Crawford, S. (2006), "Web Survey Design: Paging Versus Scrolling." *Public Opinion Quarterly*, 70 (4): 596-607.

Peytchev, A., and Crawford, S. (2005), "A Typology of Real-Time Validations in Web-Based Surveys." *Social Science Computer Review*, 23 (2): 235-249.

Rivers, D. (2006), "Sample Matching; Representative Samples from Internet Panels." YouGovPolimetrix White Paper.

Rivers, D. (2007), "Sampling for Web Surveys." Paper presented at the Joint Statistical Meetings, Salt Lake City, August.

Rivers, D., and Bailey, D. (2009), "Inference from Matched Samples in the 2008 U.S. National Elections." Paper presented at AAPOR, Hollywood, FL, May. American Statistical Association: *Proceedings of the Joint Statistical Meetings*, pp. 627-639.

Rosenbaum, P.R., and Rubin, D.B. (2004), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association*, 79 (387): 516-524.

Sakshaug, J., Tourangeau, R., Krosnick, J.A., Ackermann, A., Malka, A., DeBell, M., and Turakhia, C. (2009), "Dispositions and Outcome Rates in the 'Face-to-Face Recruited Internet Survey Platform' (the FFRISP)." Paper presented at the annual meeting of the American Association for Public Opinion Research, Hollywood, FL, May.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scherpenzeel, A., and Das, M. (2011), " 'True' Longitudinal and Probability-Based Internet Panels: Evidence from the Netherlands." In M. Das, P. Ester, and L. Kaczmirek (eds.), *Social Research and the Internet*. New York: Taylor and Francis, pp 77-104.

Schonlau, M., and Kapteyn, A. (2011), "Conducting Respondent Drive Sampling on the Web: An Experimental Approach to Recruiting Challenges." Paper presented at the MESS Workshop, Oisterwijk, The Netherlands, August.

Schonlau, M., van Soest, A.H.O., and Kapteyn, A. (2007), "Are `Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?" *Survey Research Methods*, 1 (3): 155-163.

Schonlau, M., van Soest, A.H.O., Kapteyn, A., and Couper, M.P. (2009), "Selection Bias in Web Surveys and the Use of Propensity Scores." *Sociological Methods and Research*, 37 (3): 291-318.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K.H., Marcus, S.M., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S.H. (2004), "A Comparison between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey." *Social Science Computer Review*, 22 (1): 128-138.

Schuman, H. and Presser, S. (1981), *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Shih, T.-H., and Fan, X. (2008), "Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis." *Field Methods*, 20 (3): 249–271.

Smyth, J.D, Dillman, D.A., Christian, L.M., and McBride, M. (2009), "Open-Ended Questions in Web Surveys; Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality?" *Public Opinion Quarterly*, 73 (2): 325-337.

Smyth, J.D., Dillman, D.A., Christian, L.M., and O'Neill, A.C. (2010), "Using the Internet to Survey Small Towns and Communities: Limitations and Possibilities in the Early 21st Century." *American Behavioral Scientist*, 53(9): 1423-1448.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J.W., and Terhanian, G. (2001), "The Record of Internet-Based Opinion Polls in Predicting the Results of 72 Races in the November 2007 US Elections." *International Journal of Market Research*, 43 (2): 127-135.

Terhanian, G., Bremer, J., Smith, R., and Thomas, R.K. (2000), "Correcting Data from Online Surveys for the Effects of Nonrandom Selection and Nonrandom Assignment." Harris Interactive: white paper.

Terhanian, G., Smith, R., Bremer, J., and Thomas, R.K. (2001), "Exploiting Analytical Advances; Minimizing the Bias Associated with Internet-based Surveys of Non-Random Samples." *Proceedings of the ESOMAR/ARF Worldwide Online Measurement Conference and Exhibition*, Athens, Greece, June. Amsterdam: ESOMAR, pp. 247-272.

Toepoel, V. (2011), "Panel Recruitment via Facebook." Paper presented at the Internet Survey Methodology Workshop, The Hague, August.

Toepoel, V., and Couper, M.P. (2011), "Can Verbal Instructions Counteract Visual Context Effects in Web Surveys?" *Public Opinion Quarterly*, 75 (1): 1-18.

Toepoel, V., Das, M., and van Soest, A. (2009), "Design of Web Questionnaires: The Effects of the Number of Items per Screen." *Field Methods*, 21 (2): 200-213.

Toepoel, V., and Dillman, D.A. (2011), "Words, Numbers and Visual Heuristics in Web Surveys: Is There a Hierarchy of Importance?" *Social Science Computer Review*, 29 (2): 193-207.

Tourangeau, R., Conrad, F.G. Couper, M.P., Redline, C., and Ye, C. (2009), "The Effects of Providing Examples: Questions About Frequencies and Ethnicity Background." Paper presented at the annual meeting of the American Association for Public Opinion Research, Hollywood, FL, May.

Tourangeau, R., Couper, M.P., and Conrad, F.G. (2004), "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly*, 68 (3): 368-393.

Tourangeau, R., Couper, M.P., and Conrad, F. (2007), "Color, Labels, and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly*, 71 (1): 91-112.

Tourkin, S., Parmer, R., Cox, S., and Zukerberg, A. (2005), "(Inter) Net Gain? Experiments to Increase Response." Paper presented at the annual meeting of the American Association for Public Opinion Research, Miami Beach, FL, May.

Van der Horst, W., Snijders, C., and Matzat, U. (2006), "The Effect of Progress Indicators in Online Survey Compliance." Paper presented at the General Online Research conference, Bielefeld, Germany, March.

Vehovar, V., Lozar Manfreda, K., and Batagelj, Z. (2001), "Errors in Web Surveys." *Proceedings of the 53rd Session of the International Statistical Institute*, Seoul, Korea, August.

Vonk, T., Willems, P., and van Ossenbruggen, R. (2006), *Uitkomsten Nederlands Onlinepanel Vergelijkingsonderzoek (NOPVO)*. Amsterdam: Markt Onderzoek Associatie, www.moaweb.nl/nopvo

Wejnert, C., and Heckathorn, D.D. (2008). "Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research." *Sociological Methods & Research*, 37(1): 105-134.

Werner, P. (2005), *On the Cost-Efficiency of Probability Sampling Based Mail Surveys with a Web Response Option*. Department of Mathematics, Linköpings University, Sweden: Ph.D. dissertation.

Yan, T., Conrad, F.G., Tourangeau, R., and Couper, M.P. (2011), "Should I Stay or Should I Go: The Effects of Progress Feedback, Promised Task Duration, and Length of Questionnaire on Completing Web Surveys." *International Journal of Public Opinion Research*, 23 (2): 131-147.

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levindusky, M.S., Simpser, A., and Wang, R. (2009), "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." Stanford University: unpublished paper.

Yeager, D.S., Carter, A., Tewoldemedhin, H., and Krosnick, J.A., (2010), "Study of Non-Probability Sample Internet Surveys' Estimates of Consumer Product Usage and Demographic Characteristics of Consumer Product Users." Paper presented at the annual conference of the American Association for Public Opinion Research, Chicago, May.

Yoshimura, O. (2004), "Adjusting Responses in a Non-Probability Web Panel Survey by the Propensity Score Weighting." *Proceedings of the Survey Research Methods Section*. Alexandria, VA: American Statistical Association, pp. 4660-4665.

Erakunde autonomiaduna
Organismo Autónomo del

EUSKO JAURLARITZA
GOBIERNO VASCO

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

www.eustat.es