PART 5 / IMPROVING DATA QUALITY

GREAT RESULTS FROM AMBIGUOUS SOURCES

CLEANING INTERNET PANEL DATA

Theo Downes-Le Guin
Joanne Mechling
Reg Baker

PREFACE

Rapidly increasing reliance on Internet panels as a sample source has led to data quality problems. Validation and data cleaning procedures common from telephone and in-person interviewing have not kept pace with these new methods. This paper details common forms of data quality problems resulting from the use of Internet panels, such as indiscriminate selection of responses in qualifying questions to maximize chance of qualification, and how these problems can be addressed in survey design and analysis.

THE PROBLEM

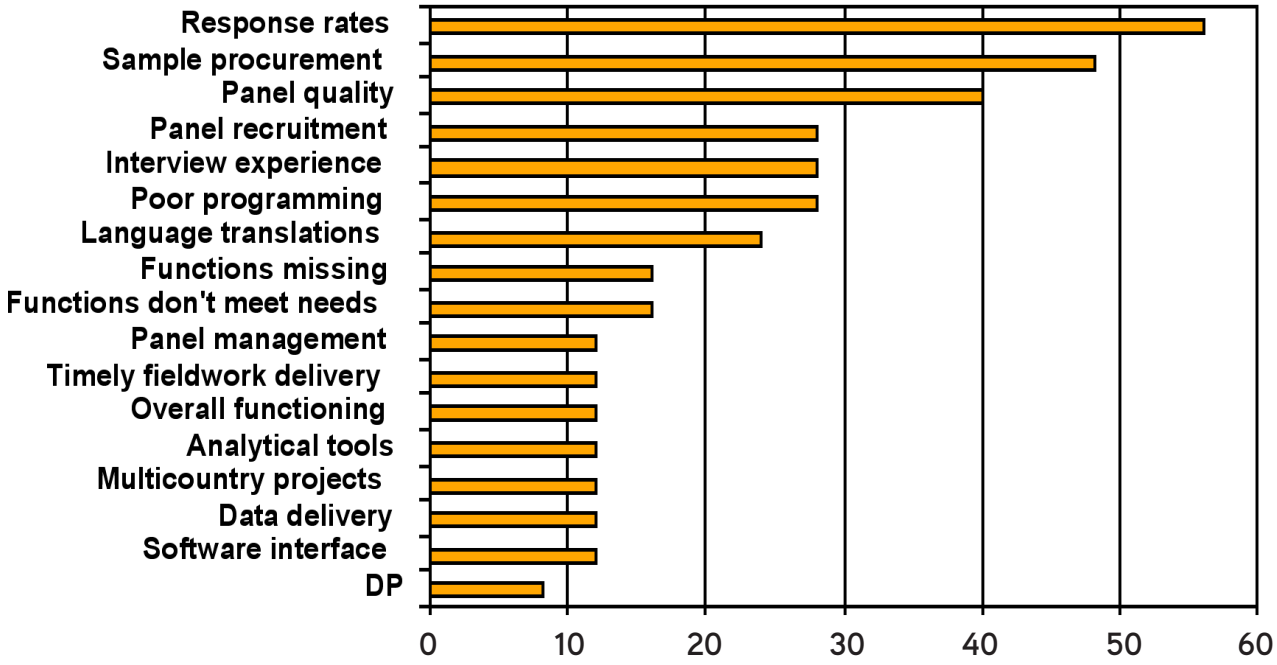
Online research today faces the prospect of a perfect storm – a collision of forces which individually are manageable but when combined may well threaten the validity of the online methodology that has evolved over the last ten years. At the core of that methodology is the substitution of online access panels for probability samples. While few would argue that panel research can match the statistical precision achievable with probability samples, our industry has come to believe that studies relying on panels are, in many cases, at least "good enough" to replace increasingly expensive and unacceptably slow probability-based surveys, often with embarrassingly low response rates. The bargain has seemed a good one, but that could change. There are at least five such worrisome forces now at work.

The first force is increasing client concerns about sample quality. After a period of relatively easy acceptance of the validity of online research, clients have begun to ask some difficult questions about panel quality. Anecdotally, we experience this on a regular basis in our businesses. Clients are asking more and tougher questions about the online panels that we use. Industry prognosticators have begun to cite it as an emerging concern.[1] More empirically, a recent survey by Cambiar identified panel quality as among the top concerns of clients, along with response rates, sample procurement, panel recruitment, and panel management (see figure 1).

The second force is the tendency for panel providers to overuse the very resource on which their business depends. While most major panel providers place restrictions on the frequency of participation by any one panelist, these restrictions vary widely. We know anecdotally that many panel members receive multiple survey invitations per week and even sometimes per day. Conceptually we understand that there is a likely though unknown limit to the number of people in any population who will join panels and respond to surveys on a regular basis. Individuals with characteristics of interest to specific clients may be especially difficult to come by. Examples include: high income households; IT decisions makers in large companies; 18-24 year olds; and physicians in some specialties.

Even accounting for high non-response in probability samples, the relatively high coverage level of probability sample sources and the sampling procedure makes it less likely that an individual will be "over-surveyed." However, in a study of seven prominent US online

FIGURE 1
CLIENT CONCERNS



Source: Cambiar, 2005

panels Krosnick et al (2005)[2] found that the median number of surveys taken in the previous year varied from six to 31. The comparable number for RDD telephone was one. A series of methodological studies conducted by Market Strategies with the same panel provider and the same target respondent over the period from 2001 to 2005 found a substantial in increase in survey taking experience and an equally substantial fall-off in response rate (see figure 2).

The third force is the impact of frequent survey taking on response quality. There is increasing evidence that responses to key questions on issues such as propensity to buy may be impacted either by panel tenure or recent survey taking experience. For example, in a paper presented at the predecessor to this conference in 2005 Coen and his colleagues[3] documented this phenomenon (see figure 3.).

The obvious implication is that outcomes on key measures for an individual survey may vary depending on the distribution of survey taking experience in the sample. As panel tenure increases and survey invitations increase these variances can be expected to become more extreme.

A fourth force is client preferences for longer and more complex surveys. Clients always have pushed for longer surveys with the cost of administration being a key limiting force. On the Web, length of interview is a less powerful cost driver, especially when panelist incentives are only loosely scaled to the interview task and length, or when incentives are altered mid-fieldwork in response to panelist participation rates. Web also has enabled a broader use of complex methodologies (for example, conjoint) which have tended to increase survey difficulty as well as extend their length. We know from research in other survey modes that as a survey wears on respondents are sometimes less likely to give full cognitive effort, a phenomenon known as "satisficing" (Krosnick, 1991).[4] Two recent studies (Lugtigheid and Rathod (2005)[5] and Gilasic (2006))[6] have documented the degree to which respondent behavior changes

FIGURE 2
SURVEY PARTICIPATION AND RESPONSE RATES

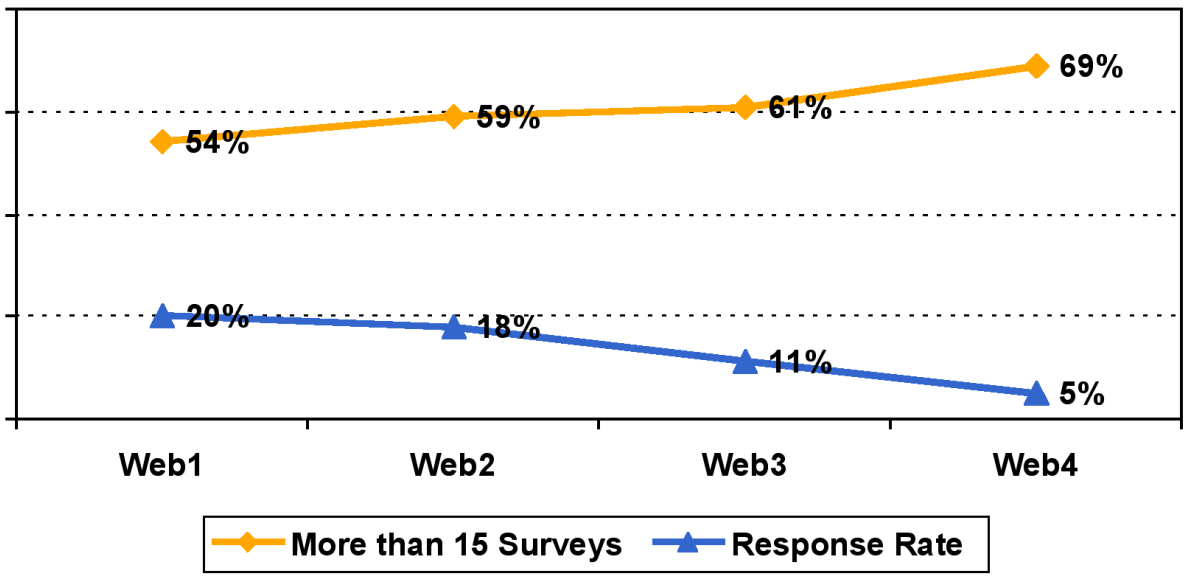
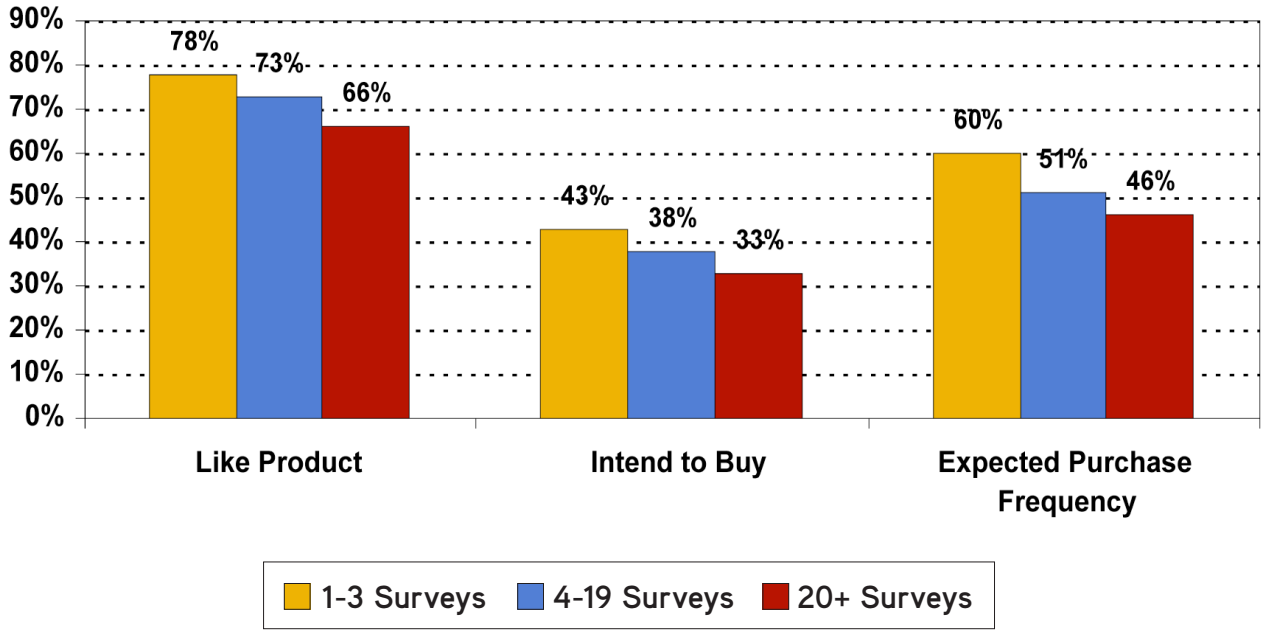


FIGURE 3
EFFECT OF SURVEY FREQUENCY ON RESPONSE



Source: Coen et al., 2005

as survey length increases. Specifically, these studies have shown that the further a respondent gets into the questionnaire the more likely he or she is to:

- Spend less time answering individual questions;
- Select Don't Know or other non-substantive responses;
- Key fewer characters in open ends.

Lugtigheid and Rathod suggest that these effects begin to get serious after about 20 minutes.

Finally a fifth force that, while not specific to panels, nonetheless has come to the fore with the emergence of online research as a prominent methodology. That force is mode effects. In many cases Web is being substituted for telephone, a transition that means two important aspects of the interview experience change. Administration by an interviewer is replaced by self administration. Questions are read and answered on a computer screen, rather than heard and responded to aurally. Both of these changes have been shown to affect how respondents answer (See, for example, Dillman (2001)).[7]

POTENTIAL SOLUTIONS

To date researchers have tended to put much of the responsibility for maintaining response quality from panel respondents on the backs of panel vendors. Although the response has been arguably dilatory, most major panel providers now disclose (at least on request) their internal guidelines for maintaining panelist integrity and response quality, and many are actively developing internal or cross-industry efforts at validation, such as Luth Research's proprietary validation system, GMI's PureSample initiative, and e-Reward's Dynamic Profile Enrichment.

At the same time, organizations such as ESOMAR, EFAMRO, CASRO, and the IMRO subcommittee or the US Market Research Association have developed guidelines or standards both for the conduct of online research and for panel development and maintenance. In addition, the Technical Committee established to develop ISO 20255, The International Standard for Marketing, Social, and Public Opinion Research, has spawned a working group with the specific mission to develop a standard for panels.

These attempts at industry standards and self-regulation are admirable, but even the most stringent standards or most ethically focused of panel companies cannot guarantee that among their millions of panel members there are not those who will sometimes misrepresent their qualifications to earn an incentive or give considerably less than full cognitive effort to every question on every survey. It seems clear to us that we are at a point where individual researchers on both the client and full-service vendor sides must assume a greater share of the responsibility and develop techniques to ensure the highest possible quality on studies using online panels. (The exception to this is the cases of sophisticated fraud in which panelists using either painstaking manual processes, or software automation, to create multiple panel profiles and populate multiple surveys in order to receive multiple incentives; this form of fraud can be difficult for panel providers' clients to detect and can be managed largely through use of physical address verification by the panel provider.)

In order to assume greater responsibility for minimizing sampling and response error in panel-based studies, researchers (whether client- or vendor- side) first need to understand the factors that can contribute to data quality issues. In panel-based surveys these factors can occur at the point of qualification (when unqualified or under-qualified panelists can become respondents), during survey taking (when satisficing behaviors can occur), and during data cleaning (which represents the last point at which the researcher can identify and correct for data quality issues caused by panelists).

Harris Interactive has created a useful though not mutually exclusive taxonomy for types of undesirable panelist behaviors, which includes fraudulents (those who intentionally misrepresent themselves to maximize earned incentives), inattentives (similar to what we term satisficers in this paper, those who

provide inadequately well-thought out responses) and hyperactive respondents who participate in numerous surveys and belong to multiple panels.[8] The first two of these categories can be identified on a study-by-study basis; the third can only be monitored by panel providers or estimated in a given study.[9] Whether focused on properly qualifying panelists or cleaning up the mess that poorly-qualified panelists can create, the goal is to identify and flag response behaviors and patterns that indicate varying degrees of satisficing, up to and including intentional misreporting by a completely unqualified panelist whose sole objective is to receive an incentive for participation. Four categories of response behaviors that betray likely sampling and response error are commonly observed:

- Indiscriminate selection of responses in qualifying questions to maximize chance of qualification.
- Obvious internal inconsistencies for responses to questions on related topics.
- Very speedy interview completion, as evidenced by programmed time stamps that ideally account for respondent suspensions and periods of server-side inactivity, to reduce the chance that a sophisticated respondent simply leaves the survey inactive for a period of time to increase his or her apparent time to completion.
- Low attention to task on open-ended questions, resulting in gibberish, cryptic text or apparently meaningful but repeated "copy and paste" responses.
- Low attention to task on tabular-style or grid questions, resulting in extensive "straightlining," in which a respondent selects the same response in a vertical pattern on the screen (for example, selecting all ones in a nine-point rating scale).

A simple but systematic process to address these behaviors can result in a dramatic improvement in data quality. In essence, the researcher defines a set of criteria (based on heuristics of what defines a "good" respondent as suggested above) unique to each study, and either screens out and/or eliminates from the dataset any response or respondent who does not meet these criteria. These criteria may be absolute filters (for example, deleting cases in the tenth percentile of interview length, adjusted for those who legitimately follow a shorter survey path due to skip patterns), or criteria such as internal consistency of responses that only in combination suggest that a panelist is misrepresenting himself.

In the following paragraphs we offer three case studies that illustrate the extent of data quality problems in typical panel-based surveys and how these problems can be addressed in screening, survey implementation and data cleaning.

Case Study: Maximizing the Chance of Qualification
In this US only sample, we sought respondents who had been involved in choosing and/or influencing the decision to buy printing devices from two vendors in the past six months for their companies. Sample was provided by a commercial Internet panel with a significant presence in the US and overseas, with a final sample of N=933 respondents post-cleaning. Knowing that the likely incidence of product ownership and respondent influence might be overstated by some panelists, we closely examined responses to qualifying questions post hoc. In so doing we observed that:

- Forty-six respondents reported having more printer brands represented among their installed base than they had actual printers.
- Nineteen respondents reported having more PC brands represented among their installed base than they had actual PCs.
- Eighty-five respondents reported that they had more printers than PCs.
- One hundred thirty respondents reported had PC-to-employee ratios higher than one-to-one; including respondents who reported 200 PCs for their two employees, 499 PCs for five employees and 356 PCs for 50 employees.

These relatively simple checks of consistency required no changes in questionnaire design and only minimal review of qualifying data, but resulted in deletion of more than 10% of the original data due to concerns about response quality. While effective, this approach is arguably the least cost effective form of remediation, especially when compared keeping unqualified or under-

qualified panelists from participating in the first place, similar to closing the barn door after the horse has escaped (or in this case, perhaps, closing the barn door after an undesirable horse has entered).

Case Study: Internal Inconsistencies

In this US-only study, we used a financial services client's customer list paired with a commercial Internet panel as the sample sources. Respondents were intended to be business decision-makers who are involved in determining what forms of payments their business accepts, with a focus on businesses that accept electronic payment forms such as debit cards. The screening questions included a multiple response list of 11 payment methods, in which we asked respondents to indicate which their businesses currently accepted as forms of payment. For respondents whose businesses qualified as accepting electronic payments, we followed up with a question about the proportion of their payments in a given period of time that came from each of the payment types they accept.

As table 1 indicates, the prevalence of inconsistency between the screening question and subsequent questions about electronic payment acceptance was minimal in the customer lists, but widespread in the Internet panel. Nearly three in 10 respondents from the Internet panel who had chosen multiple forms of electronic payments in the qualifying question subsequently reported that their typical monthly transactions employed no electronic forms of payment.

Given that the focus of the research was on business' use of electronic payments, we opted to exclude those respondents who reported not using electronic payments in a typical month. After discussions with the client, respondents for whom electronic payments constituted less than 5% their monthly transactions were also excluded due to the client concerns about misreporting; typically, if a business has invested in the infrastructure of accepting electronic payments, electronic payment transactions will account for a larger proportion of customer payments than cash and traditional checks.[10] Thus a total of 42% of cases from the Internet panel sample were deemed unusable, as compared to less than 3% from the client's customer list.

Case study: Selection of very low-incidence qualifying criteria

In this multinational (North America, Western Europe, Asia Pacific) study, we sought business and home decision-makers for a low-incidence form of printing device generally priced at a level that is attractive to businesses but less so to consumers. In North America and Europe, sample was provided purely by commercial Internet panels; in Asia Pacific, sample was provided from a variety of sources including vendor panels and home and business directories, with participants

TABLE 1

	Client-supplied customer list (N=1855)	**Commercial Internet panel** (N=867)
Although respondent had indicated the business accepted electronic payments, s/he told us that of the total number of customer payments or sales transactions received by his/her business in a typical month...		
• None were electronic transactions	<1%	28%
• At least some, but less than 5%, were electronic transactions	2%	14%

recruited first by phone for qualifying and then directed to the Web to complete the full survey. The post-cleaning sample size was N=1,216 consumer and business decision-makers.

As in the case studies above, we used a combination of consistency checks and common-sense review of screening and questionnaire data to develop criteria for case deletion. A number of criteria were used, but the most discriminating criteria were:

- The number of home technology products a respondent reported owning, including products inserted specifically to catch panelists attempting to maximize chance of participation.
- Profiling and enumerating of the installed printer base in respondents' homes/businesses.

Close examination of these results showed large differences in the performance of different sample sources, with Internet panels in general, and US panels specifically generating the highest rate of deleted cases as compared to client-supplied customer lists and phone-recruited Web completions. In table 2, we note that a total of 30% of cases were deleted, accounting for all criteria in the US Internet panel subsample, with only 5% deleted from the Asia-Pacific phone subsample. Almost half of these cases in the US came solely from respondents who claimed ownership of 10 out of 10 items in a screening question, including a Segway Human Transporter (an expensive electronic device that likely ships in the hundreds of units a year for non-fleet sales).

CONCLUSIONS

We began this paper with a meteorological metaphor that suggests a looming crisis in the research industry. Is it a gross exaggeration to characterize the survey panel ecosystem as imperiled, at least in terms of the ability of Internet panels to offer an appropriate balance of survey costs and errors for the average commercial research buyer? Just as research vendors and clients alike adopted Internet panels very quickly and with a minimum of methodological fastidiousness, it is conceivable that the tide will turn back toward phone and even mail or pseudo-ethnographic approaches as a response to the frustration with sample and data quality.

At the same time, it is hard to imagine anything approaching wholesale abandonment of the panel model. The research industry's compact with respondents has changed significantly over the past 10 years, evolving away from one based on the willingness of respondents to exchange their time and opinions in the short term for the promise of improved goods and services in the long term, toward a straightforward exchange of information for goods or cash. To be sure, curiosity, helpfulness and the desire for information received in the process of participating in research are still significant reasons that some populations participate in research on some topic areas. But the fundamental premise of the panel remains intact: we can guarantee a level of participation and quality of response by explicitly promising a return on the respondents' investment of time (whether that

TABLE 2

Criterion (# cases collected)	**US client** (N=105)	**US panel** (N=458)	**EU panel** (N=408)	**Asia phone** (N=245)
In the household technology products ownership profiling: Segway (asked in US only) or Sony AIBO owner who had 10+ items selected	n=0	n=65 (14%)	n=0	n=0
Four or more photo printers at home	n=6 (6%)	n=28 (6%)	n=16 (4%)	n=2 (1%)
Four or more large format photo printers at home	n=0	n=24 (5%)	n=12 (3%)	n=2 (1%)
Total cases we ended up deleting	N=10 (10%)	n=138 (30%)	n=65 (21%)	n=13 (5%)

return be information, entertainment, gifts or cash). After a long grace period, the researcher/respondent interaction has begun to submit to the norms of the market economy.

Fortunately for practitioners, the panel ecosystem is still largely intact and minimal efforts toward understanding and correcting for common data quality issues in panels can yield big results. The process of understanding the range of fraudulent, inattentive and hyperactive behaviors, flagging (on a study-by-study basis) these behaviors during and after screening, and cleaning datasets is both simple to employ and common-sensical for anyone versed in the norms of data cleaning from other modalities. Indeed, based on its own experience, Doxus estimates that the entire process, once codified and disseminated amongst a research staff, and automated syntax created to review interim datasets, these processes generate an additional "overhead" of perhaps two to four hours in questionnaire design and four to six hours in data cleaning for a survey of average sample size and questionnaire complexity.

Once codified, vendors and clients can work with panel providers to improve the sustainability of the ecosystem. Many end clients, once aware of the issues, are more than willing to pay the small (often less than 1%) price premium required to take these steps to maintain data quality. And panel vendors are more than happy to replace or remove panel access fees for respondents whose data they agree is suspect. More importantly to the long-term health of the panels, most larger providers use information provided by their clients as part of algorithms that determine how panelists will be treated going forward – ranging from expulsion, to permanent and temporary quarantine, to gentle warnings.

Finally, researchers need to be judicious in how they approach survey design intended to "trap" fraudulent and inattentive panelists. Panel providers, research firms and end clients all have a vested interest in keeping the survey experience appealing for conscientious panelists, and many of the approaches at our disposal, such as very low-incidence response options, ratings verifications in tabular questions, and response-sensitive prompts (for example, pop-ups along the lines of "did you really mean to select all Don't Knows in this table?) increase questionnaire length. As well, these techniques may be so obvious as to simultaneously annoy the majority of conscientious respondents while teaching those who are less conscientious exactly how to avoid a new generation of approaches designed to identify and exclude them.

Footnotes

1. *Research Business Report,* January 2006.

2. John Krosnick, Norman Nie, and Douglas Rivers, "Comparing Major Survey Firms in Terms of Survey Satisficing: Telephone and Internet Data Collection," AAPOR Annual Conference (2005).

3. Terrence Coen, Jacqueline Lorch, and Linda Piekarski, "The Effects of Survey Frequency on Panelists' Responses," ESOMAR Worldwide Panel Research, (2005).

4. Jon Krosnick, "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology,* (1991), 5, pp. 213-236.

5. Arthur Lugtigheid and Sandra Rathod, *Questionnaire Length and Response Quality: Myth or Reality,* Survey Sampling International, 2005.

6. Mirta Gilasic, "The Effects of Questionnaire Length on Quality of Data in a Web Survey," Joint Program in Survey Methodology, University of Maryland (2006).

7. Don A. Dillman and Leah Melani Christian, "Survey Mode as a Source of Instability across Surveys," *Field Methods,* (2005), 17, pp. 30-52.

8. Renee Smith and Holland Hofma Brown, "Assessing the Quality of Data from Online Panels: Moving Forward with Confidence," *Harris Interactive* (2005).

9. To this end, Doxus has included a question at the end of all panel surveys during the past nine months aimed at identifying panelists who belong to more than one panel and thus received more than one invitation to the same study. These results, to be presented in a forthcoming paper, suggest that for IT professionals, we can expect 1% - 3% "repeat respondents" on average.

10. As well, analysis of within-respondent responses indicated that the excluded cases were established businesses, reducing or eliminating the possibility that the business was not doing larger volumes of electronic payments simply because it was setting up for greater volumes of electronic payments at a later date.

The Authors

Theo Downes-Le Guin is Principal. Doxus LLC, United States.

Joanne Mechling is Senior Methodologist, Doxus LLC, United States.

Reg Baker is Chief Operating Officer, Market Strategies, Inc., United States.