Big Data and Survey Research: Supplement or Substitute?

Timothy P. Johnson and Tom W. Smith

Abstract The increasing availability of organic Big Data has prompted questions regarding its usefulness as an auxiliary data source that can enhance the value of design-based survey data, or possibly serve as a replacement for it. Big Data's potential value as a substitute for survey data is largely driven by recognition of the potential cost savings associated with a transition from reliance on expensive and often slow-to-complete survey data collection to reliance on far less-costly and readily available Big Data sources. There may be, of course, serious methodological costs of doing so. We review and compare the advantages and disadvantages of survey-based vs. Big Data-based methodologies, concluding that each data source has unique qualities and that future efforts to find ways of integrating data obtained from varying sources, including Big Data and survey research, are most likely to be fruitful.

Keywords Survey research • Big Data • Data quality • Design-based data • Organic data

1 Introduction

As response rates and survey participation continue to decline, and as costs of data collection continue to grow, researchers are increasingly looking for alternatives to traditional survey research methods for the collection of social science information. One approach has involved modifying scientific survey research methods through the abandonment of probability sampling techniques in favor of less expensive non-probability sampling methodologies (c.f. Cohn 2014). This strategy has

T.P. Johnson (✉)
Survey Research Laboratory, University of Illinois at Chicago, 412 S. Peoria St., Chicago, IL 60607, USA
e-mail: timj@uic.edu

T.W. Smith
General Social Survey, NORC at the University of Chicago, 1155 E 60th St., Chicago, IL 60637, USA
e-mail: Smith-tom@norc.org

P. Thakuriah et al. (eds.), *Seeing Cities Through Big Data*, Springer Geography,
DOI 10.1007/978-3-319-40902-3_7

become popular enough that the American Association for Public Opinion Research (AAPOR) recently felt it necessary to appoint a Task Force to investigate the issue and release a formal report (Baker et al. 2013). Others have explored the usefulness of supplementing, or replacing completely, surveys with information captured efficiently and inexpensively via “Big Data” electronic information systems. In this paper, we explore the advantages and disadvantages of using survey data versus Big Data for purposes of social monitoring and address the degree to which Big Data can become a supplement to survey research or a complete alternative or replacement for it.

Survey research originally evolved out of social and political needs for better understandings of human populations and social conditions (Converse 1987). Its genesis predates considerably the pre-electronic era to a time when there were few alternative sources of systematically collected information. Over the past 80 years, survey research has grown and diversified, and complex modern societies have come to increasingly rely on survey statistics for a variety of public and private purposes, including public administration and urban planning, consumer and market research, and academic investigations, to name a few. In contrast, Big Data became possible only recently with the advent of reliable, high speed and relatively inexpensive electronic systems capable of prospectively capturing vast amounts of seemingly mundane process information. In a very short period of time, Big Data has demonstrated its potential value as an alternative method of social analysis (Goel et al. 2010; Mayer-Schӧnberger and Cukier 2013).

Before proceeding further, however, it is important to define what we mean exactly by survey research and “Big Data.” Vogt (1999: 286) defines a survey as “a research design in which a sample of subjects is drawn from a population and studied (often interviewed) to make inferences about the population.” Groves (2011) classifies surveys as forms of inquiry that are “design-based,” as the specific methodology implemented for any given study is tailored (or designed) specifically to address research questions or problems of interest. In contrast, Webopedia (2014) defines Big Data as “a buzzword. . .used to describe a massive volume of both structured and unstructured data that is so large that it’s difficult to process using traditional database and software techniques.” Thakuriah et al. (2016), more carefully define Big Data as “structured and unstructured data generated naturally as a part of transactional, operational, planning and social activities, or the linkage of such data to purposefully designed data.” In addition to these attributes, Couper (2013) observes that Big Data is produced at a rapid pace. In contrast to design-based data, Groves classifies Big Data as being organic in nature. Although similar to survey data in the systematic manner in which it is collected, organic data is not typically designed to address specific research questions. Rather, such data, referred to by Harford (2014) as “digital exhaust,” is a by-product of automated processes that can be quantified and reused for other purposes. There are, of course, exceptions, such as the National Weather Service’s measurements, which are design-based and otherwise fit the definition of Big Data.

Although they do not fit today’s electronic-based definitions of Big Data, there are several examples of survey-based data sets that are uncharacteristically “big” by any reasonable standards. Examples of Big Surveys include national censuses, which routinely attempt to collect information from millions of citizens. The U.S. micro decennial Census is an example of this. Also included here is the infamous *Literary Digest* Poll, which attempted, and failed badly, to predict the outcome of the 1936 Presidential election, based on more than two million postcard responses collected from individuals sampled from published telephone directories and automobile registration lists (Squire 1988). The *Literary Digest* had been conducting similar straw polls since 1908, but did not run into trouble with a failed election prediction until 1936. The *Literary Digest* experience taught the still young survey research community of the 1930s that big does not necessarily mean better. Subsequent to that experience, survey statisticians worked to develop sampling theory, which enabled them to rely on much smaller, but more carefully selected, random samples to represent populations of interest.

2 What Distinguishes Surveys from Big Data?

While censuses and the *Literary Digest* examples share with today's Big Data large observation-to-variable ratios, they do not have Big Data's electronic-based longitudinal velocity, or rate of data accumulation. Rather, even Big Surveys are only snapshots that represent at best a brief moment in time. Perhaps even more importantly, the structures of these design-based data sources are carefully constructed, unlike many sources of Big Data, which are known for their "messy" nature (Couper 2013). Hence, there are several important differences between design-based survey data, and the organic data sources that represent Big Data. These include differences in volume, data structures, the velocity and chronicity with which data are accumulated, and the intended purposes for which the data are collected.

2.1 Volume

Big Data is big by definition. As Webopedia (2014) suggests, Big Data represents "a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques." Most of the information generated in the history of our planet has probably been produced in the past several years by automated Big Data collection systems. Google's search database alone collects literally billions of records on a daily basis and will presumably continue to do so into the foreseeable future, accumulating an almost impossibly large amount of organic information. Prewitt (2013: 229) refers to this as a "digital data tsunami." Survey data, by contrast, is many orders of magnitude more modest in volume, and as mentioned earlier, is becoming more expensive and difficult to collect.

2.2 Data Structures

By data structures, we mean the ratio of observations to variables. Big Data commonly have higher ratios (i.e., vastly more observation points than variables), and surveys have much lower ratios (i.e., many more variables but for vastly fewer observations). Prewitt (2013) describes survey data as case-poor-and-variable-rich, and Big Data as case-rich-and-variable-poor.

2.3 Velocity

Data velocity is the speed with which data is accumulated. Big Data's velocity, of course, means that it can be acquired very quickly. Not so with surveys, which require greater planning and effort, depending on mode. Well-done telephone surveys can take weeks to complete, and well-done face-to-face and mail surveys can require months of effort. Even online surveys require at least several days of effort to complete all "field" work. Where government and business decisions must be made quickly, Big Data may increasingly become the most viable option for instant analysis. Indeed, many complex organizations now employ real-time "dashboards" that display up-to-the-minute sets of indicators of organizational functioning and activity to be used for this purpose, and one of the stated advantages of Google's Flu Index (to be discussed below) and similar efforts has been the almost real-time speed with which the underlying data become available, vastly outperforming surveys, as well as most other forms of data collection. Big Data is collected so quickly, without much in the way of human intervention or maintenance, that its velocity is sometimes compared to that of water emitting from a fire hose. Survey research will continue to have difficulty competing in this arena.

2.4 Data Chronicity

Data chronicity refers to time dimensions. The chronicity of Big Data is much more continuous (or longitudinal) than that of most common cross-sectional surveys. With few exceptions, survey data are almost invariably collected over relatively short time intervals, typically over a matter of days, weeks or months. Some data collection systems for Big Data, in contrast, are now systematically collecting information on an ongoing, more or less, permanent basis. There is an often incorrect assumption that the methods, coverage and content of Big Data remains static or unchanging over time. In fact, Big Data systems are often quite changeable and hence there is a danger that time series measurements may not always be comparable.

2.5 Intended Purpose

Design-based survey data are collected to address specific research questions. There are few examples of Big Data being intentionally constructed for research purposes, mostly by governmental agencies interested in taking, for example, continuous weather or other environmental or economic measurements. Most Big Data initiatives, rather, seem driven by commercial interests. Typically, researchers have a good deal of control over the survey data they collect, whereas most analysts of Big Data are dependent on the cooperative spirit and benevolence of large corporate enterprises who collect and control the data that the researchers seek to analyze.

3 Relative Advantages of Big Data

The main advantages of Big Data over survey data collection systems are costs, timeliness and data completeness.

3.1 Costs of Data Collection

As mentioned earlier, Big Data has an important advantage in terms of data collection costs. Surveys, particularly those using an interviewer-assisted mode, continue to become increasingly expensive, whereas the costs of using available Big Data collected for other purposes may be less expensive. The cost of original collection of Big Data, though, is often very high. As research funding becomes more difficult to obtain, the economic attractiveness of Big Data make it difficult to not seriously consider it as an alternative data source.

3.2 Timeliness

As discussed earlier, the velocity of Big Data greatly exceeds that of traditional survey research. As such, it theoretically provides greater opportunities for the real-time monitoring of social, economic and environmental processes. It has been noted, however, that the processing of Big Data can in some cases be a lengthy and time-consuming process (Japec et al. 2015). In addition, being granted real-time access by the original collectors of this information is not always allowed.

3.3 Data Completeness

Missing data at both the item and unit levels is a difficult problem in survey research and the errors associated with it preoccupy many researchers. Big Data sets do not typically share this problem. Because most Big Data sets are based on varied data collection systems that do not rely directly on the participation of volunteers, and subjects are typically not even aware that they are contributing information to Big Data systems (on this point, see the section on *Ethical Oversight* below), non-observations due to failure to contact individuals, or to their unwillingness or inability to answer certain questions, or to participate at all, is not a problem. But Big Data is also not perfect, as we would expect for example that monitors and other recording devices will occasionally malfunction, rendering data streams incomplete. As with surveys, the information missing from Big Data sets may also be biased in multiple ways.

4 Relative Advantages of Survey Research

Advantages of survey research data over Big Data include its emphasis on theory, the ease of analysis, error assessment, population coverage, ethical oversight and transparency.

4.1 The Role of Theory

Some have argued that the we are facing “the end of theory,” as the advent of Big Data will make “the scientific method obsolete” (Anderson 2008). Although some of the survey research reported in the popular news media is descriptive only, much of the research conducted using survey methods is theory-driven. Survey data are routinely employed to test increasingly sophisticated and elaborate theories of the workings of our social world. Rather than allowing theory to direct their analyses, Big Data users tend to be repeating some earlier criticisms of empirical survey research by inductively searching for patterns in the data, behaviors that left earlier generations of survey researchers vulnerable to accusations of using early high-speed computers for “fishing expeditions.” Fung (2014) criticizes Big Data as being observational (without design) and lacking in the controls that design-based data typically collect and employ to rule-out competing hypotheses.

4.2 Ease of Analysis

The sheer size of many Big Data sets and their often unstructured nature make them much more difficult to analyze, compared to typical survey data files. There are numerous packaged data management and statistical analysis systems readily available to accommodate virtually any survey data set. Big Data, in contrast, typically requires large, difficult-to-access computer systems to process, and there is a shortage of experts with the knowledge and experience to manage and analyze Big Data (Ovide 2013). The time necessary to organize and clean Big Data sets may offset, to some extent, the speed advantage with which Big Data is accumulated.

4.3 Measurement Error

The error sources associated with survey data are reasonably well understood and have been the subject of robust, ongoing research initiatives for many decades (Groves et al. 2009; Schuman and Presser 1981; Sudman and Bradburn 1974). We know that the Literary Digest poll was discredited by several error sources, including coverage and nonresponse errors that have been well documented (Lusinchi 2012; Squire 1988). Errors associated with Big Data, however, are currently not well understood and efforts to systematically investigate them are only now beginning. Prewitt (2013: 230) observes that "there is no generally accepted understanding of what constitutes errors when it is machines collecting data from other machines." Measurement error is an important example. Survey measures are typically the subject of considerable research and refinement, with sophisticated methodologies readily available for the design, testing, and assessment of measurement instruments (Madans et al. 2011; Presser et al. 2004). Big Data shares many of the challenges of secondary analyses of survey data in which specific indicators of the construct(s) of interest may not always be available, challenging the analyst's creativity and cleverness to sometimes "weave a silk purse from a sow's ear." Indeed, those analyzing Big Data must work with what is available to them and there is seldom an opportunity to allow theory to drive the design of Big Data collection systems. There is also concern that those who generate Big Data are sometimes unwilling to share details of how their data are collected, to provide definitions of the terms and measures being used, and to allow replication of measurements and/or analyses based on their measurements.

One interesting example is the Google Flu Index. In 2009, a team from Google Inc. and the Centers for Disease Control and Prevention (CDC) published a paper in *Nature* that described the development of a methodology for examining billions of Google search queries in order to monitor influenza in the general population (Ginsberg et al. 2009).[1] They described a non-theoretical procedure that involved

[1] In 2008, a team of academic investigators and Yahoo! Employees published a similar paper (Polgreen et al. 2008).) That team, however, had not continued to report on this topic.

identifying those Google search queries that were most strongly correlated with influenza data from the CDC; a large number of models were fit during the development of the flu index. They reported the ability to accurately estimate weekly influenza within each region of the U.S. and to do so with only a very short time lag. Shortly thereafter, the flu index *underestimated* a non-seasonal outbreak, and researchers speculated that changes in the public's online search behaviors, possibly due to seasonality, might be responsible (Cook et al. 2011). Despite an ongoing effort to revise, update and improve the predictive power of Google Flu Trends, it also greatly *overestimated* influenza at the height of the flu season in 2011–2012 (Lazer et al. 2014a) and especially in 2012–2013 (Butler 2013). Lazer et al. (2014a) also demonstrated that Google Flu Trends had essentially overestimated flu prevalence during 100 of 108 weeks (starting with August 2011). A preliminary analysis of the 2013–2014 season suggests some improvement, although it is still *overestimating* flu prevalence (Lazer et al. 2014b).

Couper (2013) has made the interesting point that many users of social media, such as Facebook, are to some extent motivated by impression management, and we can thus not be certain of the extent to which information derived from these sources accurately represents the individuals who post information there. Social desirability bias would thus appear to be a threat to the quality of Big Data as well as survey data. The fact that a significant proportion of all Facebook accounts, for example, are believed to represent fictitious individuals is another cause for concern. One estimate from 2012 suggests the number of fake Facebook accounts may be as many as 83 million (Kelly 2012). Hence, concerns with data falsification also extend to Big Data.

4.4 Population Coverage

The *Literary Digest* Poll was big, but many believe it did not provide adequate coverage of the population to which it was attempting to make inferences. Rather, it likely over-represented upper income households with political orientations decidedly unrepresentative of the Depression Era U.S. citizenry. Clearly, volume could not compensate for or fix coverage error. Big Data faces similar problems. For Big Data that captures online activities, it is important to be reminded that not everyone is linked to the internet, not everyone on the web uses Google search engines, Twitter and Facebook, and everyone who does certainly does not do so in a similar manner. Among those who do interact with the web, the manners in which they do are very diverse. The elderly, who are less likely to engage the internet, are particularly vulnerable to influenza, yet none of the Google Flu Index papers referenced here address this issue. A related concern is the problem of selection bias. As Couper (2013) has observed, Big Data tends to focus on society's "haves" and less so on the "have-nots." In addition, in Big Data there can be a problem with potential violations of the "one-person-one-vote" rule. As Smith (2013) has commented, a large preponderance of some social media activities, such as Twitter and Facebook, are the products of the activities of relatively small concentrations of individuals, further calling in to question the adequacy of their coverage. Indeed, many Big Data systems have what Tufekci (2014) refers to as a denominator problem "created by vague, unclear or unrepresentative sampling." Others have expressed concerns regarding the danger that Big Data "can be easily gamed" (Marcus and Davis 2014). Campbell (1979) wrote more than 40 years ago about the corruptibility of social data as it becomes more relevant to resource allocation decisions. Marcus and Davis (2014) discuss several Big Data examples of this. Design-based, "small data" surveys, in comparison, go to great lengths to insure that their samples adequately cover the population of interest.

4.5 Ethical Oversight

Unlike survey researcher's insistence on obtaining informed consent from respondents prior to data collection, and emphasis on the distribution of de-identified data only, many Big Data operations routinely collect identifying information without the consent, or even the knowledge, of those being monitored. In comparison to the careful ethical reviews and oversight academic and government-based survey research routinely receives, the ethical issues surrounding Big Data are not yet well understood or recognized. There is little transparency or oversight in Big Data research, much of it being conducted by private groups using proprietary data.

Unfortunately, recent events, such as Facebook's mood experiments (Albergott and Dwoskin 2014), are reminiscent of some of the ethical transgressions of past generations that led to ethical review requirements for federally funded research (Humphreys 1970; Milgram 1974). For example, in 2014, Kramer et al. (2014) published in the *Proceedings of the National Academy of Sciences (PNAS)* findings from a field experiment that examined the extent to which emotional states could be manipulated by altering the content that Facebook users were exposed to. They demonstrated that reductions in displays of emotionally negative postings from others resulted in both reductions in the amount of positive posting and increases in negative emotional postings among the Facebook users being monitored. The paper's authors reported that 689,003 individuals and more than three million Facebook postings were studied as part of the experiment. Shortly after the paper's publication, *PNAS* published an "Editorial Expression of Concern and Correction," acknowledging potential contradictions between established ethical principles of research conduct—specifically the degree to which the Facebook study subjects had sufficient opportunity to provide informed consent and/or to opt out of the research—and the Facebook data user policies, under which users agree to corporate use of their data at the time they establish their personal account. In response to ambiguities such as these, some have called for a new Big Data Code of Ethical Practices (Rayport 2011). The National Science Foundation recognized this need and launched a Council for Big Data, Ethics, and Society in early 2014 "to provide critical social and cultural perspectives on big data initiatives" (see: http://www.

datasociety.net/initiatives/council-for-big-data-ethics-and-society/). There is no consensus, however, regarding the ethical issues surrounding cases such as the Facebook experiments (Puschmann and Bozdag 2014).

4.6 Transparency

Transparency of methods is central to the ability to replicate research findings. There is a need for greater and more general understanding of how Big Data sets are constructed (Mayer-Schönberger and Cukier 2013). Big Data is not yet transparent, and most Big Data is proprietary and commercially controlled, and the methods employed to analyze these data are seldom described in a manner that would facilitate replication. In fact, commercial interests often dictate against transparency. The Google Flu Index, for example, has never revealed the 45 or so search terms it uses to make its prevalence estimates. Lazer et al. (2014b) have accused Google of reporting misleading information regarding the search terms they employ. While survey research is far from perfect when it comes to transparency of methods, there is general recognition of its importance. Most high-quality professional journals demand disclosure of survey methods. In 2010, AAPOR launched a Transparency Initiative, designed "to promote methodological disclosure through a proactive, educational approach that assists survey organizations in developing simple and efficient means for routinely disclosing the research methods associated with their publicly-released studies" (see: http://www.aapor.org/). In addition, codebooks, methodological reports, and other forms of documentation are considered to be standard products of any reputable survey, and have been so for many decades. The documentation requirements of social science data archives, such as the Inter-University Consortium of Social and Political Research (ICPSR; see http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html) are very stringent. Documentation of internet data, by comparison, is extremely limited (Smith 2013).

5 Supplement or Substitute?

Lazer and colleagues (2014a: 1203) have coined the term "Big Data Hubris" to refer to "the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis." Others share this sentiment. The British sociologists Savage and Burrows (2007: 890) have considered the historicity of survey research and suggest that its "glory years" were between 1950 and 1990. Taking the long view, one has to wonder as to whether or not surveys might merely represent one of the first generations of social research methods, destined to be replaced by more efficient methodologies in an increasingly digital world? Just as the horse-drawn carriage was replaced by more advanced forms of transportation, might we be now witnessing the passing of a traditional methodology?

Only time will tell. Big Data, in its current stage of evolution, though, does not appear capable of serving as a wholesale replacement or substitute for survey research. Even Savage and Burrows (2007: 890) acknowledge that there are some niches “in which the sample survey will continue to be a central research tool because of the limits of transactional data” (i.e., Big Data). They cite crime victimization surveys, which consistently demonstrate victimization rates well in excess of estimates derived from administrative records. There are no doubt many other examples. But, Big Data is an important new and highly valuable source of information about our social world, one with the potential to help us examine and better understand social problems, including many of those being addressed in this book. So how do we reconcile small surveys with Big Data?

Several observers, including another AAPOR Task Force concerned specifically with the rise of Big Data (Japec et al. 2015), see important opportunities for surveys and Big Data to be supplements or adjuncts to one another (Butler 2013; Couper 2013; Marcus and Davis 2014; Smith 2011; 2013); for Big Data to contribute rich context to surveys, and for surveys to help make sense of patterns uncovered, but not well understood, in Big Data. Combining multiple data sources to take advantage of the strengths of each and to help compensate for the limits of each approach, seems to be what the future holds for these largely unique data resources. Smith and Kim (2014) have proposed a multi-level, multi-source (ML-MS) approach to reducing survey-related errors through a coordinated effort to more systematically link survey data with information from multiple auxiliary sources, including Big Data. These linkages would take place at each possible level of analysis, from high levels of geographies through unique paradata sources that are themselves by-products of survey data collection activities, such as contact attempts and even computer key-stroke data from interviewers and/or respondents (c.f., Kreuter 2013). In addition to private Big Data, administrative data files from governmental sources would also be linked to develop better understandings of social phenomena and the strengths and limitations of the various data sources themselves. As the former U.S. Census Bureau Director Robert Groves (2011: 869) has commented: “combining data sources to produce new information not contained in any single sources is the future.”

References

Albergott R, Dwoskin E (2014) Facebook study sparks soul-searching and ethical questions. Wall Street J. [Online] 30th June. http://online.wsj.com/articles/facebook-study-sparks-ethical-questions-1404172292. Accessed 3 Aug 2014

Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine, 23rd June. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory. Accessed 29 Jul 2014

Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, Gile KJ, Tourangeau R (2013) Report of the AAPOR task force on non-probability sampling. http://www.aapor.org/AM/Template.cfm?Section=Reports1&Template=/CM/ContentDisplay.cfm&ContentID=5963. Accessed 1 Aug 2014

Butler D (2013) When Google got flu wrong. Nature 494:155–156

Campbell DT (1979) Assessing the impact of planned social change. Eval Program Plann 2:67–90

Cohn N (2014) Explaining online panels and the 2014 midterms. New York Times. [Online] 27 July. http://www.nytimes.com/2014/07/28/upshot/explaining-online-panels-and-the-2014-midterms.html?_r=0. Accessed 1 Aug 2014

Converse JM (1987) Survey research in the United States: roots and emergence 1890-1960. University of California Press, Berkeley

Cook S, Conrad C, Fowlkes AL, Mohebbi MH (2011) Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One 6:e23610

Couper MP (2013) Is the sky falling? New technology, changing media, and the future of surveys. Surv Res Methods 7:145–156

Fung K (2014) Google flu trends' failure shows good data > big data. Harvard Business Review/HBR Blog Network. [Online] 25 March. http://blogs.hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/. Accessed 17 Jun 2014

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinsji MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014

Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with web search. PNAS 107:17486–17490, http://www.pnas.org/content/107/41/17486. Accessed 15 Aug 2015

Groves RM (2011) Three eras of survey research. Public Opin Quart 75:861–871

Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2009) Survey methodology, 2nd edn. Wiley, New York

Harford T (2014) Big data: are we making a big mistake? Financial Times. [Online] 26 March. http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz39NlqxnU8. Accessed 17 Jun 2014

Humphreys L (1970) Tearoom trade: impersonal sex in public places. Duckworth, London

Japec L, Kreuter F, Berg M, Biemer P, Decker P, Lampe C, Lane J, O'neil C, Usher A (2015) AAPOR report on big data. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf. Accessed: 27 Jul 2015

Kelly H (2012) 83 million Facebook accounts are fakes and dupes. CNN Tech. [Online] 2 August. http://www.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/. Accessed 3 Aug 2014

Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. PNAS 111:8788–8790, http://www.pnas.org/content/111/24/8788.full. Accessed 2 Aug 2014

Kreuter F (2013) Improving surveys with paradata: analytic uses of process information. Wiley, New York

Lazer D, Kennedy R, King G, Vespignani A (2014a) The parable of Google flu: traps in big data analysis. Science 343:1203–1205

Lazer D, Kennedy R, King G, Vespignani A (2014b) Google flu trends still appears sick: an evaluation of the 2013-2014 flu season. [Online] http://ssrn.com/abstract=2408560. Accessed 26 Jul 2014

Lusinchi D (2012) "President" Landon and the 1936 *Literary Digest* Poll: were automobile and telephone owners to blame? Soc Sci Hist 36:23–54

Madans J, Miller K, Maitland A, Willis G (2011) Question evaluation methods: contributing to the science of data quality. Wiley, New York

Marcus G, Davis E (2014) Eight (no, nine!) problems with big data. New York Times. [Online] 7 April. http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=0. Accessed 7 Aug 2014

Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, New York

Milgram S (1974) Obedience to authority: an experimental view. Harper, New York

Ovide S (2013) Big data, big blunders. Wall Street Journal. [Online] 10 March. http://online.wsj.com/news/articles/SB10001424127887324196204578298381588348290. Accessed: 30 Jul 2014

Polgreen PM, Chen Y, Pennock DM, Nelson FD (2008) Using internet searches for influenza surveillance. Clin Infect Dis 47:1443–1448

Presser S, Rothgeb JM, Couper MP, Lessler JT, Martin E, Martin J, Singer E (2004) Methods for testing and evaluating survey questionnaires. Wiley, New York

Prewitt K (2013) The 2012 Morris Hansen lecture: Thank you Morris, et al., for Westat, et al. J Off Stat 29:223–231

Puschmann C, Bozdag E (2014) Staking out the unclear ethical terrain of online social experiments. Internet Policy Review 3(4). http://policyreview.info/articles/analysis/staking-out-unclear-ethical-terrain-online-social-experiments. Accessed: 5 Aug 2016

Rayport JF (2011) What big data needs: a code of ethnical practices. MIT Technology Review, May 26. http://www.technologyreview.com/news/424104/what-big-data-needs-a-code-of-ethical-practices/. Accessed 2 Aug 2014

Savage M, Burrows R (2007) The coming crisis of empirical sociology. Sociology 41:885–899

Schuman H, Presser S (1981) Questions and Answers in Attitude Surveys. Wiley, New York.

Smith TW (2011) The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. Int J Public Opin Res 23:389–402

Smith TW (2013) Survey-research paradigms old and new. Int J Public Opin Res 25:218–229

Smith TW, Kim J (2014) The multi-level, multi-source (ML-MS) approach to improving survey research. GSS Methodological Report 121. NORC at the University of Chicago, Chicago

Squire P (1988) Why the 1936 *Literary Digest* poll failed. Public Opin Quart 52:125–133

Sudman S, Bradburn NM (1974) Response effects in surveys: a review and synthesis. Aldine Press, Chicago

Thakuriah P, Tilahun N, Zellner M (2016) Big data and urban informatics: innovations and challenges to urban planning and knowledge discovery. In: Thakuriah P, Tilahun N, Zellner M (eds) Seeing cities through big data: research methods and applications in urban informatics. Springer, New York

Tufekci Z (2014) Big questions for social media big data: representativeness, validity and other methodological pitfalls. In ICWSM'14: Proceedings of the 8th international AAAI conference on weblogs and social media, forthcoming. http://arxiv.org/ftp/arxiv/papers/1403/1403.7400.pdf. Accessed 26 Jul 2014

Vogt WP (1999) Dictionary of statistics & methodology, 2nd edn. Sage, Thousand Oaks, CA

Webopedia (2014) Big data. http://www.webopedia.com/TERM/B/big_data.html. Accessed 29 Jul 2014