

# Getting Data for (Business) Statistics: What's new? What's next?

Ger Snijkers

Statistics Netherlands / Utrecht University, e-mail: [g.snijkers@cbs.nl](mailto:g.snijkers@cbs.nl)

## Abstract

This paper discusses new developments in (business) data collection methodology and their implications. A model for business data collection serves as a framework.

**Keywords:** business surveys, multi-source/mixed-mode data collection, internet

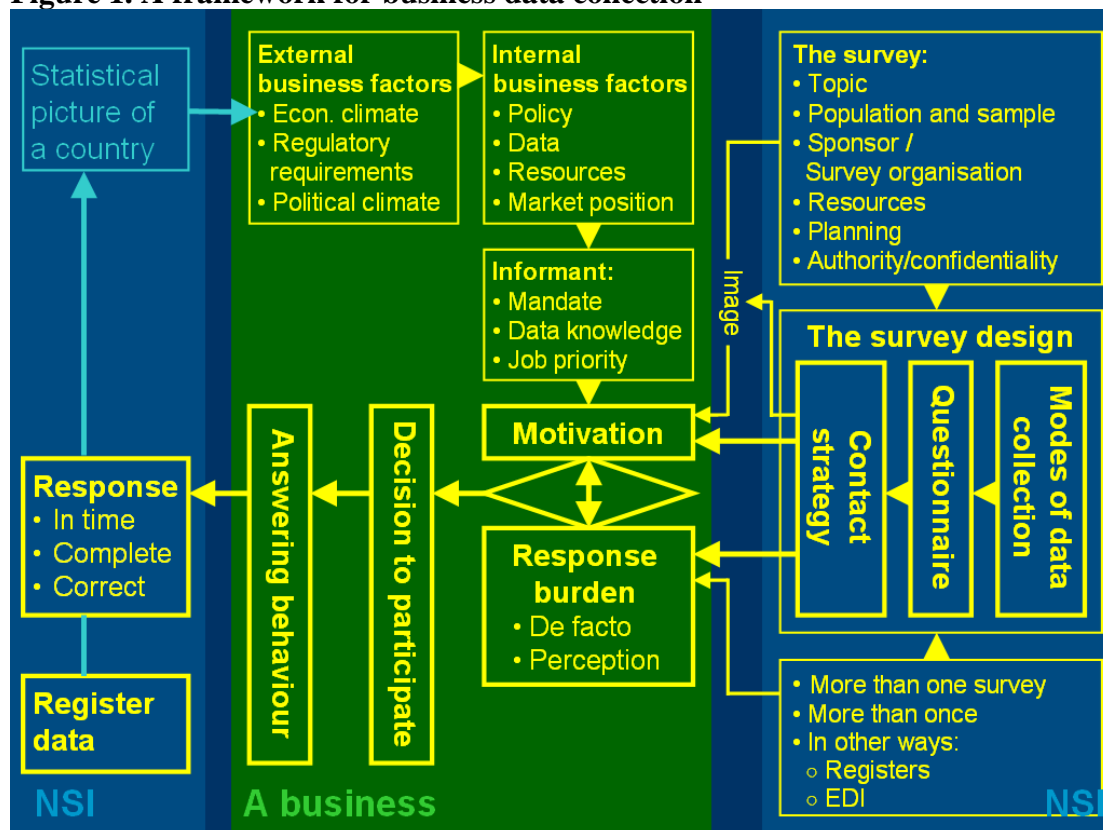
## 1. A framework for business data collection

Following a conceptual framework for survey participation in social surveys (Groves & Couper, 1998), a framework for business survey participation (Willimack, Nichols & Sudman, 2002), and a conceptualization of total business survey burden (Hedlin et al., 2005), and based on the process of getting data for business statistics in NSI's, I developed a new framework, in which actors and aspects of the data collection process are incorporated (Snijkers, 2007). My aim was to get a concise model that describes the data collection process for business surveys in general. The model, however, is not a process model but a causal model with response as the target variable. The model is presented in figure 1, and discussed by Snijkers (2007, 2008).

In a simple situation, one specific survey is conducted with sampled businesses according to a survey design. With the survey design the survey organisation tries to control the data collection process, i.e. tries to influence the behaviour of the business respondent –the response behaviour– in order to get the required response. In general the situation is much more complex: NSI's conduct many surveys, or should I say, collect data on many subjects in many ways, and businesses get many requests to deliver data. They question now is, what can NSI's do to get the data they want as quickly and complete as possible at good quality and at the lowest costs as possible, both for the survey organisation and the business respondents. The model in figure 1 identifies the elements that are relevant with regard to this question. This model also helps in identifying aspects of the data collection process with regard to new developments, and the consequences of these developments.

In this paper I will describe a number of (relatively) new developments that are relevant for business data collection. I feel however that these developments are not only relevant for business data collection but for data collection in general. It is not my aim to cover all developments, but merely discuss a few relevant examples. In the last section I will discuss the implications of these developments. But I will start with a brief discussion of the (not so long ago) past in section 2.

**Figure 1. A framework for business data collection**



## 2. The past: From stove pipes to business data collection methodology

Until some 20 years ago data collection for business statistics was simple. Data for business statistics (as published by NSI's) were collected using paper forms. These forms are characterised by detailed items to be answered by business respondents. Furthermore, a lot of technical terms are used, and explained by many and long definitions. These definitions stem from government or Eurostat regulations, which often are not in line with the information that is available in business administrations. Generally, these surveys are mandatory. Every survey collected data for a specific set of statistics for a specific branch of industry, and was designed independently of other surveys (stove pipes). NSI's were more or less autonomous with regard to the data collection design.

The result was a whole set of surveys and questionnaires sent to businesses without much design coordination. Aspects that to some extent were coordinated were the sample units and definitions of variables (e.g. size class and economic activity). The focus was on the internal production process. What happened within businesses was a black box (in figure 1: the green box –A business– is black).

At the beginning of the 1990 the idea emerged that the “lack of published methods and communication among researchers was a stumbling block for progress in solving business surveys’ unique problems” (Cox et al., 1995, p. xiii). This was the start of the First International Conference on Establishment Surveys that took place in 1993 (ASA, 1993; Cox et al., 1995). By now, we’ve had three ICES conferences: the

Second in 2000 and the last one in 2007 (ASA, 2000, 2007). And the preparations are being taken for ICES IV.

When looking at the papers presented at these conferences a shift in the research issues under focus can be noticed: from stove-pipes to general business survey methodology. At ICES I (ASA, 1993) Most papers addressed issues dealing with surveying specific branches of industry, like agriculture, energy, health care, trade, finance, education, manufacturing industry, construction industry, etc. Also a lot of attention was given to business frames, sampling and registers, classification systems, data analysis and estimation. Only little attention was given to general data collection methodology issues like questionnaire design, data quality, and non-response. The picture that emerges from these issues is a stove-pipe approach with single mode survey designs in which the survey organisation is central.

At ICES III (ASA, 2007) many papers discussed survey data collection methodology issues, including issues like questionnaire design and pre-testing, survey participation (non-response, response burden) and contact strategies, mixed-mode designs and electronic data collection, data quality (bias), and the opening of the black box (understanding the response process in businesses). Again the use of administrative data was discussed. And again the traditional issues like business frames and sampling, weighting, outlier detection, data analysis and estimation were addressed.

In about 15 years we see a shift from stove pipes and single-mode designs to general business data collection methodology, discussing systematic and standardised methods for business data collection in general, moving towards multi-source/mixed-mode designs (Snijkers, 2007, 2008) in which the respondent is put central, which implies tailoring the design to the respondent.

To tailor adequately it is necessary to open the black box. Here, I will mention another development: Cognitive Aspects of Survey Methodology (CASM). This movement started in the 1980's the USA and Germany (Tanur & Fienberg, 1992; Snijkers, 2002), and was the beginning of pre-testing of questionnaires in laboratories, at first for social surveys but later on also for business surveys (see e.g. ICES II: ASA, 2000; Willimack et al., 2004).

The application of cognitive research to information processing in survey responding (Jobe & Mingay, 1991, p. 178), “offered a view of the survey respondent as a question-and-answer system that carried out a series of mental operations, such as comprehension of what was required in response, retrieval of relevant information from memory, and decision-making to arrive at and provide answers to the survey interviewer’s inquiry.” According to Jobe and Mingay (1991, p. 178), “modelling the respondent’s mental operations represented a vast change over the simple stimulus-response conception of respondent behaviour, that from the beginning of modern survey-taking governed the principles employed in designing survey instruments.”

One benefit from the CASM movement thus is the modelling of these mental operations in four stages (see e.g. Tourangeau, Rips & Rasinski, 2000). This model has been extended to business surveys (see e.g. Willimack et al, 2004). Another is the emergence of laboratories for pre-testing and usability testing (web questionnaires).

Since the beginning of CASM and the 1<sup>st</sup> ICES conference, many new developments have taken place, technological developments (like the introduction of the internet in surveys) but also new insights in behavioural and other sciences. In the next section I will discuss a few examples of these developments and insights. I will use the model as shown in figure 1 as framework.

### 3. What's new?

In figure 1, the aspects listed at the right hand side deal with and most of them are in control by the survey organisation. I will discuss most of these aspects step by step, starting with the box at the right upper side. This box lists preconditions of the data collection, like the topic and the resources, as well as the sampling design.

The challenges NSI's face today is that there is a demand for *more and integrated information*. It's not separate statistics for every branch of industry anymore, but there is a demand for statistics that provide an overview of specific themes, like globalisation, global warming, knowledge-based economy. Also these statistics have to be published in a shorter period of time, with less money, and at less compliance costs. The policy of the Dutch government here is that citizens and businesses provide the same data to the government only once, and that these data are forwarded internally to all government agencies that need these data.

The fact that businesses provide their data only once to the government has immediate consequences for the setting up and maintenance of the *sampling frame*. For the statistical business register data from the Tax Office, the Chambers of Commerce, and the Industrial Insurance Boards need to be used. In the Netherlands, a unified integrated and exhaustive business register is coming up soon (Ritzen, 2007).

As for the sampling design an issue that is discussed recently in the Netherlands is the introduction of *survey holidays*, which is the rotating of units over time. A survey holiday system has been implemented for two large, burdensome annual surveys: the Structural Business Survey and the Survey on Investments. When businesses are sampled for one of these surveys in year  $t$ , they will be excluded from the sample in year  $t+1$ . Large and relevant businesses are not included in this system, since their data are necessary in order to get unbiased estimates. Another system has been implemented in the UK, e.g., and discussed in Norway. The introduction of such a system in the field needs to be done carefully. The concept of 'survey holiday' may be interpreted by businesses in various ways. Business respondents may interpret a survey holiday as a survey pension, or they may think that they will not get any surveys at all, for the next year, while in fact they are only excluded for one or two surveys. The aim of this system is reduction of response burden and making sure that some businesses are not selected frequently (Boonstra, Smeets & Smeets, 2006).

Sampling is also being discussed within the area of social surveys in order to reduce non-response. Here, I am referring to *access or online panels*. This is a system in which many units are contacted and asked to take part in a panel. Samples for specific surveys are drawn from this panel. Ideally this panel is constructed using a random sample from a population, but also non-probability selection mechanisms exist. In the Netherlands online panels are very popular, since they are cheap and fast using the

internet. Bethlehem and Stoop (2007) discuss the methodological flaws of online panels, like under-coverage (people without internet are not included in the sample) and self-selection (since selection probabilities are unknown, no unbiased estimates can be computed). In general they conclude that access panels can be used as a host survey under the condition that they are constructed according to strict methodological specifications. I am not aware of the use of these panels in businesses data collection, but if they do not yet exist, they probably will soon.

As for the *mode of data collection*, the introduction of electronic data collection (disk-by-mail, e-mail, and the internet) has changed business data collection dramatically. I will discuss the usage of the internet in two ways: data collection with and without questionnaires. A lot of papers discuss web data collection using questionnaires (see e.g. the proceedings of ICES3: ASA, 2007). A major change is the shift from single-mode (paper) to mixed-mode data collection. Nowadays, most EU NSI's offer both paper and web questionnaires for some business surveys, but still only a few (Norway, Sweden, Finland) offer web as the primary mode for all business surveys. Web data collection proves to be faster and cheaper, and in general improves data quality; a major concern still is mixed-mode effects.

*XBRL* is an Electronic Data Interchange (EDI) technology for primary business data collection without questionnaires. *XBRL* (eXtensible Business Reporting Language) is an internet technology offering businesses the possibility to extract data from their business files by pressing one button, once the right matchings have been made. In the Netherlands *XBRL* has been introduced recently in joint cooperation by the Tax Office, the Chambers of Commerce and Statistics Netherlands (Roos, 2008).

Modern technologies also make other sources of data possible, like the use of *GPS* (Global Positioning Data) data in transportation statistics. With the modern navigation systems in cars and trucks, these data become available (Daas, Roos & Puts, 2008).

Other sources to be used in business data collection are administrative data. The use of *administrative data* in statistics has been discussed in the literature for some time (see e.g. ICES I: ASA, 1993). In a number of countries, e.g. Finland (ASA, 2007: ICES III End Panel Discussion) and Portugal (Chumbeau et al., 2008), registers are leading, but in many countries in practice surveys are still leading and the use of registers is now introduced step-by-step.

I now move to the *questionnaire*. The introduction of electronic data collection offers new possibilities to the questionnaire, as is discussed by Couper (2009). Web questionnaires are self-administered and computerised instruments, making customization (tailoring) and control (e.g. incorporating routing, and edit and consistency checks) possible. Because of the ever growing power of computers the possibilities of web questionnaires are increasing. Instead of using only text, also images and pictures, spoken language, animations and video pictures can be included in the questionnaire. With the addition of the human voice, images of people, and videos the presence of an interviewer can be introduced.

An aspect of the survey design that needs more attention is the communication of the survey (the *contact strategy*). Jones et al. (2008) identify pre-survey communication (introducing the survey and seeking cooperation), survey field period communication

(making contact, seeking cooperation, handling questions by respondents, and respondent chasing), and post survey field period communication (respondent chasing, enforcement actions, and data validation). In every stage a number of modes of communication can be used: the traditional letters, brochures, and telephone contacts, but also new modes like e-mails and the internet. The communication mode does not necessarily have to be the same as the data collection mode. Thus, like the data collection, also the contact strategy shifts from single mode to mixed-mode.

In every contact, it is important to think carefully about the message to be expressed (what). To get the message across, the way it is expressed (how: tone-of-voice, layout, compliance principles), the selection of the right contact within a business (addressee), and the moment of communication (when) are of relevance. An aspect that needs special attention is the introduction of mixed-mode designs.

With the introduction of the internet the communication with respondents also has become more complex. More modes of communication can be used, and the initiative for communicating in the field and post field stages no longer is in control by the NSI: businesses can send e-mails and check the internet. In the present information society, businesses may be influenced by other (indirect survey) communications, like what they see in the news papers, on TV, and on the internet. Indirect outings are in control by NSI's are e.g. outputs published on the internet.

What has made communication also more complex is that we no longer can take survey cooperation by businesses for granted, even in case of mandatory surveys. Snijkers, Berkenbosch and Luppens (2007; see also Snijkers, 2008) have studied the argument used by businesses to comply with survey requests in the Netherlands and related those arguments to the compliance principles as defined by Cialdini (see Groves, Cialdini & Couper, 1992). Most businesses cooperate because the surveys are mandatory. Also businesses would like to know what the data are used for and what their interest is in complying. And businesses are sensitive for getting something in return, e.g. benchmark information. To my knowledge, not many studies have researched the effect of incentives in business surveys (Snijkers, 2008). As for their internal process, they also would like to have an overview of data requests (a survey calendar).

The issue here is how to motivate business respondents to cooperate. I am now opening the black box (see figure 1). Research studying the internal processes in businesses with regard to data collection and surveys is relatively new. As figure 1 shows, this deals with motivation, response burden, the decision to participate and the completion of a questionnaire.

*Motivation* is dependent on a number of aspects as modelled in figure 1. The aspects include external factors, internal factors and characteristics of the informant. These relationships are discussed by Willimack et al. (2002) and Snijkers (2008).

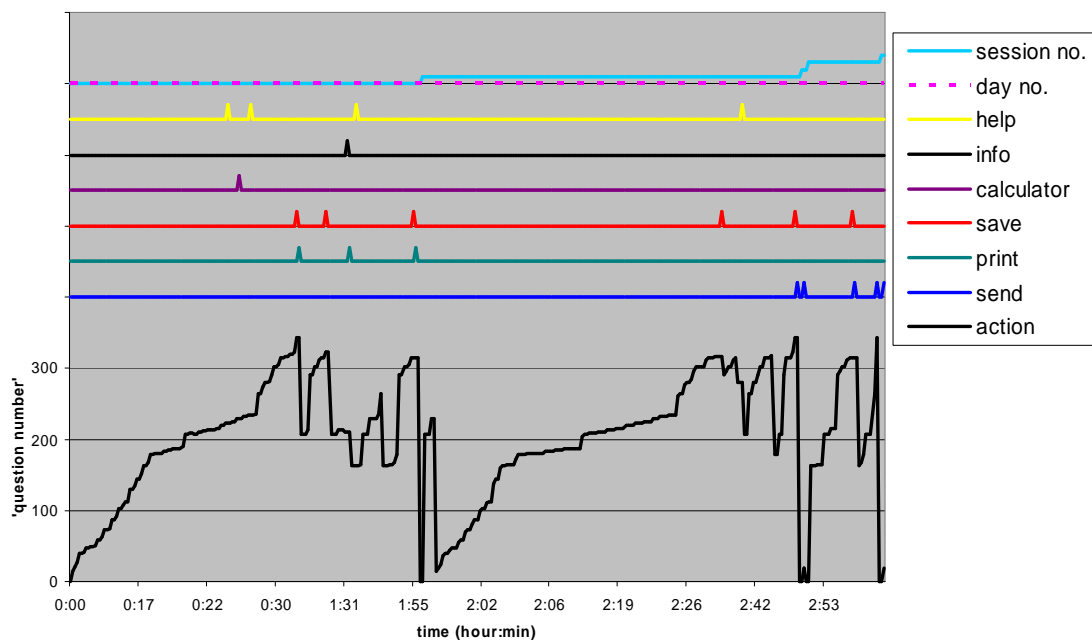
Motivation is an important factor that affects the decision to participate; the other is *response burden*, i.e. the work that has to be done to comply with a data request, and more importantly the perception of work. Response burden is an issue that is typically related to business surveys, and is not new. Statistics Netherlands has been working on the reduction of response burden since 1994. Relatively new, however, is

systematic research into this concept. The NSI's of Sweden, Norway and the UK (Hedlin et al., 2005) have conceptualised this concept and published methods on how to assess perceived response burden. Dale and Haraldsen (2007) have edited a handbook for monitoring and evaluating business response burden.

The next steps in the response model are the *decision to participate* and the *completion of the questionnaire*. To evaluate the total response process, I will discuss *paradata*. The concept of paradata is introduced by Couper in 1998 (Couper, 2009) and refers to data about the process. Couper and Lyberg (2005) discuss the use of paradata in survey research. This can be illustrated by two examples from research at Statistics Netherlands. Response analyses for the monthly Short Term Statistics show that response rates increase immediately after reminding the respondent. Also these analyses show that the effect is faster for web questionnaires (with e-mail reminding) than for paper (with paper reminder letters) (Hoekstra, 2007).

Key stroke or audit trails are paradata about the completion process. Figure 2 shows an example of how a web questionnaire for the annual Structural Business Survey has been completed (Morren & Snijkers, 2008). The graph shows that it took about 3 hours all together to complete the questionnaire. This was done in a number of sessions in one day. The respondent started at the beginning of the questionnaire (question number 1 on the y-axis), and went through the questionnaire in about 30 minutes, then he went back and forth for a number of times. Apart from this conscientious completion profile, we have also discovered a printing profile (starting with printing the questionnaire, stopping, and filling it in a short period of time in another session), and quick 'n' dirty profile (revealing a minimal effort to complete the questionnaire). On average, 266 actions in 2 sessions were needed to complete the questionnaire; 43% of the respondents printed the questionnaire. The average completion time was about 1 hour and 10 minutes.

**Figure 2. Audit trail of a conscientious respondent**



#### 4. What's next: Implications of new technologies?

In the light of integrated statistics, we need content matter specialists, who are experts in these matters. We will see a shift from content matter specialists who actually are single survey managers to content experts, who manage integrated sets of statistics. Their statistics will be estimates based on multiple sources and mixed-mode surveys, using advanced statistical modelling. In case additional information is needed, a survey can be commissioned. This will, however, not be a single-purpose survey but more likely an omnibus survey, in which overlap over questionnaires is reduced to a minimum.

Going over all developments, the picture emerges that it is no longer possible to design a single survey without taking all other data collections into account. Survey holidays and survey calendars imply that one sample frame for all surveys needs to be used and samples need to be coordinated. Also the definitions of units need to be coordinated with other organisations like the Tax Office and the Chambers of Commerce, as well as definitions of important variables (metadata).

Coordination of definitions of units and variables are also important conditions for using administrative data and XBRL. The use of registers also brings about changes in the production process, and systems for managing all available information need to present. NSI's are becoming dependent on providers of these data. In combination with survey data, business data collection will become a *multi-source/mixed-mode design* (Snijkers, 2007; Snijkers, 2008).

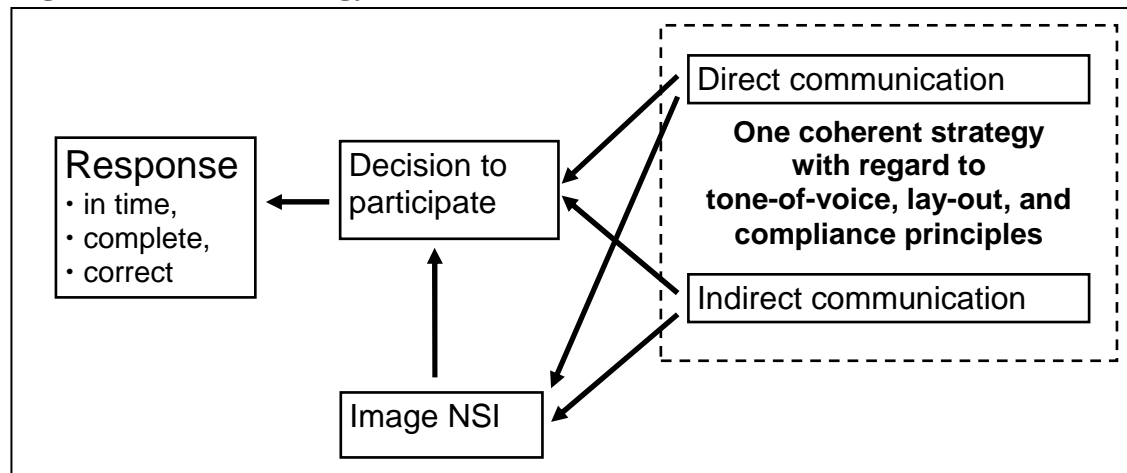
With the introduction of mixed-mode designs (both data collection mode and contact mode), survey complexity increases. For every mode a data collection process has to be developed and controlled (process management), and for every unit the mode of data collection and contact has to be registered (case management). Also questionnaires need to be developed for every mode, which are tailored to the mode but generate the same information. Methodologists must be competent in all modes (Dillman et al., 2009).

Web questionnaires are different from paper questionnaires. Snijkers, Onat and Vis (2007) show that completing a questionnaire on the computer is very different from completing a paper questionnaire. These differences include the facts that a web questionnaire is not just a passive measuring instrument, and respondents expect it to help them. Also reading from the PC-screen is very different than from paper, as is navigating and obtaining an overview. With a computer people are less patient than when reading from paper. Therefore, for web questionnaires it is important to design the questionnaire according to the task-by-task principle, so people can move forward in small steps. A detailed overview of designing effective web surveys is presented by Couper (2009). Research in the Netherlands shows that web designers overestimate the computer skills of users (van Deursen & van Dijk, 2008).

As for communicating the survey, in my view, it is important to seriously take the arguments by businesses into account. Research with regard to persuasion strategies in surveys, like the application of the compliance principles as developed by Cialdini (see e.g. Groves et al. 1992) is needed. I feel that insights from social psychology, communication sciences and market research can serve survey methodology. The goal

is to get a coherent and effective communication strategy for communication modes, as is shown in figure 3.

**Figure 3. Contact strategy communication model**



In order to open the black box and tailor the multiple-source/mixed designs to internal business processes, it is my belief that insights from other sciences like administrative and organisational sciences can help business survey methodology. Thus, the model in figure 1 can be improved. At ICES III, Willimack (2007) has made an effort in this direction. I feel that the next step should be a new CASM movement, bringing together various scientific disciplines to discuss business survey data collection.

## References

- ASA (1993, 2000, 2007) *Proceedings of the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> International Conference on Establishment Surveys*, American Statistical Association.
- Bethlehem, J., & I. Stoop (2007) Online Panels – A Paradigm Theft? In: *The Challenges of a Changing World*, Trotman, M. (ed.), ASC, Berkeley, UK.
- Boonstra, H.J., M. Smeets, & V. Smeets (2006) *Survey holidays for CBS surveys* (in Dutch: Periodieke vrijwaring van CBS-enquêtes). Statistics Netherlands, Heerlen.
- Chumbeau, A., C. Neves, & H. Pereira (2008) Simplified Business Information: Improving Quality by using Administrative Data. In: *Proceedings of the 4<sup>th</sup> European Conference on Quality in Official Statistics*, Rome, Italy.
- Couper, M.P. (2009) *Designing Effective Web Surveys*. Cambridge University Press.
- Couper, M.P., & L. Lyberg (2005) The Use of Paradata in Survey Research. *Proceedings of the 55th Meeting of the ISI* (CD-rom), Sidney, Australia.
- Cox, B.G. et al. (1995) *Business Survey Methods*. Wiley, New York.
- Daas, P., M. Roos, & M. Puts (2008) *New technologies in data collection* (in Dutch: Waarnemingsinnovatie: Nieuwe bronnen en mogelijkheden). Statistics Netherlands, Heerlen.
- Dale, T., & G. Haraldsen (eds.) (2007) *Handbook for Monitoring and Evaluating Business Survey Response Burden*, Luxembourg: Eurostat.
- Dillman, D.A., J.D. Smyth, & L.-M. Christian (2009) *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method, 3rd Edition*. Wiley, Hoboken.
- Groves, R.M., R.B. Cialdini, & M.P. Couper (1992) *Understanding the Decision to Participate in a Survey*. *Public Opinion Quarterly*, Vol. 56, No. 4, 475-495.

- Groves, R.M., & M.P. Couper (1998) *Nonresponse in Household Interview Surveys*. Wiley, New York.
- Hedlin, D., T. Dale, G. Haraldsen, & J. Jones (2005) *Developing Methods for Assessing Perceived Response Burden*. Statistics Sweden, Stockholm, Statistics Norway, Oslo, and UK Office for National Statistics, London.
- Jobe, J.B., & D.J. Mingay (1991) Cognition and Survey Measurement: History and Overview. *Applied Cognitive Psychology*, Vol. 5, No. 3, 175-192.
- Jones, J., A. Lewis, S. Woodland, G. Jones, & J. Byard (2008) Communicating with Survey Respondents at the UK Office for National Statistics. *Paper presented at the Joint Statistical Meeting*, Denver.
- Hoekstra, M. (2007), *Mode-effects in business surveys*. (in Dutch: Analyse van mode-effecten bij bedrijfsenquêtes). Statistics Netherlands, Heerlen.
- Morren, M., & G. Snijkers (2008) Utility of audit trails in the Annual Structural Business Survey. *Paper presented at the 19<sup>th</sup> Non-response Workshop*, Ljubljana, Slovenia.
- Ritzen, J. (2007), Statistical Business Register: Content, Place and Role in Economic Statistics. In: *Proceedings of the 3<sup>rd</sup> International Conference on Establishment Surveys (ICES-III)*, ASA, 179-191.
- Roos, M. (2008) *Using XBRL in a statistical context. The case of the Dutch Taxonomy Project*. Statistics Netherlands, Heerlen.
- Snijkers, G. (2002) *Cognitive Laboratory Experiences: On Pre-testing Computerised Questionnaires and Data Quality*. Ph.D.Thesis, Utrecht University.
- Snijkers, G. (2007) Collecting Data for Business Statistics: A Response Model. *Proceedings of the 56th Meeting of the ISI* (CD-rom), Lisbon, Portugal.
- Snijkers, G. (2008) Getting Data for Business Statistics: A Response Model. In: *Proceedings of the 4<sup>th</sup> European Conference on Quality in Official Statistics*, Rome, Italy.
- Snijkers, G., B. Berkenbosch, & M. Luppens (2007), Understanding the Decision to participate in a Business Survey, *Proceedings of the 3rd International Conference on Establishment Surveys (ICES-III)*, ASA, 1048-1059.
- Snijkers, G., E. Onat, & R. Vis (2007) The Annual Structural Business Survey: Developing and Testing an Electronic Form, In: *Proceedings of the 3<sup>rd</sup> International Conference on Establishment Surveys (ICES-III)*, ASA, 456-463.
- Tanur, J.M., & S.E. Fienberg, 1992, Cognitive Aspects of Surveys: Yesterday, Today and Tomorrow. *Journal of Official Statistics*, Vol. 8, No. 1, pp. 5-17.
- Tourangeau, R., L.J. Rips, & K. Rasinski (2000) *The Psychology of Survey Response*. Cambridge University Press, Cambridge.
- Van Deursen, A., & J. van Dijk (2008), *Computer skills of Dutch citizens* (in Dutch: Digitale vaardigheden van Nederlandse burgers). Twente University, Enschede.
- Willimack, D.K. (2007), Considering the Establishment Survey response Process in the Context of the Administrative Sciences. In: *Proceedings of the 3<sup>rd</sup> International Conference on Establishment Surveys (ICES-III)*, ASA, 892-903.
- Willimack, D.K., E. Nichols, & S. Sudman (2002) Understanding Unit and Item Nonresponse in Business Surveys. In: *Survey Nonresponse*, Groves, R., et al. (eds.), Wiley, New York, 213-227.
- Willimack, D.K., et al. (2004), Evolution and Adaptation of Questionnaire Development, Evaluation, and Testing Methods for Establishment Surveys. In: *Methods for Testing and Evaluating Questionnaires*, Presser, S., et al. (eds), Wiley, Hoboken, 385-407.