1Using the Theory of Socially Distributed Cognition to Study the Establishment Survey Response Process

Boris Lorenc[1]
Statistics Sweden[1]

Abstract

The aim of this paper is to introduce a new approach to understanding and analyzing the response process in establishment surveys, complementary to the existing individual and hybrid approaches. It is a cognitive approach, but with a new unit of analysis: that part of the establishment that takes part in responding to a survey. One of the key characteristics of the approach is its focus on the path that requested information needs to traverse in order to correctly reach the statistics producer's database. This is referred to as the data perspective of the approach, distinguishing it from the others. Some key concepts of the socially distributed cognition approach are introduced in the paper and illustrated, followed by a pair of studies inspired by the approach. Finally, a scheme for evaluation of establishment survey questionnaires with respect to data availability, also inspired by the socially distributed cognition approach, is presented.

Keywords: propagation of representational states, PRS, established practices of survey participation

1. Introduction

Concerns about the quality of data collected in establishment surveys are not new, and neither are the attempts to generalize and give theoretical accounts of the response process that produces these data. The terms in which these accounts are cast are influenced, among other factors, by concurrent developments in related fields of study. Early accounts of the establishment surveys response process, in 1960-1970's, relied in part on theories of organizational behavior, to be, in the 1980-1990's, amended with cognitive theories about the response process (cf. Edwards and Cantor, 1991).

Working on a better understanding and modeling of the response process in establishment surveys, researchers seem to have taken the individual response process as the base upon which they layered the complexities implied by the sampling unit being an establishment (Edwards and Cantor, 1991; Willimack and Nichols, 2001). The results were models of the response process that were quite general in their outreach but at the same time do appear somewhat eclectic. For instance, the hybrid model (Willimack and Nichols, 2001; Willimack, Nichols and Sudman, 2002) seems to combine the accounts of the response process from the viewpoints of a producer of statistics, an establishment and an individual respondent.

In giving any theoretical account, choice of the unit of analysis makes an essential step. This in turn may be guided in part by what the theory that the researcher sees as plausible for the phenomenon under study recommends and in part by what the goal of the analysis is. While the hybrid model, for instance, does not seem to have any unit of theoretical analysis explicitly singled out (as it's goal presumably is presentation of an all-around model of the response process in establishment surveys), implicitly one of the units of analysis appears to be an individual cognitive system, the one of a single respondent who comprehends the request, retrieves the data, judges and communicates the response (steps 4-7 of the model).

The aim of this paper is to introduce a new approach to understanding and analyzing the response process in establishment surveys, complementary to the existing individual and hybrid approaches. It is a cognitive approach in that it has information processing in its focus. It relates in that respect to the individual cognitive 4-step response model (Tourangeau, 1984). But, it takes another unit of cognitive analysis than an individual, namely the establishment. In that, it relates to theories of organizational behavior, mentioned at the outset. The focus is, though, on representations and this mixture of organizational and cognitive aspects is what characterizes the Socially Distributed Cognition (SDC) theory, originally created by Edwin Hutchins (Hutchins, 1995).

Benefits of the approach include existence of an explicit unit of analysis and a developed theoretical frame of reference behind it. In particular, the concept of *propagation of representational states* provides a valuable contribution in understanding and analyzing the establishment response process. Representations are in this approach observable and provide for an objective way of analysis.

The approach is geared towards surveys concerning "hard data on record", especially recurring such surveys, where *practices of survey participation* are established and

propagated through time over different individual respondents and their tools. The approach is, by the same token, of little relevance for surveys of, for instance, attitudes of employees.

The SDC theory, like any other theory, brings some aspects of the investigated phenomenon into focus and disregards other aspects. Motivation for introducing it to establishment surveys methodology lies in the potential of the approach to be a complement to the other approaches (e.g. the hybrid model), providing a new way of looking at things and opening up new topics, thus improving our understanding of the response process in establishment surveys. Illustrations of this are given in the text.

The outline of the rest of the paper is as follows. Ideas and principles of SDC are introduced in general terms in Section 2. The especially important concept of propagation of representational states is in Section 3 illustrated with two examples involving a survey of schools. Sections 4 and 5 present two studies that were motivated by (or, grounded in) the principles of SDC as applied to establishment surveys—as illustrations of the kind of topics that are picked up by this approach and of the proposals that resulted. A round up and a discussion are the subject of the final Section 6.

2. Principles of Socially Distributed Cognition

The unit of cognitive analysis is traditionally an individual. It is above all a single human, but advances in understanding of the mental life and its relation to physical processes widened this view into allowing at least higher apes but in some views other animals and even artificial structures to be considered cognitive. Thus, a general, formalized notion of cognitive processes has been derived in terms of

> *representational structures in the mind and computational procedures that operate on those structures* [Thagard, 2007].

SDC does not depart in this respect from the standard approach. But, SDC theorists (e.g. Hutchins, 1995; Hollan, Hutchins and Kirsh, 2000) argue that identification of a unit for cognitive analysis does not need always to go along the lines of physical boundaries of individuals but that the identification should rather be conducted in terms of the flow of representational structures. For instance, for one of the types of phenomena that these authors studied (namely, memory processes related to flying commercial airplanes), it proved beneficial to view the airline cockpit as the unit of cognitive analysis because the memory process involved "a rich interaction between internal processes, the manipulation of objects, and the traffic in representations among the pilots. A complete theory of individual memory by itself is insufficient to understand how this memory system works" (Hollan et al., 2000:176).

In the SDC approach to the response process in establishment surveys, the unit of cognitive analysis is the part of the enterprise that responds to a survey. This choice is motivated by the complex interactions between employee(s) and the establishment's information system(s) that as a rule take place when providing a valid establishment survey response. Taking into account only individual cognitive processes is not sufficient. Karin Goldenberg seems to echo the same view when she, arguing for going beyond the individual level when studying data quality in establishment surveys, states: "Although the respondent may understand the question perfectly well, the establishment's records may not contain the needed information, the data may not be aggregated in a way that meets survey requirements, or the respondent may not have the knowledge necessary to prepare the answer" (Goldenberg, 1994:1357).

If cognitive processes are to be viewed as distributed, then the SDC theory argues that there are three senses in which that they are (here with reference to the establishment survey response process): (*i.*) distributed between people in an establishment taking part in a survey (e.g. accountants and HR-personnel; administrators and teachers); (*ii.*) distributed between the people and the tools they use (e.g. records, invoices, computerized systems and other record-keeping devices); (*iii.*) distributed through time, in the form of established practices of survey participation (e.g. respondent roles, use of notes and spreadsheets).

The representations of concern for an establishment survey, within the SDC approach, relate to "hard data". The point of departure is the assumption that representations of the hard data of interest exist somewhere in the 'cognitive system' (i.e., the new unit of analysis) of the establishment and there to correctly reflect the relevant state of affairs in the world for that establishment. Main concern of the analysis is then the question of what needs to happen so that the hard datum in the 'cognitive system' of the establishment is correctly mapped onto the statistics producer's database.

To a considerable degree the cognitive processes—when these are understood in the SDC sense, which the scare quotes in the preceding paragraph indicate—take place in the open, observable, social world. Thus, they are in principle amenable for more objective study than what representations in the classical view, involving a single human, are. For instance, retrieval in an establishment survey may take place by having an employee perform a process on any of the three levels identified by Willimack et al. (2002). The usual cognitive processes of encoding,

comprehension, retrieval, judgment and formatting (Tourangeau, 1984; Eisenhower, Mathiowetz and Morganstein, 1991) may be distributed across individuals, and/or information systems as well as in time.

3. Propagation of Representational States

3.1 Theoretical background

A representation is something that stands—for somebody, in some respect—for something else (Peirce, 1931 -58(1897), 2:228). A written letter "A" may stand for the sound \a'\, a picture of Paris may represent Paris. Mental or inner representations, standing for phenomena in the real world, are considered to be the content of humans' internal, cognitive processes. It is traditionally taken that we operate upon mental representations when we perceive, judge, recall, make decisions or perform other cognitive activities.

Although this approach to cognition is not unquestioned (e.g. Núñez and Freeman, 1999, contains some alternatives), the SDC theory does not depart from it completely: cognition is there taken to be "computation realized through the creation, transformation, and propagation of representational states" (Hutchins, 1995:49). Nevertheless, while assuming the existence and importance of internal processes, SDC focuses on external representations, considering them more amenable for scientific, objective investigation.

Thus, propagation of representational states (PRS) is one form of cognitive activity in a cognitive system. A technical note: the more involved term "representational state" is used in SDC and here rather than the simpler "representation" because from the definition of representation it follows that context and observer take part in determining what a representation is.

The concept of PRS provides a tool for analyzing the response process in establishment surveys. I proceed to illustrate PRS for two of the variables collected in a survey. It will be seen that in one case the propagation is successful while in the other it is much less so. Consequences for data quality are then briefly mentioned.

3.2 Illustrations

The two examples come from a survey of schools done each year by Statistics Sweden on behalf of the Swedish National Agency for Education ("the schools survey"). This survey is actually a census. In addition to pure enumeration, breakdown of the pupil population by gender, class and several aspects of pupils' language education enrollment and attendance are requested. Among the language items are the following:

(a) whether the pupil is entitled to instruction in her or his native language,
(b) whether the entitled pupil attends classes in her or his native language, and finally,
(c) whether the attending pupil gets the instruction within the curriculum proper or not.

The response process, as we shall see, is a task that often involves more than one person. Nevertheless, there is one person that has to be, by request of Statistics Sweden, designated "contact person" (for editing purposes, amongst others). No aggregate statistics about that person's position in the school is produced, but impressions from contacts with schools during a qualitative study (Lorenc, 2006) would suggest that it is an administrative clerk (each school has a staff of one or two persons dealing purely with administrative matters), the school's principal, staff member from the school administration section of the school district or municipality, or a teacher. The data collection method in the schools survey was originally a paper self-administered questionnaire (SAQ), which since 2001 has been complemented by a web SAQ.

In a preparatory qualitative study of the response process, I visited 7 schools that took part in the schools survey (Lorenc, 2006). Regarding technical tools, it was notable that all visited schools had some kind of a computerized record of pupils. Five had one of the two dominating programs in the school administration software market, one had one of the smaller programs on that market, and one used a commonly used commercial spreadsheet program.

The minimum of the data that the schools seem to have had in the records and held current were the names of the pupils and their dates of birth (relevant for remuneration).

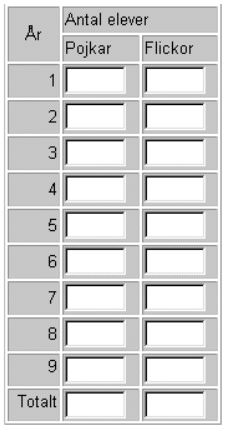

Figure 1: Questionnaire item 1, requiring a breakdown of the school's pupils by class and gender.

None of the schools had in their records all the data asked for in the survey. A data gathering process beyond consultation of the records was implicit for completing the survey.

3.2.1 Simple response process

Figure 1 presents the questionnaire item 1, which requires breakdown of the school's pupils by class and gender. (Depicted is the web form item, but the paper form item is virtually the same.) In illustrating the response process for that item, I will make several simplifying assumptions. I will assume—based on the observation above—that pupils actually enrolled in a school are correctly mapped onto the school's record. (Unnecessarily to say, this may not hold.) Further, I will assume that gender is coded explicitly in the record. (Complications may, however, arise when it is not, requiring manual processing, which has been observed during the visits.) Finally, as majority of the schools use one of the two dominating programs, which in turn can produce the class and gender breakdown automatically, through a simple menu choice, I will for simplicity assume that all the schools have access to such a record and that all respondents are able to produce such a breakdown on screen or paper.

Then, essential steps of PRS from the record onto the survey database are:

I. from the record onto the screen or paper,
II. from the screen or paper onto the respondent,
III. from the respondent onto the form (paper or web),
IV. from the form onto the survey database.

Challenges to correct propagation are here modest, nevertheless they do exist. For instance, one of the two larger record programs displays boys' data in the right column of the pair, while the survey form has the column for boys to the left.

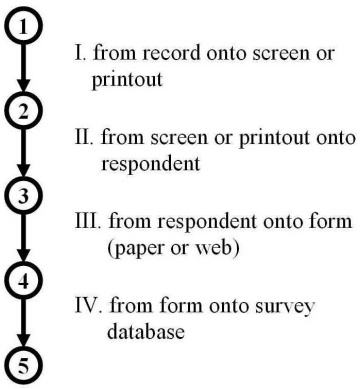

Figure 2: A graph of propagation of representational states (PRS) for questionnaire item 1.

A graph of PRS from a school's record onto the statistics producer's database is given in Figure 2. Here, each node denotes a representational state, while each arc is a propagation step. In this way, PRS allows formalization and graphical presentation of the response process.

3.2.2 Complex response process

Next example concerns the language items (a)-(c). The web questionnaire item that collects this information is depicted in Figure 3: the headings are class, language, and—under the broad heading of number of pupils—four double columns of which the first three relate to the language items (a)-(c). In each pair the left column concerns boys and the right girls. (The fourth double column is unrelated to the other three and is left out here in order to simplify the exposition.)

A school is informed of a pupil's native language (if other than Swedish) through the application form sent in by the parents. It may be noted that this initial propagation fails whenever parents, for whatever reason, do not report a native language or report it erroneously. Information about entitlement to instruction in one's native language is in some schools stored in the record. In any case, in order to submit information about actual participation (second and third double column in the body of the questionnaire table depicted in Figure 3), a distributed process of data collection is often initiated because this information is not recorded or kept up to date.

Key persons for collecting native language data not in the record are native language teachers. Contact persons, who in general also are respondents for the rest of the questionnaire, have here a twofold role: to coordinate data collection and to compile the collected data (cf. Willimack and Nichols, 2001:3).

A teacher of a native language is assumed, in virtue of having been given the task, to know which pupils attend

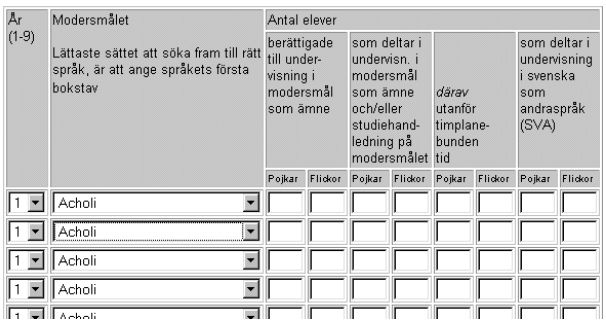

Figure 3: Questionnaire item 2, requiring aspects of native language education per class and language (for details, see text).

classes he or she gives; in he qualitative study, I had no contacts with native language teachers, nor have I otherwise assessed the validity of this assumption. In the course of the data collection process, the teacher may have received from the contact person a paper copy of the relevant table from the questionnaire and been asked to supply the information for the pupils he or she knows about. One often applied strategy for contact persons was to collect this information from native language teachers on the individual pupil level (use the form in Figure 3 to fill in names of the pupils entitled and attending), and then to compile it in the processing step. A benefit of this approach is that it enables checking that a pupil has been included exactly once in the reporting.

Most often, there were as many teachers of native languages in a school as there were native languages taught in it, which may be several or even many. Including teachers of the usual modern foreign languages and of Swedish as the second language, that may have provided information for the other language items, this may imply quite a large group of people being involved. No actual data can, however, be presented here about these numbers for this survey.

Essential PRS steps through the cognitive system and into the statistics producer's database are here as follows (I will assume, for the sake of easier comparison with the previous example, that the propagation into the record was correct, although this may be questionable):

I. from the record onto the data collection coordinator or a printout,
II. from the data collection coordinator or a printout onto the native language teacher providing information on those entitled to participate,
III. from the school's/statistics producer's/established criterion for threshold of attendance that yields the "the pupil attends" judgment to the native language teacher,
IV. from actual attendance of the pupils onto the native language teacher,
V. possibly, from the teacher onto an external record (e.g. an attendance list),
VI. possibly, from the external record onto the teacher,
VII. from the native language teacher onto the paper copy of the questionnaire (supplied by the data collection coordinator),
VIII. from the paper onto the data collection coordinator while compiling the data,
IX. possibly, from the data compiler onto auxiliary notes while processing the data (e.g. tallying), and back,
X. from the data compiler onto the schools survey questionnaire draft or onto questionnaire.

It ought to be noted that existence of the "attendance list" mentioned in step V above is hypothetical (i.e. neither observed in the current study nor mentioned by the respondents) but plausible.

The response process is also depicted as a graph in Figure 4. Its central element is integration of three sources of information that a native language teacher involved in the response process needs to perform if a correct propagation of representation is to be achieved: entitlement, actual attendance and criteria for level of accumulated actual attendance whose attainment warrants the judgment "the pupil attends". While integration (node 5) itself is a mental process in the standard sense, availability of the

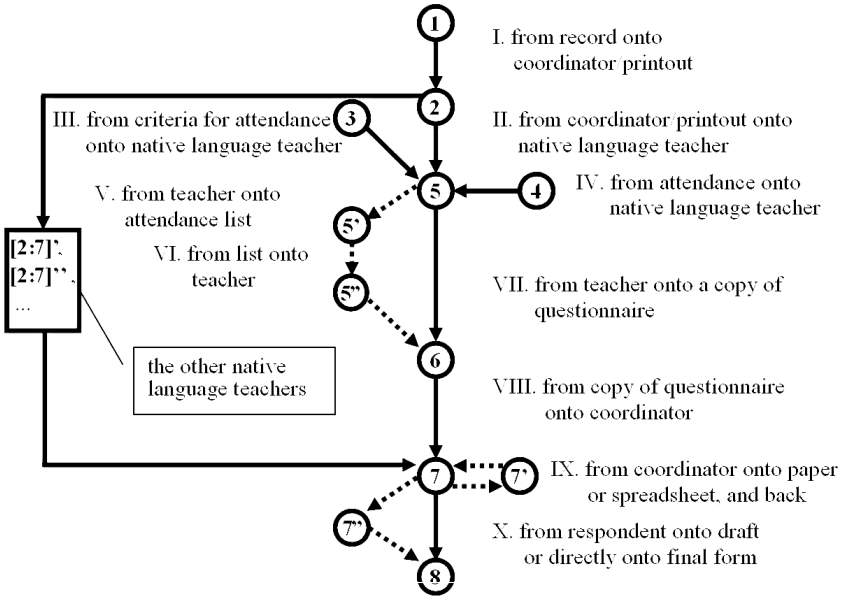

Figure 4: A graph of propagation of representational states (PRS) for questionnaire item 2.

data for it is a result of a social process, thus amenable to the SDC analysis. Contributing to the complexity is the fact that a complete response by a school on this item usually involves several native language teachers. Overall correct propagation would thus entail that in the steps II to VIII correct propagation was achieved with respect to all the native language teachers involved.

3.3 Discussion and further considerations

Comparison of Figures 2 and 4 indicates fairly larger complexity of the response process for item 2 than for item 1. While there are yet no data to corroborate the claim, the larger complexity presumably also leads to lesser quality of the item 2 variable: errors of omission and commission of some pupils, varying understanding of instructions by different native language teachers, etc. A SDC analyst notices also fast that if the criteria for what constitutes "attendance" are not defined, then strong additional variance is introduced in the data, further diminishing data quality for the statistics producer.

Regarding the methodology, simple reflection indicates that granularity of representing PRS as a sequence of nodes and arcs is dependent on the specific context of use. A high-quality model would tend to break down the path in such a way that no more than one error source is active on a single arc. But, such a graph could prove to be too complex, preventing comprehension by its user. Balancing between the extremes, at least an effort to separate important error contributions into different arcs should be made. Building a model of a specific response process might require an iterative procedure that involves collecting data on actual occurrence of errors.

To each arc in a graphical representation of the information propagation may thus a probability of correct propagation be attached. In addition to empirical data, in the development phase of a survey expert estimates may be used. This enables even some statistical modeling. For the sake of demonstration, let us in the graph in Figure 2 attach the following probabilities to the example at hand: p_1 (probability of occurrence of error in arc 1, i.e. the arc connecting the nodes 1 and 2, for a specific school) 1.00, p_2 0.98, p_3 0.95 and p_4 1.00. Then, with independence of error occurrence in each arc from that in the others, we have the probability of correct propagation of this datum into the statistics producer's database as $p_1 \times p_2 \times p_3 \times p_4 = 0.93$.

4. Study I: Reporting Averages

In addition to providing the possibility for graphical illustration and analysis, SDC helps open up some research topics and bring out some aspects of the response process that in some other approaches may remain less visible. The starting point in any case is the question: what path(s) the information has to pass in order to end up correctly registered in the statistics producer's database.

A methodological study was initiated in cooperation with a governmental agency that conducted a survey the goal of which was to estimate the extent and effects of occurrence of late payments that small and middle-sized enterprises (SME's) in Sweden are exposed to by their organizational customers.

Several questions in the survey were of the form:

- What percent of your invoices <have been paid after due-in date> in the previous 12 months,
- On average, how many <days after due-in date have the late payments arrived> in the previous 12 months,

where expressions in the angle brackets gave the subject content of the item.

Questions were by expert reviewers deemed easy to understand, data to generally exist in records and to be accessible to respondents and retrievable. For the cases where the accounting system was not able to provide the answer with a simple mouse click, an analysis suggests that PRS would in this case need to consist of the respondent performing following steps: pass through the invoices issued in the reference period and either (for the proportion questions) categorize each into 'paid in time' or 'paid late', and then count the number of occurrences of the latter kind, or (for questions regarding amounts) successively add the sought-for quantity related to each invoice, to finally in both cases divide the resulting sum by the number of issued invoices.

As there was a concern that some proportion of respondents would not go through this process, a methodological follow-up study was carried out. It aimed at measuring some of the variables in the main study using another technique. A sample of 300 respondents in the main survey was re-contacted and asked to provide exact information regarding a small sample of 10 of their invoices issued in the reference period: date of issue, due-in date, date of arrival of the payment, and—in the case of a delayed payment of an invoice—whether the business has taken or not specified debt recovery actions regarding that payment.

The assumption behind the study was that if information in the enterprise accounting system does not pass through the above mentioned steps, this may manifest itself in a substantial discrepancy between the results of the main study and the 'golden standard' of the follow-up study.

The results obtained in the two studies are presented in Table 1. Comparing the two series, an agreement within

Table 1: Results for variables that were the same in the main and follow-up surveys

	Item	*Main Study*	*Follow-up*
1.	Average credit period (days)	27	27
2.	Percent of invoices with credit period longer than 30 days	14	18
3.	Percent of invoices with credit period of at least 60 days	3	4
4.	Percent of invoices paid late	14	50
5.	Percent of invoices paid more than 10 days late	4	13
6.	Percent of invoices paid more than 30 days late	1	3
7.	Average length of delay of late payments (days)	10	9

sampling uncertainty can be observed except for one variable: percent of late paid invoices. In the main study it was estimated to be 14% and in the follow-up 50%.

The main hypothesis for the difference—waiting to be substantiated or rejected in a qualitative study subsequently launched within Cognitive Laboratory of Statistics Sweden—is a "gray zone" hypothesis: that respondents allow a certain period of time to pass after the expiration of the credit period before they determine a payment to be really delayed. The hypothesis has two variants, a "psychological" one claiming that the respondents perceive all delays but classify only payments late at least *t* days as delayed, and a "system" one whereby accounting tools that the respondents use have settings to alarm the user of a late payment after a chosen delay of *t* days.

From the data of the follow-up study, it was possible to estimate the delay *t* above to be about 9 days. By modeling the length of delay (number of days between dates of expiration of credit period and placement of debt on the enterprise's account) as an exponential distribution, which is not unreasonable from the subject matter point of view, and by the memoryless property of the exponential distribution, the average length of delay is the same whether all the delays are taken into account or only those longer than 9 days—the obtained results are in agreement with this model.

In summary, data of the follow-up study do not provide evidence that the PRS failed. One explanation could be that information of the kind asked for in the main survey is of vital importance for a SME and thus in general has already been appropriately encoded in the respondents mind at the time that the request to participate in the main survey came (i.e., the PRS steps assumed necessary are continuously performed and the results updated). With respect to the only variable on which there was a substantial difference, the mentioned hypothesis implies that "late payment" is not a hard fact but rather a decision made after taking specific circumstances into account. Details and further results concerning this study are expected to be given in a forthcoming report (Lorenc, Björnram, Persson and Wibell, 2007). One general conclusion thus far is that this somewhat puzzling result seems to have as much bearing on understanding individual cognitive processes as on understanding establishment reporting.

The study on reporting averages was given in order to illustrate one kind of questions that SDC brings forth. In general, this question concerns the path that data need to pass in order to be correctly recorded in the statistics producer's database. The question is the same in spirit to the one that was posed when the 4-step cognitive model was introduced (Tourangeau, 1984), the difference being that here external representations are in focus as well as the processing that these representations go through in the context of an establishment survey responding.

5. Study II: A Model for Data Availability

Next example is intended to illustrate a SDC approach to a theoretical question. When going through the example, it should be born in mind that there can be other approaches to the investigated problem and that the obtained model is motivated specifically by the SDC approach.

It has been suggested, in the context of surveys of individuals, that unavailability of data might lead to nonresponse and possibly other quality problems in a survey (Beatty and Herrmann, 2002). While it is plausible that this relation basically holds even in establishment surveys, no formal (theoretical) approach to the problem of availability of data in establishment surveys—similar to the cited authors' approach in the field of individual surveys—could be found in the literature. Thus, a methodological study was initiated to identify levels of data availability, with the intended practical aim of using these levels for conducting fast quality evaluation instruments for questionnaire items in recurring establishment surveys at Statistics Sweden.

In the current analysis, "unavailable" will mean "not able to propagate correctly to the statistics producer's database". This is in the spirit of SDC, where the focus is, as already pointed out, on the path that data need to pass

in order to be correctly recorded in the statistics producer's database. It will be taken, throughout, that the mode of data collection is a paper or web-based SAQ.

Given the aim of the envisioned instrument, a descriptive stance (what happens in the course of supplying the response for a particular item in a survey) was taken rather than a normative stance (what would need to happen if a correct response for a particular item in a survey is to be recorded).

In the literature on establishment surveys, data retrieval and communication of data (i.e. questionnaire fill-in) are two distinct phases of the response process (e.g. Hedlin, Dale, Haraldsen and Jones, 2005; Willimack and Nichols, 2001; Willimack et al., 2002). It has also been documented that data retrieval consists of not only data collection itself but also of data compilation (Willimack and Nichols, 2001). This evidence, put in a SDC perspective, resulted in a proposal of a framework in which two main characteristics of the retrieval process are: media for propagation and operations on representations performed by the respondent.

Each process has a corresponding scale, with a number of levels (Tables 2 and 3). In the course of identifying the levels, two aspects were taken into account: substantially different (and increasing) levels of failure in propagation and ease of identification of a particular level by the user. Task of an user (evaluator) is to pick out the level that best describes the response process for a particular item in an establishment or, somewhat abstractly, a group of establishments (cf. section 5.3, below).

5.1 Media for Propagation

Let us begin by looking at the Media for propagation scale. Level 1 (Electronic) indicates the ability of the administrative computer tool to provide the desired report without involvement of any respondent (or, as referred to at Statistics Sweden, contact person), except to the extent that that respondent would match the request to the available command and execute the command. (It is presumed that this procedure fulfills the request for the whole survey and not for only a single item.)

The second level (Respondent) indicates involvement of a human respondent in PRS for that item. Typically, no lookup is needed: information would be already propagated to (i.e. represented in) the mind of the respondent by the time the data request came, and at that time it would be propagated further onto the form. Request regarding number of employees can in small companies be one such item.

Table 2: Media for propagation scale

Level	*Content*
1.	Electronic
2.	Respondent
3.	Electronic + Respondent
4.	Respondent′ + Respondent
5.	Electronic + Respondent′ + Respondent
6.	Other way of data collection + Respondent
7.	Other way of data collection + Respondent′ + Respondent
8.	None

The third level (Electronic + Respondent) indicates involvement of a human respondent that uses a computerized administrative tool to provide the requested information. Example 1 of Section 3.2 represents such an involvement of a respondent.

The fourth level (Respondent′ + Respondent) indicates that the respondent is, for the specific item, contacting another respondent, who in turn provides the requested information according to the process of level 2. Typical example is that the respondent does not know the requested datum but knows who knows it and gets it from that person.

The fifth level (Electronic + Respondent′ + Respondent) indicates that the respondent is, for the specific item, contacting another respondent, who in turn provides the requested information according to the process of level 3. Typically, the respondent cannot access the relevant electronic administrative tool or does not know how to use it, and so instead contacts another respondent who provides the required information from the system.

The sixth level (Other way of data collection + Respondent) indicates that the respondent is no longer searching for required information in an electronic administrative tool, but collecting the data elsewhere: by observation (in e.g. the warehouse), by interviewing other employees, and so on.

The seventh level (Other way of data collection + Respondent′ + Respondent) indicates that the respondent is contacting other respondent(s) who in turn collect the data elsewhere than in an electronic administrative tool: by observation, by interviewing, and so on. Example 2 of Section 3.2 represents such an involvement of respondents.

Table 3: Scale for operations on representations, performed by the respondent

Level	*Content*
1.	None (solely electronic processing)
2.	Re-write
3.	Arithmetic operations
4.	Estimation (on at least one component, if there is more than one)
5.	Guessing (on at least one component, if there is more than one)
6.	Satisficing

The eight level (None) indicates that no PRS is really taking place. Propagation has nothing to start from or has been broken somewhere along the path.

5.2 Operations on Representations

On the scale for operations on representations performed by the respondent, level 1 (None (solely electronic processing)) indicates no contact between the data and the respondent, and thus that respondent performs no operation on the data. For instance, data are collected from the accounting system on a file (using a predefined command) and then the file is put onto the statistics producer's server. While there is in general no fix correspondence between the levels of the operations on representations scale and the media scale, operations level 1 would often be found in conjunction with media level 1 (Electronic).

Level 2 (Re-write) indicates that the administrative system provides information of adequate granularity and thus that providing this information to statistics producer requires no further operations than to reproduce it on another medium. Example 1 of Section 3.2 exemplifies this.

Level 3 (Arithmetic operations) connotes that at least two representations have propagated to the respondent, and indicates that these in themselves do not suffice to fulfill the request, but instead need to be brought into relation. Addition over accounts or time periods would be a typical example, exclusion of some posts from an existing cumulative account another. It ought to be noted that whenever at least one of the components in an arithmetic operation is not a fact that correctly propagated to the respondent, but rather an estimate or a guess, than the whole operation is moved onto the corresponding higher level (4 or 5).

Level 4 (Estimation) indicates that the intended representation (if it existed) did not reach the respondent, and is replaced by another one of relatively high quality. For instance, instead of supplying the requested information on the value of delivered goods in a reference period (of which there is, at least in Sweden, no good formal account of in many SME's), the companies may supply information on invoiced goods (which is well recorded in accounting systems of most companies). It ought to be noted that this level, as well as the next one, implies that PRS has broken and is replaced by an approximation.

Level 5 (Guessing), like the preceding level, indicates that the intended representation did not reach the respondent, and is replaced by another. On this level the quality of the substitute is questionable, hidden in mental operations that the respondent—if asked—would likely have difficulties in accounting for. There is no implication that the supplied datum must be inaccurate, just that an account of the PRS that leads to the substitute is difficult to provide.

Level 6 (Satisficing) indicates that the response was provided on the bases of situational clues rather than on the basis of some substantial data that propagated to the respondent.

5.3 Envisioned Usage

Assuming that data availability in establishments has a consequence for data quality, it is in the statistics producer's interest to gauge this availability in the target population. The two presented scales are conceived to be a part of an evaluation instrument for assessing data availability in enterprises in the context of recurring surveys.

Two uses are envisioned: field use, involving real respondents in conjunction with site visits, and a proxy use. The proxies are the statistics producer's employees responsible for data collection in a specific survey. This latter category of employees is in general rather well informed about the practicalities of a specific data collection and provision in a particular survey. Among other sources of this knowledge are the contacts initiated by the respondents in cases of difficulties arising when taking part in the survey or during data editing re-contacts. Main motivation for using proxies lies in expected reduction of cost and duration of the evaluations, on the expense of the loss of inaccuracy due to lack of closer insight and the impact of generalization.

5.4 Discussion

The two aspects—media for propagation and operations on representations—do not cover all the aspects of the establishment surveys response process, but instead pick up those closely related to SDC. Presence of mental processes (notably, the judgment step) is only implicit, through the success or failure of some PRS or through the impact on the level of operations on representations: substantial treatment of mental processes is left to the classical approach. This is an expected consequence of the use of SDC.

An aspect of the enterprise surveys response process close to SDC that has thus far been left out from the formal treatment is the role of a respondent (or "contact person", in Statistics Sweden terminology) as a coordinator of data collection efforts. (An example is given in section 3.2.2, where coordination and compilation is for a cluster of items carried out by the school's administrator and where the respondents—or "respondents′", by the denotation of section 5.1—are native language teachers). Whether this omission has a consequence for validity of the model depends on whether coordination has a consequence for PRS and data quality beyond that that different respondents are involved.

The division into levels is best seen as not accomplished yet, as it could benefit by further refinements: for instance, in operations level 3, a distinction between doing an addition manually and using an electronic adding machine might have a sufficiently large consequence for the response process and data quality to warrant creating two levels of the scale.

The two scales have thus far undergone limited validity and reliability checks, so further studies remain to be carried out in that direction.

Neither of the two scales has a direct correspondence with the cognitive state scale of Beatty and Herrmann (2002), but there are resemblances. Data *available* in an enterprise survey can be said to be those that can be produced directly (the shortest path of PRS is across levels 1 or 2 of the media for propagation scale) and not operated much upon by the respondent (levels 1 or 2 of the operations on representations scale). *Accessible* data can be said to be those that can be retrieved after some data collection activity (shortest path across any of the levels 3 to 5 of the media scale) and possibly after some arithmetic processing (level 3 of the operations scale). For an item that requires involvement of others to perform direct data collection (shortest path across level 6 or 7 of the media scale), like Example 2 of Section 3.2, it appears less clear whether to view it as *accessible* or *generatable* (pursue of agreement between different approaches would tip the scales towards the latter choice).

Nevertheless, the gist of both Beatty and Herrmann's scale and the current ones is the same: the further down on a scale a questionnaire item is, the likelier it is that data quality issues will be associated with the item (low accuracy, nonresponse). For the current scales, though, this assumption needs to be corroborated by empirical evidence.

6. Summary

A theory is a way of looking at things. Any formal theory brings some aspects of the investigated phenomenon into focus and disregards other aspects. The SDC approach to understanding the response process in establishment surveys focuses on propagation of representational states that occurs in the course of this process. It brings to the fore the question "What needs to happen so that a certain state of affairs in the world would correctly be reflected in the statistics producer's database?" It can be said that the approach takes a *data perspective* on the process.

SDC is a coherent theory, developed in another practical field but directly applicable to establishment responding. In the context of establishment surveys, it should best be viewed as a complement to the existing approaches. It contributes by providing accounts of the phenomena not treated thus far, giving new ways to describe the response process and opening new research topics. Examples of this include graphical analysis of the response process (with the possibility of quantification of propagation success), research issues opened by studying PRS, and attempts to provide more detailed accounts of the response process, as presented in the text.

Not treated in this paper, but otherwise encountered in the survey literature, is the phenomenon of established practices of survey participation (Bavdaz, 2006; Willimack, Nichols and Sudman, 1999). While current establishment survey literature does not treat this in any systematic manner, SDC has a concept and a place for this phenomenon as well in its theoretical apparatus.

In endeavoring to answer that question, the surveyor would achieve a better understanding of respondents' (i.e. establishments') environments: whether and where in the establishments the requested information exists, whether there is a way in which the information can be mapped onto the database, what the particular path that the information needs to traverse is, what risks of error are associated with specific steps of the path and so on. On the one hand, by specifying the methods of data collection, designing forms and formulating instructions, the surveyor may exert influence on the choice of paths

that will be used for propagation of information, and thus promote those that are less prone to error. On the other hand, in finding answers to the questions raised, the surveyor would hopefully get to know better the environments of the establishments and the circumstances in which people in them (primary and secondary respondents) are coping in their everyday activities.

The dream of every cognitive scientist, I believe, is to look inside the head and see how the mind works, how it processes information. (Advanced attempts at this are in fact currently carried out.) SDC is amazing in the sense that is provides the opportunity to inexpensively and objectively study how a cognitive unit—the establishment—processes information.

Acknowledgments

The work on this study was in part supported by The Bank of Sweden Tercentenary Foundation grant K2000-5063 to Department of Statistics, Stockholm University. Study I was co-sponsored by Statistics Sweden and Ministry of Enterprise, Energy and Communications Sweden. The views expressed in the paper are those of the author and need not reflect the views of any of the above mentioned organizations. I am grateful to Lilli Japec, who proposed the line of research that led to the model presented in Study II, as well as to the colleagues that gave feedback and otherwise contributed to the two studies presented here: Anette Björnram, Anna-Greta Erikson, Johan Erikson, Eva Furubjelke, Lisa Levinsson, Reijo Ljungberg and Andreas Persson.

References

Bavdaz, M. (2006). "The Response Process in Recurring Business Surveys". *Proceedings of European Conference on Quality in Survey Statistics*, Cardiff, 24 - 26 April 2006.

Beatty, P., and Herrmann, D. (2002), "To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse". In Groves et al. (2002), pp. 71-87.

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

Edwards, W.S. and D. Cantor (1991). "Towards a Response Model in Establishment Surveys". In Biemer et al. (1991), pp. 211-33.

Eisenhower, D., N.A. Mathiowetz and D. Morganstein (1991). "Recall Error: Sources and Bias Reduction Techniques". In Biemer et al. (1991), pp. 127-44.

Goldenberg K.L. (1994). "Answering Questions, Questioning Answers: Evaluating Data Quality in an Establishment Survey". In *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 1357-62.

Groves, R.M., D.A. Dillman, J.L. Eltinge and R.J.A. Little (Eds.) (2002). *Survey Nonresponse.* New York, NY:Wiley.

Hedlin, D., T. Dale, G. Haraldsen and J. Jones (2005). *Developing Methods for Assessing Perceived Response Burden*. LEG-report to Eurostat.

Hollan, J.D., E. Hutchins and D. Kirsh (2000). "Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research". *ACM Transactions on Computer-Human Interaction*, 7:2, pp.174-96.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Lorenc, B. (2006). *Social Distribution of the Response Process in a Survey of Schools*. Research Report 2006:6. Stockholm: Department of Statistics, Stockholm University.

Lorenc, B., A. Björnram, A. Persson and K. Wibell (2007). *Asking About Averages: Approaches and an Assessment*. Research report. Örebro: Statistics Sweden.

Núñez, R. and W.J. Freeman (Eds.) (1999). *Reclaiming Cognition: The Primacy of Action, Intention and Emotion*. Thorverton, UK: Imprint Academic.

Peirce, C.S. (1931-58). *Collected Papers of Charles Sanders Peirce*. 8 vols. C. Hartshorne, P. Weiss, and A. Burks (Eds.). Cambridge, MA: Harvard University Press.

Thagard, P. (2007) "Cognitive Science". In *The Stanford Encyclopedia of Philosophy* (Summer 2007 Edition), E.N. Zalta, (Ed.), URL = <http://plato.stanford.edu/archives/sum2007/entries/cognitive-science/>.

Tourangeau, R. (1984). "Cognitive Sciences and Survey Methods". In T. Jabine, M. Straf, J. Tanur and R. Tourangeau (Eds.) (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, DC:National Academy Press, pp. 73-100.

Willimack, D.K. and E. Nichols (2001). "Building an Alternative Response Process Model for Business Surveys". In *Proceedings of the Annual Meeting of the American Statistical Association* [CD-ROM]. Alexandria, VA: American Statistical Association.

Willimack, D.K., E. Nichols and S. Sudman (1999). "Understanding the Questionnaire in Business Surveys". In *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp.889-94.

Willimack, D.K., E. Nichols and S. Sudman (2002). "Understanding Unit and Item Nonresponse in Business Surveys". In Groves et al. (2002), pp. 213-28.