Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

Paradata from Concept to Completion

Fritz Scheuren[1]

What follows is an annotated version of the PowerPoint presentation of my talk given at the fall 2005 Methodology Conference. The slides, even after annotation, are incomplete in several respects as a record of the experience. Notably missing is the Q and A that followed. Some of this was clarifying, for example David Binder's comments. Clarifying corrections were, of course, taken. The many other good points made could not be dealt with in this way. Suffice it to say this presentation needs to be supplemented by the experiences of those now deeply involved in the process, several of whom spoke later in the Conference or at the companion US Federal Committee on Statistical Methodology Conference (www.FCSM.gov) in Washington. No software recommendation was made in the talk but several other presenters had used the ***Nesstar*** *system. This may be a tool for those just starting out and interested in employing the ideas here for their own surveys. A fuller reference to* ***Nesstar*** *is included at the end.*

The word "paradata" may still be unfamiliar to some of you, although the underlying ideas should be well known to all. Mick Couper coined the word a few years ago (e.g., Lyberg and Couper 2005).

Paradata are data about the survey process. These can be simple summaries like --

Refusal rates,
Noncontact rates,
Item nonresponse rates,

or even the fraction of the data that had to be obtained via a proxy interview.

Paradata can be employed at the macro data or summary level, as above, or at the individual or micro data level. Some micro data items that could be provided on every data record in the operating file might be --

Language of the interview,
Who was present at the interview,
Calls attempts until contact was established,
Whether the interview was initially refused?

The length of a face-to-face interview is now typically captured. In a CAPI or CATI setting, even the length of time it can take to answer each question may be obtained and available, at least on an internal file of the survey.

Paradata, in short, are part of the computerized metadata that surround many large-scale government and other surveys -- for instance the monthly Canadian Labour Force Survey.

Why talk about paradata now, especially when it is already so ubiquitous? Well, perhaps, for that very reason. Certainly, while widely available and extensive, paradata are seldom being used at their full power, either during survey operations or, more commonly, at the survey inference stage. For some specific reasons for reminding everyone of this observation, see figure 1.

[1] Fritz Scheuren, NORC University of Chicago

Figure 1
Why Talk About Paradata Now?

- Increase in computing
- Impact on data access
- Shift to greater client focus
- Growing need to quantify all survey impacts on inference
- Growing abandonment of sampling-nonsampling split (Scheuren 2005b)

The changes caused in part by cheaper and faster computing have made the survey going paradigm a much more shared art. Clients can do a lot more of the work that used to be the exclusive province of data producers.

Paradata are used or could be useful at virtually every survey going stage. Figure 2 gives a few of these.

Figure 2
Paradata Uses at Different Survey Stages

- Managing the survey
- Planning for improvements
- In quality profiles
- As part of metadata
- Increasingly in inference

In what follows we take up each of these ideas in turn. The presentation is basically historical with examples drawn mainly from my own practice.

Figure 3
Managing and Planning A Survey

- Traditionally within department
- Mainly functional planning
- Increasingly now done overall
- Quality costs quantified more
- Time and cost tradeoffs shared

In figure 3 above there is a list of how the work on paradata evolved from something done just within a department or survey function, to an activity that is increasingly done in a unified way.

Part of the reason for this is that surveys have often become more expensive due to the challenges of falling response rates and rising labour costs. That has forced data producers to analyze their methods more and more critically.

One of the most effective new tools that has grown up in the last several decades is what can be characterized as the "Quality Profile." In the terminology of this talk, a quality profile is a report that tells the story of a survey from start to finish, mainly using macro paradata summaries.

Figure 4 itemizes the reasons quality profiles are shared with clients and other data producers, both across surveys and for the same survey across time. There have been quite a few such profiles beginning with what was then called an "Error Profile" of the Current Population Survey (CPS). This was done as part of the work of the US Federal Committee on Statistical Methodology (www.FCSM.gov). The principal authors were Barbara Bailar and Camilla Brooks.

Figure 4
Quality Profiles Shared

- Began as an "Error Profile"
- Summarized survey goals
- Used aggregate paradata
- Better producer accountability
- To clients and other producers
- Especially over time (meta-analysis)

Since I was a part of that effort, I can say with some certainty, although memory can play tricks, that we had not seen the role of a quality profile, as a tool in conducting meta-analyses. We did see its value in achieving better producer accountability and in supporting the growth of the client's role in survey analysis.

We still lack a fully worked out attempt at conducting survey meta-analyses that focus on process. Partial efforts abound, however. The systematic building of a combined computer file of all of a survey's paradata was also not thought of back then. Yet we can do both of these now. And we should!

We are not, in my view, learning fast enough from our practice. The growth in computing and in mathematical statistical ideas has far outstripped our employment of these new tools. Maybe that was inevitable, but a better and more systematic use of paradata can help us here; and I advocate it strongly.

Still, we have made much progress, as most of you know. Offering in figure 5 a chronology of the creation of computerized meta-data may allow us to see this.

Figure 5
Incorporation Into Metadata
A Chronology

- Only record layout produced
- Codebooks provided
- Question frequency tables
- Codes (e.g., missing items)
- Quality profile aggregates
- Systems thinking breakthrough

When I began my statistical career, data producers were generally only keeping electronic copies of file, then called tape layouts. Moving these layouts from paper onto the data file themselves was thought of as an advance (and it was). Lots of other functional documentation was created, of course, but was generally not systematically being brought together. You had to be almost inside a producer culture to really appreciate its depth (Scheuren 2005a).

In the 1960s and 1970s computerized codebooks and accompanying (univariate) question-by-question frequency counts were still new. The creation of public use files (e.g., Mulrow and Scheuren 2000) aided in this change but we still, as a profession, have not typically brought all of the survey steps together into a single database.
Deming has enjoined us to employ "system's thinking." Even so, we have not acted on his advice. This is so despite the fact that for a long time now we have been able to.

One of his 14 quality points or principles (Deming 1986) is to break down barriers between departments. Yet we seldom, in large surveys, build a single combined database for a survey. Information on the frame is not always brought forward, for example. Interviewer details, even just (de-identified) codes that links interviewers to their work, are not passed on and placed on the final file, etc.

Why is this our practice? Well, I do not really know but let me speculate and offer three reasons:

Clients do not ask for much of what we could provide

Costs are always a factor and creating such a file is of unproven value

Neither data producers nor clients have fully conceptualized inference uses

There are some notable exceptions, such as codes for imputed data but, generally, we have what may be characterized as a "Catch 22."

Figure 6
Potential Paradata Uses During Inference

- Descriptively for better qualitative interpretation
- As in quality profiles
- Cognitive revolution (Groves)
- Recognition of bias and variance impact
- Quantitative use at inference, even by clients

We seldom create combined files and hence have neither learned how to do so efficiently nor explored their full value at inference. Figure 6 above lists a few of the possibilities were we to change our practice.

What have clients gotten historically in the way of data? Typically, as figure 7 displays, before about 1960, mainly survey aggregates.

Figure 7
Typical Inferences Before About 1960

$$\int f(Y,M,R,C,D)dMdRdCdD = g_1(Y)$$

Classically, this function $g_1(Y)$ was just a vector or matrix of totals, say, in tabular form.

The data producer adjusted for or "integrated out" of the data (Y) or "fixed" all the response errors (R) that they could, missing data problems (M) were also "taken care of," as were coverage issues (C). Using the sample design variables (D), survey weights were obtained and applied in developing tables of aggregates $g_1(Y)$ that were then published.

Now clearly the implication of this classical approach was that the data producers know best and that clients could not add much value. From almost the beginning this was too limiting. For example, tools like regression were required for some problems and the iterative nature of such an application made it necessary for data producers to create some form of microdata product.

Figure 8 represents that stage of development -- a major advance. In fact, the microdata product $g_2(Y,D)$ has spawned a rich literature on how to analyze complex surveys; a great step forward for our profession.

Figure 8
Typical Inferences Before About 1960

$$\int f(Y,M,R,C,D)dMdRdC = g_2(Y,D)$$

The function $g_2(Y,D)$ is a much richer microdata product that begins shifting the inference burden.

Of course, this formulation of the inference problem leaves a great deal of the work using paradata still in the hands of the data producer, as is shown in figure 9 on the next page.

Notice that if, as is usually the case, for the microdata to be provided publicly they have to be in a de-identified form. This means that full design detail (D) may have to be sacrificed, with the need to make a compromise on the direct estimation of sample variances.

Figure 9
Initial Separation from Inference

- Implicit assumption that data editing and other measurement errors are resolved by producer
- Missing data, after adjustment, is ignorable
- Only survey weights are needed for unbiasedness
- Variances can be calculated by using only "D" Design variables
- Often "D" reduced to de-identify cases

Paradata did not figure much, if at all, at this early public microdata stage. But increasingly clients wanted to obtain data about what had been done to impute for missing item nonresponse.

Here we begin to see a start in the use of paradata. Putting "flags" on the microdata for item missingness can be, as I have characterized it in figure 10, considered a revolution. Rubin's seminal work (1987) on multiple imputation was one, but only one of the uses of this.

Figure 10
Typical Inferences After
Missingness Revolution

$$\int f(Y,M,R,C,D)dRdC = g_3(Y,M,D)$$

This product $g_3(Y,M,D)$ connects us to the introduction of missingness into survey inference.

As figure 10 illustrates, at most there could be but two more steps in shifting or sharing the inference task as between the data producer and the client. One of these, shown in figure 11 below, moves the response concerns (R) that exist, or part of them anyway, over to the left hand side of the equation. See Scheuren (forthcoming). Bob Groves also has been among those emphasizing the need for this change.

Coverage issues and unit nonresponse have so far remained with data producers but there is no reason why these cannot be shared with clients too.

We are probably not ready to take this last step just yet, nor need we do so now. There is enough to concentrate on elsewhere. But still we need to begin to better document our surveys so that they are amenable to later meta-analyses.

Figure 11
Emerging Inferences with Full
Use of Paradata

$$\int f(Y, M, R, C, D)dC = g_4(Y, M, R, D)$$

$g_4(Y, M, D, R)$ reflects the beginning of the full use of paradata for inference, as advocated in this paper.

Three early examples, speaking from my own practice, may be worth mentioning. They span a number of years and have many common elements that I will discuss at the end.

First were the 1966-67 Survey of Economic Opportunity Files (completed in 1971). These were done just as I was finishing up my graduate education. They are now found in the US National Archives. They are that old. Next were the 1973 CPS-IRS-SSA Exact Match Files. While basically finished in 1980, they have been updated periodically and are still in use at the US Social Security Administration. Finally, there are the 1997 and 1999 files of the National Survey of America's Families (finished in 2000), available at the Urban Institute.

Some of the common attributes of these files are to be found in figure 12 below. The 1973 CPS-IRS-SSA files, unlike the others, also had administrative data added to the survey results.

Figure 12
Some Common Attributes

- All had records for both interviews and noninterviews
- All had a code to link but not identify interviews done by the same interviewer
- All had coding to identify natural groups known from frame, like ultimate cluster (but not identifiable PSUs)
- All had extensive coding for missing data and a code to flag imputations

Still there were elements that were desired but not done. Surveys of interviewers were completed but the results never were added to the microdata survey files. This has controversial elements but may have been done elsewhere. I would love to learn about such examples.

Only modest use of the paradata variables was made at inference, almost none by clients. The theory surrounding the use of these variables was left to others but is emerging now (e.g., Beaumont 2005)

In ending, let me hazard some next steps for us as a profession:

> Operations. I see lots more that could be done during survey operations if we simply move the survey file along from one stage of the process to the next, enlarging it as we go, breaking down thereby in part barriers between departments.
>
> Inference. Deming's dictum of "system's thinking" clearly must extend to clients as well. Many US federal statistical agencies offer workshops for their users. Supporting client use of paradata could go a long way to speeding up our growth as a profession and increase the quality of our practice.
>
> Learning. I look for a lot more use of paradata by people here today. In fact I know many of you are busy right now on this. Of course, statisticians besides those in Statistics Canada can have this fun too. I would love to play a role but, in any case, all my very best to you.

AFTERWORD

In the materials handed out at the Conference I also provided a quality spoof (available separately in the September 2005 AMSTAT NEWS) that played off the word "paradata." I ask Mick Couper's indulgence here but the spoof perhaps made some useful points that are covered in this more usual presentation.

REFERENCES

Biemer, P. and L. Lyberg (2003), *Introduction to Survey Quality*. New York: Wiley. For a fully operational example of top survey quality, see the work of Arthur Kennickell and his colleagues in the Survey of Consumer Finances, found at http://www.frb.gov/ .

Beaumont, Jean-François (2005), "On the Use of Data Collection Information for the Treatment of Unit Nonresponse through Weight Adjustment", *Survey Methodology,* 31, 227-231

Couper, M. and L. Lyberg (2005), "The Use of Paradata in Survey Research", International Statistical Institute Meetings, April 2005, Sydney.

Deming, W. E. (1986), *Out of the Crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Dippo, C. and Sundgren (2000), "The role of metadata in statistics". See also Colledge, M. and Boyko, E. (2000), "Collection and classification of statistical metadata; the real world of implementation". Both of these papers were presented at the Second International Conference on Establishment Surveys in Buffalo.

Ishikawa, K. (1990), *Introduction to Quality Control*. Tokyo: 3A-Corporation.

Jabine, T. (1994), *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS).* National Center for Education Statistics.

Juran, J. M. (1988), *Juran on planning for quality*. New York: Free Press.

Kalton, G., Winglee, M., Krawchuk, S., and Levine, D. (2000), *Quality Profile for SASS Rounds 1-3: 1987-1995, Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS).* National Center for Education Statistics.

Mulrow, J. and Scheuren, F. (2000), A Confidentiality Fable, *JSM Proceedings, 2000*.

Nesstar Publisher, a metadata tool, can be used to publish data and the accompanying documentation to a catalogue on a Nesstar Server. From here these resources can be made available to the wider community via Nesstar WebView.

Rubin, D. (1987), *Multiple Imputation*, Wiley: New York.

Scheuren, F. (2005a), "Reminisce on Multiple Imputation", *The American Statistician.*

Scheuren, F. (2005b), "Seven Rules of Thumb for Nonsampling Error in Surveys," National Institute of Statistical Science (NISS) Total Survey Error Conference, Washington, March 2005.

Scheuren, F., "Macro and micro paradata for quality assessment in surveys", report not yet published, *Journal of Official Statistics.*