Macro and Micro Paradata for Survey Assessment

Fritz Scheuren, Urban Institute
Fritz Scheuren, 1402 Ruffner Rd. Alex., VA 22302
scheuren@aol.com

Key Words: Quality, Client Communication Systems, Statistical Service Systems, and Metadata.

Abstract: At present we typically assess quality by relying heavily on summary process measures or macro paradata that is a by-product of sample selection and survey administration (e.g., item and unit nonresponse rates, response variances, sampling and coverage errors). These measures are an outgrowth of the randomization-based approach to survey sampling that triumphed in government agencies around the world in the two decades after the seminal 1934 paper of Neyman. When most of our present measures were developed, therefore, nonsampling error problems, like nonresponse, were less serious or less well understood.

Now there is a widespread belief that the randomization paradigm needs to be replaced by a more explicitly model-based approach that continues to incorporate features like random selection. In this paper we will illustrate how the current macro paradata measures might be revised or reinterpreted in the light of the concerns just mentioned. The changing partnership between data producers and data analyzers also increases the need for a greatly expanded use of micro paradata – i.e., process details known on each case (such as how many attempts it took to get an interview, whether the interview was in English or Spanish, etc.) To illustrative the paper will employ novel macro paradata summaries of coverage and unit nonresponse, plus micro paradata measures of response variation. The examples will all be taken from the National Survey of America's Families (NSAF) but we believe the perspective we advocate would be actionable in other settings too.

1. Introduction and Background

1.1 Introduction. Paradata can be macro or micro. Macro paradata is very common and widely used. Most of the familiar examples of macro paradata measures are in the form of global process summaries, like overall response or coverage rates. The widespread dissatisfaction with existing quality measures seems to lie in this area. The present paper stresses that while producer-based marginal summaries can and should be improved, they are simply not adequate alone, when the development of information is increasingly a joint undertaking of the data producers and their clients.

Micro paradata is less familiar than macro paradata. A notable exception is that good survey systems keep track of and flag, on individual records, data items that have been imputed. Seldom are other micro paradata items provided to survey clients. Confidentiality concerns can be legitimately cited for some of this practice. The main reasons may well be lack of end researcher interest and inability by

producers in seeing any clear value added. The extra expense involved and simple inertia are among the other reasons why clients seldom obtain or use micro paradata. Instead, at present we assess data quality by macro paradata measures that are an outgrowth of the randomization-based approach to survey sampling that triumphed in government agencies around the world in the two decades after the seminal 1934 paper of Neyman (Neyman 1934).

When most of our present measures were developed, therefore, nonsampling error problems, like nonresponse, were less serious or less well understood. Arguably, the traditional summary measures we still use often, implicitly at least, assume a quasi-randomization model for nonsampling errors (Oh and Scheuren 1983).

Now there is a widespread belief (e.g., Särndal and Swensson, 1992) that the randomization paradigm needs to be replaced by a more explicitly model-based approach that continues to incorporates features like random selection but appeals directly (rather than just implicitly) to models when making inferences. The need to fully incorporate the new quality improvement approaches (Deming 1986) into our measures is also essential. Without a doubt, a holistic approach is needed, similar in focus to the goal of a quality profile report (e.g., Jabine 1994, NCES 2000) -- particularly as part of systematic quality planning (Juran 1988).

Finally, and of increasing importance, the client perspective needs to be recognized, both organizationally (e.g., Petska and Scheuren 1992, Brackstone 1999)) and in the development of metadata and paradata systems (e.g., Dippo and Sundgren 2000). We are producing a service and not a product (e.g., Fellegi 1999); thus, simply providing producer measures of product quality will not be enough, no matter how good they are. In assessing the quality of a service, there are three elements to look at, which we have expressed in equation form as –

$$\text{Quality} = \text{Producer} + \text{Client} + \text{Joint}.$$

In most of the survey literature on quality the focus has been on the first term, when it may well be the third (interaction) term that is key (Batcher and Scheuren 1996).

1.2 Barriers to Improved Assessment. What are the biggest barriers to improving quality assessments in surveys? Reciting a few of the perceived barriers, in addition to those touched on above, may however be of value to help frame what comes next:

(1) Need for change is not widely felt among producers. Probably the biggest barrier is that the need to make a change is not felt by data producers. No compelling case has been made to rethink a process that seems to have been highly success historically. Producers, even if they wanted to change, have not been given a workable alternative. Clients often do not even realize that they are not fully utilizing survey results. They may ignore nonsampling errors altogether or treat them just verbally, without any quantification or sensitivity analyses.

(2) Producers do not usually make it possible for clients to do their own assessments of nonsampling error. Because of confidentiality concerns, some survey public use files, e.g., CPS and SIPP, do not even make it possible to directly estimate sampling error. Existing, so-called, quality measures are often very simple by-products of survey operations and do not use either modern computational power or, indeed to any serious degree, the power of modern statistics.

(3) Two-way communication and real listening systems are entirely too rare. Producers need to better learn the language of their customers and especially to build better listening mechanisms, like customer surveys, that get needed feedbacks. Right now, much of the producer-client communication that goes on is systematically carried out in only one direction. Failure to build adequate listening systems may be among the most serious barriers to producer improvement. The idea that the creation of information is done jointly with clients has not penetrated deeply enough into producer operating systems.

(4) Many survey organizations do an extremely good job of providing detailed information about their processes and running training classes for users. All kinds of interesting general tidbits can be had, except what they might mean to the interpretation of a particular client result. The need for a "just-in-time" knowledge, like that possible with web-based metadata systems, can be a partial answer but this movement is still at an early stage (e.g., Dippo and Sundgren 2000).

(5) The slowness of large survey systems to change is discouraging. Budget pressures and the need to focus on, in some settings at least, growing nonsampling problems, especially unit nonresponse, absorb us. Whatever the reason, there is not enough creative energy being expended on finding new and better ways to measure quality. The gap is especially noticeable in aiding end-users in the development of the quality measures that "fit their use" -- even though we (and they) cannot fully predict what that use will be ahead of time. Survey products and documentation have had a one-size fits all character that needs to be re-examined.

(6) Quality is a value and goal for many and not a tool. Quality, clearly, is an overused word that has lost much of its motivational punch. Ironically, the hype around the word quality may be a major barrier to improving the many sound existing survey assessment tools. After all, who would tell you they are not trying to do a quality job?

(7) Fixing individual assessment measures is a trap, unless the overall assessment system is fixed too. In an information age, producers alone cannot judge fitness for use; hence, the system of assessment must be open to clients. Simply reciting a long list of places where failures of different sorts occurred can create a misleadingly bad impression of quality. Conversely, as noted already, substituting words for quantification can mislead in the opposite direction,. Clients need to know when the survey system may be said to have "failed safe." Some errors are corrected and have no bearing on final usefulness. For example, adjustments are typically made for nonresponse and coverage, but generally there are no measures of the improvement these adjustments may make.

To summarize, the need for change is not widely felt among producers, perhaps because two-way

communication with clients and real listening systems are entirely too rare. Efforts to build just-in-time metadata systems have begun to help, but these are still not mature yet. The slowness of large survey systems to change is discouraging but it has begun and will accelerate as the information age continues. Unfortunately, the word “quality” has largely lost its value and in the remainder of this paper, it will be avoided to the extent possible. Finally, the listing of a litany of defects is often readily available in good surveys but how meaningful is it?

What clients really want to know is how to deal with defects remaining after all survey processing and this is often the one thing that our current systems cannot readily help them learn. A major effort is made in most government surveys to eliminate inconsistencies, impute missing items, adjust for nonresponse and coverage errors. None of these efforts get “scored” in the usual quality measures. For example, the weighted response rate is generally employed as a summary measure of how good a survey is, without regard to the post-survey adjustment efforts made.

What about alternatives? Put another way, there is no use complaining unless something better exists. The rest of this paper illustrates alternatives that are actionable in some settings. A consistent approach will be taken, where explicit models, sometimes computationally challenging, will be used for the nonsampling error being measured. The use of explicit models may, in and of itself, help clients, who often operate in a model dependent world. And with producer help, clients might be able to better integrate producer knowledge, as imbedded in models, with subject matter knowledge, also expressed in models.

One admission might be made before going into the illustrations provided.. The paper is quite incomplete and covers nothing like a full set of alternative paradata measures to the traditional quality indicators now in wide use. This may discourage readers expecting more. To them, an apology is in order. However, other readers may welcome the fact that few fully worked alternatives exist. That is where the challenge is -- to make for our time and technology as complete a survey-going paradigm as Morris Hansen and others did in the 1940s and 1950s.

1.3 What Comes Next? Ideally, the paper should be divided into four parts to deal in turn with errors arising from coverage, nonresponse, measurement and sampling. Only for the first three of these do we attempt to present something new.

The coverage illustration being used in Section 2 examines a new way to estimate residential telephone coverage by using survival analysis. The approach taken is computationally intense and makes full use of modern statistical ideas, borrowing heavily from the work of Lawless (e.g., 1982). It illustrates how we can adapt ideas invented elsewhere to pressing quality measurement problems and thereby improve existing macro paradata measures.

The nonresponse section (Section 3) uses a model to motivate a new macro paradata response rate; it may not be readily accepted. In fact, we are not fully ready yet to accept it ourselves. What the presentation does, however, is represent a fresh approach and, if people could only become "unstuck"

from the conventional implicit quasi-randomization models (Oh and Scheuren 1983) still in use, we might as a community be led to better ideas. We again borrow our approach – this time from the older work of Sekar and Deming (1949) and employ a novel capture/recapture model of nonresponse to attempt to distinguish ignorable from nonignorable nonresponse.

The section on measurement or micro paradata (Section 4) could be the most immediately actionable in a wide variety of settings, since it directly employs existing quantities and can be used in virtually all surveys right now. This is unlike the coverage modeling (of Section 2) which is specialized to an RDD environment or the capture/recapture models of Section 3 that require more development and potentially additional data collection. As will be seen, what we attempt in Section 4 could be fairly readily duplicated in many good surveys, without a great deal of new work. In major government surveys the kinds of micro paradata items needed are certainly going to be available.

There ought to be a section on sampling error but no concrete examples came to mind that would be all that new. This area, in any case, is much further along than the others. So, instead, the paper ends with a few tentative concluding comments (Section 5).

1.4 Background on NSAF. In this paper all the examples are taken from the National Survey of America's Families (NSAF). Because of this fact, the remainder of the present section (Section 1) will be spent on background details about that survey.

In 1996, in response to welfare reform legislation that devolved many activities to the states, the Urban Institute, along with Child Trends, initiated an effort entitled *Assessing the New Federalism*. NSAF is a major component of that effort. NSAF data collection is being conducted for the Urban Institute and Child Trends by Westat. There have been two rounds, for 1997 and 1999, with a third planned for 2002. Each round is obtaining detailed information on the economic, health, and social characteristics of children, adults under the age of 65, and their families. In both 1997 and 1999, interviews were conducted with over 40,000 households, yielding information on over 100,000 people.

By design, large representative samples of households are being taken in each of 13 targeted states (Alabama, California, Colorado, Florida, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, New York, Texas, Washington, and Wisconsin) and in the balance of the nation – oversampling households under 200% of the federal poverty line.

The NSAF is a dual-frame survey with two separate components. One is a random-digit dialing (RDD) survey of households with telephones. The RDD approach was adopted because it is a cost-effective way to collect the desired data. However, because households without telephones contain a significant proportion of low-income children, a supplementary area sample was conducted in person for those households without telephones. In the area sample, households within sampled blocks were screened and all nontelephone households with someone under 65 years of age were interviewed.

All interviews were conducted on the telephone by interviewers working in central interviewing facilities,

using computer-assisted telephone interviewing (CATI) technology. In-person interviewers provided cellular telephones to respondents in nontelephone households to connect those respondents to the interviewing centers for the CATI interview. Nontelephone household interviews were conducted in essentially the same way as those for RDD households. About 10 percent of each telephone interviewer's work was silently monitored for quality control purposes.

For a more complete summary description of NSAF, see the paper by Scheuren and Wang (1999). Full details on the survey are to be found in the NSAF Methodology Series found at the URL http://newfederalism.urban.org/nsaf/methodology.html.

2. Refining Coverage Measurement in RDD Surveys, like NSAF

In random digit dial (RDD) telephone surveys, some telephone numbers ring when dialed (at least that is how it sounds to the person dialing the number), even though the telephone number is not assigned for use. For these numbers it is not possible to classify the residential status for the number with certainty. We denote telephone numbers for which residential status is still not resolved at the end of the data collection period as "undetermined" telephone numbers. If we stopped at this point, we would be unable to separate differences in coverage from noncontact nonresponse.

The percentage of undetermined telephone numbers encountered in RDD surveys has been increasing over the last few years as a result of changes in the telephony system. Piekarski *et al*. (1999) describes many of the changes related to this increase, including the competition for local exchange service in some markets. They note that the number of telephone households increased only 11 percent from 1988 to 1998, but the number of telephone numbers that could be dialed in an RDD telephone survey increased by 80 percent. Even accounting for the increase in the number of households with more than one telephone number and the increased demand for business telephone numbers, many of these newly created numbers are not assigned to any user. Piekarski *et al*. (1999) also suggest that since the changes are at least partially due to competition in local markets, the percentage of telephone numbers that are undetermined may vary substantially by geography.

To describe this situation analytically, we begin by explicitly defining the residency rate for all telephone numbers and then discuss the new approach to estimating this rate. For ease of description, we refer to the t-th call attempt as "trial t." The trials do not refer to fixed points in time, since one case might be receiving its 18th call attempt while another case is receiving its 2nd.

Let

r_t =number of cases (telephone numbers) resolved as residential at trial t; and

n =total number of telephone numbers.

The residency rate at trial t is R_t, and this rate is estimated by

$$\hat{R}_t = \frac{\sum_{k=1}^{t} r_k}{n},$$

where k denotes a trial (i.e., call attempt number) at which there are non-censored cases. The status of a telephone number, as residential or nonresidential, is determined on the t-th call attempt if it can be unambiguously classified on this attempt. Subsequent call attempts to obtain cooperation from the household do not affect R_t, so these additional call attempts are not included in this analysis.

A typical pattern for resolving RDD samples as residential shows that a large proportion of cases is usually resolved within the first few call attempts, and then the curve flattens out. Note that $\{\hat{R}_t\}_{t=1,2,3,\ldots}$ is a nondecreasing sequence that converges to the asymptote R_∞, the overall residency rate. If the residential status of all cases was resolved by some trial T, then $\hat{R}_T$ could be used as an estimate of the overall residency rate. However, in practice, it is neither feasible nor cost-effective to resolve the residential status of all cases. Even after a large number of calls, some cases will remain undetermined, with a status of "no answer" or "no answer, answering machine." Some of these cases are nonworking numbers (numbers that have not been assigned); others include telephone numbers permanently connected to home computers, etc.

The estimated residency rate among cases with undetermined numbers is the difference between the estimated number of residential numbers (an unbiased estimate of R_∞ multiplied by the number of telephone numbers) and the resolved number of residential telephone numbers, divided by the number of undetermined numbers. For sake of illustration, suppose $\hat{R}_\infty = 0.50$, where 900 telephone numbers are resolved as residential, 800 are resolved as nonresidential and 300 are undetermined. Then we would estimate that 100 [=(.50*2,000)-900] of those 300 cases are residential, and that the residency rate for undetermined numbers is 33 percent [100/300].

All the variables needed to apply this approach are known, except the estimate of R_∞. A scheme for estimating R_∞ is to consider cases with undetermined numbers at the end of data collection as right-censored data, with a varying number of call attempts. When cast in this light, techniques from survival analysis can be used. In particular, we first employ a Kaplan-Meier estimator to obtain the overall survival function from the data, where the survival function is the probability that a telephone number is not resolved as either residential or nonresidential by a specific trial. We then partition the survival function into a separate function that describes the probability of a number being classified as residential. This function, evaluated at an infinite number of call attempts, is an estimate of R_∞.

The Kaplan-Meier estimator (also known as product-limit estimator) is a nonparametric procedure to estimate the survival function, $S(t) = \Pr\{T \geq t\}$, where T is a nonnegative random variable that denotes

the "lifetime" of the case. In our application the lifetime is the number of call attempts until the number is classified as residential or nonresidential. The Kaplan-Meier estimate (Lawless, 1982) is

$$\hat{S}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i}$$

where

i indexes the trial or call attempt at which there are non-censored "deaths". (In this context, "deaths" are cases resolved to be residential or non-residential.);

n_i is the number of cases "at risk" just prior to trial t_i. (In this context, being "at risk" just prior to trial t_i corresponds to still being called at the t_i^{th} call attempt.); and

d_i is the number of "deaths" or resolved cases at trial t_i.

If all numbers were called at least one time, S(1) would be equal to 1. However, in our applications, there are some numbers that are never dialed (e.g., listed business numbers), so we have specified the more general condition that S(0) = 1. Now an assumption of survival methods is that all of the telephone numbers eventually "die" (are classified as residential or nonresidential after an infinite number of calls), so S(0)=1. This assumption does not hold in dialing an RDD sample because some telephone numbers are never resolved irrespective of the number of calls. For example, some telephone numbers are not assigned for use but ring when dialed. These numbers are never resolved, no matter how many call attempts are made.

A second issue with applying survival methods to this problem is that the number of call attempts is a less than perfect measure of "time." For example, two call attempts made minutes apart to the same telephone number are not the equivalent of two calls made on different days at different times. The two calls made minutes apart are less likely to resolve a telephone number, but the survival function does not account for this. A third issue concerning the validity of the survival method is that the censoring of undetermined cases must be random. For example, it must not be the case that telephone numbers exhibiting the smallest likelihood of being residential are censored earlier than those exhibiting a greater likelihood of being residential. However, if auxiliary data (such as whether the telephone number is listed) are used to determine when censoring should occur, these auxiliary data could be included in the estimation of conditional survival functions, making the survival method valid.

The determination that a telephone number is residential and the determination that a telephone number is nonresidential may be thought of as the two "causes of death." The survival function given above does not estimate R_∞ because it is the survival function for the resolution of cases due to any reason. The survival functions for the two causes of death are estimated (Lawless, 1982) by

$$\hat{S}_{RES}(t) = \sum_{i:t_i \geq t} \frac{d_{RES,i}}{n_i} \hat{S}(t_i)$$

and

$$\hat{S}_{NONRES}(t) = \sum_{i:t_i \geq t} \frac{d_{NONRES,i}}{n_i} \hat{S}(t_i)$$

where

$d_{RES,i}$ is the number of cases determined to be residential at trial t_i; and

$d_{NONRES,i}$ is the number of cases determined to be nonresidential at trial t_i,

where the summations are defined only at those trials t_i where $n_i > 0$.

The overall residency rate, R_∞, is then estimated as

$$\hat{R}_\infty = \frac{\hat{S}_{RES}(0)}{\hat{S}_{RES}(0) + \hat{S}_{NONRES}(0)}.$$

With this estimate of R_∞, we can apply the approach described at the beginning of the section to estimate the residency rate for cases with undetermined telephone numbers. The estimate is

$$\hat{R}_{UN} = \frac{\left(\hat{R}_\infty \cdot n_{TOT} - n_{RES}\right)}{n_{UN}},$$

where n_{TOT} is the number of total number of cases, n_{RES} is the number resolved as residential, and n_{UN} is the number undetermined. For more on this approach, including extensions and proofs, see Brick, Montiquilla and Scheuren (2000), from which this excerpt was taken. For the 1999 NSAF this method of measuring residential status was highly successful and led to a much cleaner measure of telephone residential coverage and consequently nonresponse rates, discussed next.

3. Refining the NSAF Nonresponse Rate as a Quality Measure

3.1 Types of Unit Nonresponse. Response rates are thought to be important indicators of how good a survey is. This role is justified because of the relationship between the nonresponse rate and nonresponse bias. In simple random sampling we have, for example, the relationship –

$$\bar{y}_r = \bar{y}_n + \frac{m}{n}(\bar{y}_r - \bar{y}_m),$$

where the—

$\bar{y}_r$ reflects average values taken from respondents,

$\bar{y}_n$ the ideal average values on the entire sample (usually not fully available),

m is the number of nonrespondent cases,

n the total sample size, and

$\bar{y}_m$ reflects the values on the nonrespondent cases.

In general, as nonresponse rate increases, the potential for nonresponse bias also increases. This is true of all sampling settings, although the specific expression for bias is different in different designs.

In discussing nonresponse, it is useful to distinguish between two types – ignorable and nonignorable. To define these terms, for our purposes here, suppose there are two groups of eligible potential respondents: those who will respond with a positive probability p>0 and those who will never respond under the essential survey conditions, i.e., those for whom p=0.

Ignorable nonresponse can occur when a survey subject failed to respond at random – i.e., p>0 with the value of p unrelated to the vector of variables "$\underline{y}$," that is of interest. It can also occur when p=0, i.e., for hard core nonrespondents, but this time the average values of the vector $\underline{y}$ in the hard core group must be equal to the corresponding averages for respondents. Nonignorable nonresponse occurs in all other situations.

For units that have a positive probability p>0, the proper choice of post-survey weighting class adjustments for nonresponse can make a great deal of the nonresponse ignorable by conditioning on factors which make the nonresponse rate equal in any given cell or nearly so,. Choosing the adjustment cells entirely in this way is generally not possible but there can still be some gains. The problem is to determine how much of a gain there is and to "score" it. After all, only the potentially nonignorable nonresponse has implications for survey bias. The ignorable nonresponse, given the adjustment, can have a variance impact, if target sample sizes were not achieved, but it has no bias impact. Theoretically, the adjustment cells could also be chosen to make the mean vectors "$\underline{\bar{y}}$" of respondent and hard core nonrespondents equal. This, however, seems a great deal more challenging (Oh and Scheuren 1982).

3.2 Using the Capture/Recapture Model to Score Nonresponse Adjustments. Consider the following model and its usefulness as a way to score the improvement in bias reduction that was made by the post-survey nonresponse adjustment.

Suppose we have made two attempts to interview a subsample of the main survey sample, in our case NSAF. Suppose, unlike in the usual refusal conversion process, we attempted to interview both refusals and previously interviewed cases. We could then construct 2X2 tables, within each nonresponse adjustment weighting class. Each of these 2X2 tables would have weighted cell entries given by the values a, b, c, and d – where the "a" cases had been interviewed twice, the entries "b" and "c" once each, and the entry "d", for those not interviewed at all. Under the assumptions of the capture/recapture model (e.g., Sekar-Deming 1949) -- assumptions equivalent to ignorability when p>0 – we can estimate the capturable or ignorable portion of the d cell, denoted d_I as $d_I = bc/a$. The remainder d minus d_I is then potentially nonignorable.

Now, of course, this approach might not work in small adjustment classes. In NSAF, these could be as small as 30 cases, after collapsing. So we expected to fit a hierarchical log-linear model to the data to gain the strength of deep cross-classification, while preserving enough stability to make good estimates of d_I at least for large groups. By the way, to the extend we are unable to use as deep a cross-classification as in the actual nonresponse adjustment, arguably we underestimate the amount of ignorable nonresponse.

3.3 Experimental Results of Capture/Recapture Model. There have been two major studies of NSAF nonresponse, plus numerous efforts to compare NSAF results with those of other surveys, like the CPS and NHIS (See Reports Nos. 7-8 and 15 in the 1997 NSAF Methodology Series and Nos. 6-8 in the forthcoming 1999 NSAF Methodology Series). In one of them we have employed the model articulated in Subsection 3.2 above. This is the only result we will go into in detail here.

As already mentioned, Westat carried out the main NSAF studies for both 1997 and 1999. From the main 1999 study, we selected a sample of about 2000 cases that had not been in the 1997 sample. Both interviewed and noninterviewed cases were selected. The University of Maryland's Survey Research Center was then contracted with to administer a second survey that had broadly the same purpose as NSAF but with only a little more content than that in the NSAF screener. Before proceeding to the capture/recapture step, we eliminated noncontacts from both surveys, cases not completed (in the original survey or in the second Maryland attempt), and telephone numbers that had become nonresidential between administrations.

In the main survey, weighting classes are constructed from frame variables, such as listed/unlisted status, urbanicity (central city, suburban, non-MSA) – separately in the 13 NSAF targeted states, plus the balance of the US. When completely cross-classified there are 2X3X14 = 84 cells. Within each of these, for the subsample, we created 2X2 tables of cells a, b, c, and d. In all this meant that we had a contingency table of 2X2X84 = 336 cells, many of which were very small – some even zero.

We wanted to carry out the actual capture/recapture estimates in the smallest subgroup we could in order to increase the face validity of the assumption that the p values in any given cell were constant or nearly so. Because of the cell size limitations, we elected to first smooth the overall table by setting the highest order interactions to zero. Several models were tried to gain stability. The overall capture/recapture model estimates of the fraction of nonresponse that was ignorable did not seem to be too sensitive across the range of plausible models.

The results we obtained seemed quite in accord with our understanding of NSAF screener refusals. Overall we found that about 60% of the 1999 NSAF screener nonresponse is ignorable under the models we fit. The rate of ignorability ranged from 75% for telephone exchanges outside metropolitan areas to 60% for central cities and 55% for suburban areas. Listed telephones had ignorability rates of 67% with unlisted at 52%.

Work by Black and Safir (2000), in a companion paper also given at the Joint Statistical Meetings in Indianapolis, is supportive of the modeling done here. The NSAF study by Groves and Wissoker (1999) also is in rough agreement. All of these results are consistent in that there is little or no evidence for a serious nonresponse bias, arising from a large fraction of the refusals. This conclusion has implications for the survey sponsors, in that it confirms the inference value of the effort. It also has implications for the expensive refusal conversion process and whether or not it should be redesigned, since the bias reduction potential of such efforts is questionable, relative at least to the costs incurred. Finally, and pertinent to our purpose here, the raw weighted nonresponse rate measure being used in NSAF, could be recalibrated to reflect only the potentially nonignorable portion of the nonresponse. In the 1999 NSAF, for example, the response rate at screening was 76%, with that for the extended interview being about 85%, for an overall response rate of (76%)(85%) or about 64%. If we adjusted the screener response rate to reflect only the potentially ignorable portion, it would become [100% minus (24%)(40%)] or about 90%, making the overall nonresponse rate (90%)(85%) or approximately 76%.

4. Communicating and Using Survey Micro Paradata Directly at Analysis

So far in this paper we have looked at producer Macro paradata that could lead to a richer set of survey assessment tools. These were offered, not as substitutes to existing measures but as supplements. In this section we discuss ways that clients, with data producer help, could construct their own measures and potentially improve their own use of a given survey.

The starting point for this analysis is to talk more about "macro paradata." Paradata is a term Mick Couper coined a few years ago at the Joint Statistical Meetings. As he explained it and as I have already said, paradata is a form of metadata, but focused on the measurement process itself. Most traditional quality assessment measures are summaries taken from this source (unit and item nonresponse rates, counts of editing errors detected and so on).

Generally, except for imputation flags, micro paradata are not passed on to clients. Most micro paradata are virtually never placed on public use files and seldom passed on to researchers even inside a survey organization. Implicitly this producer behavior assumes that the producer has "integrated out" all of their analytic value. But really has the producer integrated out, to the extent possible, all the measurement impacts dM

$$\int f(D,M)dM$$

of such things as field checking, weighting classes, editing, imputation, and such context variables (such as whether the interview was in English or Spanish, how many times the household was called before contact was obtained, whether an initial refusal had to be converted, etc)? Most producers talked too, it turns out, are uncomfortable with this assumption but standard practice has not reacted by putting forth an alternative.

So what would an alternative be? In rest of this section, we will very briefly illustrate, using a 1997 NSAF regression example, that there can be paradata variables of value to clients and that such variables might be profitably added to public use files. The expense here, moreover turned out to be modest, at least in our application.

The particular regression model that we chose was for health insurance coverage. We began with a standard model of insurance coverage that had as independent variables, such things as whether the person was employed, where they lived, age, citizenship, and ethnicity. The model was run for the US as a whole and then for Massachusetts separately. We then added county-level, Massachusetts geography -- a feature of NSAF in the 13 targeted states. Not surprisingly, health insurance rates varied, in a statistically significant way between Boston and western Massachusetts. When paradata dummy variables were added to the regression, like whether the interview was in English or Spanish, how hard it was to contact a respondent and whether there had been an initial refusal, etc., the coefficients on the paradata were also found to be statistically significant. The original coefficients in the original model were also affected, albeit the changes were not, in most cases, large. Some, however, were altered in value up to half their original standard errors.
This attempt, it should be noted, to examine the regression's sensitivity to paradata items was done without more that a cursory effort. Even so, the changes in coefficient values might, for some researchers, be enough to alter the inferences that might otherwise have been made. This, of course, is a researcher decision and by providing the paradata variables the producer allows the client to construct their own quality assessments – assessments that are conditional on the analysis and not just the unconditional ones that a producer might provide in, say, a quality profile. The bottom line is that we believe that it is worth the effort to put at least some paradata on public use files (For our paradata choices in NSAF, see Reports Nos. 21 and 22 in the 1997 NSAF Methodology Series).

5. Conclusions and Recommendations

In section 1 of this paper we made an attempt to set out possible weaknesses in existing quality assessment tools. In making this case, we have been perhaps too hard on the measures now in wide use. At present we assess data quality by relying heavily on macro paradata quantities developed during sample selection and survey administration. This has been a very useful practice and should continue – perhaps being extended to provide to end-users these very same measures for them to consider in their own analyses. However, we have tried to make the case that, in NSAF at least, this is not enough.

Why was NSAF different in its use of macro and micro paradata? There were several factors, primarily due to the challenges of RDD technology (with its falling unit response rates and rising coverage problems), combined with the need for highly interpretable outcome measures that did not confound treatment and measurement effects. Our 'solution" was to invent some new macro paradata measures of coverage (section 2) and nonresponse (section 3). We also used an example (in section 4) to illustrate the potential importance of micro paradata to researchers and to motivate our first attempts at providing such data on NSAF public use files. Frankly, though, even in NSAF -- let alone more

generally, we have only scratched the surface.

We do believe that in common with much else that is going in surveys (e.g., Valliant, Dorfman, and Royall 2000), more use of models is warranted in carry out quality assessments. Especially important is being very explicit in the model being chosen. Not mentioned earlier is that when producer models are unavoidable in handling nonsampling error (as they usually are), then there should be a way that clients can do sensitivity analyses on them. To repeat, the use of explicit models may, in and of itself, help clients, who often operate in a model dependent world. And with producer help, clients might be able to better integrate producer knowledge, as imbedded in models, with subject matter knowledge, also expressed in models.

In the information age we are now in, it is clear that the partnership between data producers and data analyzers should be reexamined and possibly changed. Two-way communication systems, primarily web-based, are a clear avenue for this partnership to mature. In the presence of an increasing amount and deeper understanding of nonsampling errors, the need for assessment measures that allow producers to "score" their post-survey adjustments deserves a lot more attention as well.

Acknowledgments

We are grateful for the computational work done by Aparna Lhila and H. Lock Oh. Working with Mike Brick, Pat Cunningham, and their colleagues at Westat on NSAF has been very stimulating. Thanks also go to Johnny Blair of the University of Maryland Survey Research Center. The seminal work of Dan Kasprzyk also has been very important, including the work he did on an upcoming Report by the Federal Committee on Statistical Methodology on quality measurement. Finally, the help of Pat Doyle who organized this session cannot be underestimated. She has been a leader in this area too.

References

Batcher, M. and Scheuren, F.. (1997), CATI site management in a survey of service quality, ***Survey Measurement and Process Quality***. Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin Eds. John Wiley & Sons, Inc.: New York, 573-588.

Black, T. and Safir, A. (2000), Assessing nonresponse bias in the National Survey of America's Families, Paper presented at the Joint Statistical Meetings in Indianapolis, August 2000.

Brackstone G.(1999), Managing data quality in a statistical agency, ***Survey Methodology***, 25, 2, 139-150.

Brick, J.M., J. Montaquila, and F. Scheuren. (2000), Estimating residency rates for undetermined numbers in RDD, Paper presented at American Association for Public Opinion Research Conference.

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.

Fellegi, I. P.(1999), Statistical services - preparing for the future, ***Survey Methodology***, 25, 2, 113-128.

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.

Lawless, J. F. (1982), *Statistical models and methods for lifetime data*, John Wiley & Sons, New York.

Neyman, J. (1934), On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection (with Discussion), ***Journal of the Royal Statistical Society***, XCVII (1934), 558-625.

Oh H. and Scheuren F. (1983), Weighting adjustment for unit nonresponse. Madow WG, Olkin I, Rubin D eds. ***Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies***. New York; Academic Press, Publishers, 143-84.

Petska, T. and Scheuren, F.(1992), Turning administrative systems into information systems**.** ***Journal of Official Statistics***.

Piekarksi, L., Kaplan, G., Prestegaard, J. (1999), Telephony and telephone sampling: The dynamics of change. Paper given at the AAPOR meetings in St. Petersburg, FL. See also Tucker, c., Lepkowski, J. and Piekarksi, L. 2000. List assisted sampling methods; the effect of telephone system changes on design, Washington Statistical Society presentation, November 16, 2000.

Särndal, C.E. and Swensson, B. (1992), Washington statistical Society presentation on the weaknesses in the randomization paradigm. See also Sarndal, C.E., Swensson, B., and Wretman, J. (1991),. ***Model Assisted Survey Sampling***. New York: Springer-Verlag.

Scheuren, F. (1999), Administrative records and census taking, ***Survey Methodology***. 1999.
Scheuren, F. Ed. (2000), ***National Survey of America's Families Methodology Series***, especially reports Nos. 21 and 22 in the 1997 collection.

Sekar, C. Chandra and W. Edwards Deming. (1949), On a method of estimating birth and death rates and the extent of registration, ***Journal of the American Statistical Association***, Vol. 44, No. 245. pp. 101-115.

Valliant, R., Dorfman, A., and Royall, R. (2000), ***Finite population sampling and inference: a prediction approach,*** John Wiley & Sons, Inc.: New York.