1

Chapter

Overviews

Chair: Fred Vogel, National Agricultural Statistics Service

Linda M. Ball

Leopold Granquist

Kenneth W. Harris

1

Chapter

A Paradigm Shift for Data Editing

Linda M. Ball, U.S. Bureau of the Census

Abstract

Viewed through the current paradigm, the survey process consists of collecting, editing, and summarizing survey data. We think of survey data as the "stuff" that interviewers collect, the basic units of which are individual questionnaire items. On this view, pieces of data are either erroneous or not erroneous, you can correct erroneous data, and data editing is a manageable process for most surveys.

The author proposes that we instead view the survey process as engineering and managing socio-economic information systems. In this paradigm, a survey is an expression of a mental model about society. The basic units that make up the mental model are objects or concepts in the real world about which we wish to collect information. Our mental models fail to capture fully the complexity of those objects and concepts, and a questionnaire fails to capture fully the complexity of our mental model. It is no surprise, then, that surveys yield unexpected results, which may or may not be erroneous.

When an edit detects an "error," it often can't tell whether that "error" was simply an unexpected result or one of the host of errors in administering the questionnaire and in data processing that occur regularly in the administration of surveys. If we write "brute force edits" that ensure many errors are corrected, we may miss getting feedback on the problems with the mental model underlying the survey. If we take a more hands off approach, users complain that the data set has errors and is difficult to summarize and analyze. Is it, then,any surprise that we are usually not satisfied with the results we get from edits?

Abstract (cont'd)

We get a glimpse of the true complexity of the subject matter of a survey when we study the edits of a survey that has been around for a long time.

The longer a survey has been around, the more its edits evolve to reflect the complexity of the real world. For the same reason, questionnaires tend to become more complex over time. CATI/CAPI allowed us to climb to a new level of possible questionnaire complexity, and we immediately took advantage of it because we always knew that a paper questionnaire could not be designed to handle the complexity of the subject matter of most surveys.

One way to address unexpected results is to prepare some edits in advance and use an interactive data analysis and editing process after data collection to examine unexpected results. But there is a limit to the desirability of this because of the volume of labor intensive analysis that must be done, which interferes with the timeliness of data delivery that is so valuable to many data users and increases costs.

A better alternative may be to identify or develop a methodology for approximating the mental model that underlies the survey using information engineering techniques.

Information engineering is a family of modeling techniques specifically developed for information systems. First, one approximates the mental model using information engineering techniques. Then, he or she documents the linkage between the information model and the questionnaire. Everyone who works on or sponsors the survey helps to document the model and can propose changes to it.

The information system model improves considerably over the questionnaire and procedural edits. It provides a language for representing information and relationships (for example, entity relationship diagrams or object models), allows better economy of expression, is more stable over time, is more manageable and maintainable, serves as survey documentation for data users, and serves as a basis for database design. Data relationships would replace the data edits of the current paradigm.

By adopting an information engineering paradigm, we have at our disposal many well-established, tried and tested methods for managing what we usually call survey data (what the author would call socio-economic information systems). We can take advantage of existing training, professional expertise, and software, and we can integrate the practice of survey statistics with other information technologies.

A Paradigm Shift for Data Editing

Linda M. Ball, U.S. Bureau of the Census

Introduction

A paradigm is "a set of all inflected forms based on a single stem or theme" according to the Random House Dictionary. This paper proposes a new paradigm for data editing based on the central theme:

data edit = data relationship.

Examples are provided that illustrate how the information that is normally described in terms of "IF/THEN/ELSE" procedural logic, can also be represented in the form of a logical data model (an entity-relationship diagram). (See Allen, C. Paul, 1991, for a definition..)

The implications that this paper describes as following from this central theme are the opinion of the author, and the reader is encouraged to come to her or his own conclusions about the implications. Although the implications may be a matter of opinion, the basic premise that the data relationships can be derived from current procedural edits and can be expressed as a logical data model in the form of an entity-relationship diagram is demonstrated in this paper.

For each example the following pieces of information are provided:

❐ **Current Paradigm**

- **List of Data Items**: including a short variable name for the item, a longer more descriptive variable name, a textual description of the data item, the actual questionnaire text of the question it represents (if applicable), and the possible values the data item can have.
- **Flowchart or Pseudocode**: depicting procedural edit logic.

❐ **New Paradigm**

- **Entity-Relationship Diagram**: depicting a logical information structure that is more informative than in the current paradigm.

Example 1

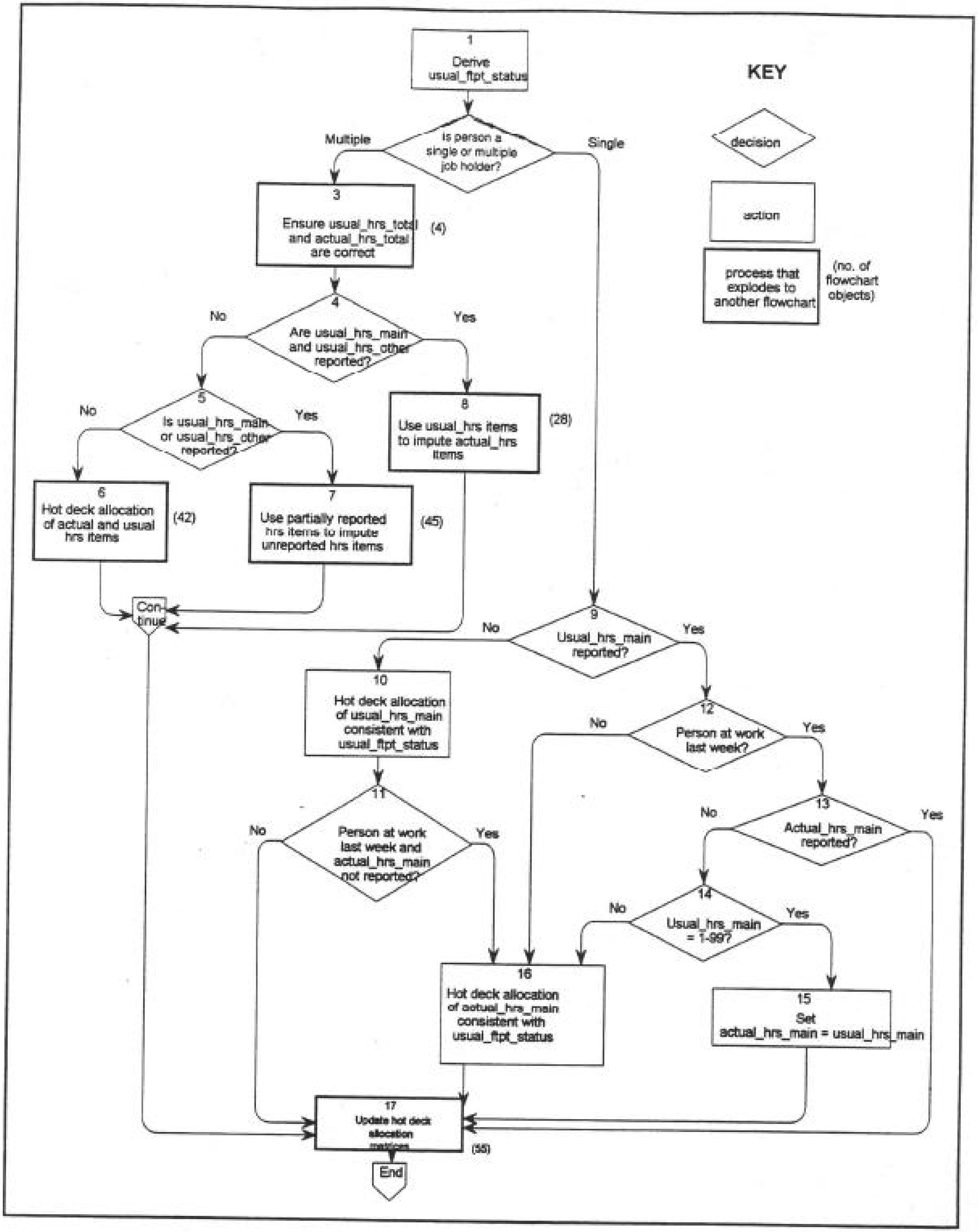
The first example is taken from the labor force section o f the *Current Population Survey*. The data items from the survey that are used in this example are shown in Table 1.

Under the current paradigm the data structure has minimal complexity and the edits have a high degree of complexity. The edits of the hours-worked items from the labor force section of the CPS questionnaire, as illustrated in Table 1 and Figure 1, take over 300 lines of pseudocode, which means some multiple of that number in FORTRAN code. This is a significantly large set of logic to document and maintain.

Table 1.--Current Paradigm: List of CPS Labor Force Data Items

Data Item		Description	Question Text (if applicable)	Values
Short Name	Long Name			
QSTNUM	QSTNUM	Unique identifier for a a questionnaire	Not applicable	1-*n* where n = the number of questionnaires in the survey
OCCURNUM	OCCURNUM	Unique identifier for a person about which the interviewer collects information	Not applicable	1-*n* where n = the number of persons interviewed at a particular address (usually 16 or less)
MJNUM	number_of_jobs	Number of jobs held last week	Altogether, how many jobs did you have?	2=2 jobs 3=3 jobs 4=4 or more jobs
HRUSL1	usual_hrs_main	Usual hours per week at main job	How many hours per week do you USUALLY work at your [main job? By main job we mean the one at which you usually work the most most hours./job?]	0-99=Number of hours v=Hours vary
HRUSL2	usual_hrs_other	Usual hours at other jobs	How many hours per week do you USUALLY work at your other (jobs/job)?	0-99=number of hours v=Hours
HRUSLT	usual_hrs_total	Sum of HRUSL1 and HRUSL2. If only one of them has a value, that value is stored in HRUSLT.	Not applicable	0-198=Number of hours v=Hours vary
HRACT1	actual_hrs_main	Actual hours at main job last week did	(So for ?) LAST WEEK, how many hours did you ACTUALLY work at your (MAIN/) job?	0-99=Number of hours
HRACT2	actual_hrs_other	Actual hours at other jobs last week	LAST WEEK, how many hours did you ACTUALLY work at your other (jobs/job)?	0-99=Number of hours
HRACTT	actual_hrs_total	Sum of HRACT1 and HRACT2. If only one of these has a value, that value is stored in HRACTT.	Not applicable	0-198=Number of hours
USFTPT	usual_ftpt_status	Usual full-time/part-time status. (derived)	Not applicable	1=Usually full time 2=Usually part time 3=Status unknown
ABSRSN	reason	Reason for absence from work last week.	What was the main reason you were absent from work LAST WEEK?	1=On layoff 2=Slackwork/business conditions 3=Waiting for a new job to begin 4=Vacation/personal days 5=Own illness/injury/medical problems 6=Child care problems 7=Other/family/personal obligation 8=Maternity/paternity leave 9=Labor dispute 10=Weather affected job 11=School/training 12=Civic/military duty 13=Does not work in the business 14=Other (specify)

Figure 1.--Current Paradigm: Flowchart Depicting Procedural Edit Logic


1
Derive usual_ftpt_status
Is person a single or multiple job holder?
Multiple
Single
3
Ensure usual_hrs_total and actual_hrs_total are correct
(4)
4
Are usual_hrs_main and usual_hrs_other reported?
No
Yes
5
Is usual_hrs_main or usual_hrs_other reported?
No
Yes
6
Hot deck allocation of actual and usual hrs items
(42)
7
Use partially reported hrs items to impute unreported hrs items
(45)
8
Use usual_hrs items to impute actual_hrs items
(28)
Con-tinue
9
Usual_hrs_main reported?
No
Yes
10
Hot deck allocation of usual_hrs_main consistent with usual_ftpt_status
11
Person at work last week and actual_hrs_main not reported?
No
Yes
12
Person at work last week?
No
Yes
13
Actual_hrs_main reported?
No
Yes
14
Usual_hrs_main = 1-99?
No
Yes
15
Set actual_hrs_main = usual_hrs_main
16
Hot deck allocation of actual_hrs_main consistent with usual_ftpt_status
17
Update hot deck allocation matrices
(55)
End
KEY
decision
action
process that explodes to another flowchart
(no. of flowchart objects)

In contrast, the diagram in Figure 2, along with supporting information in Table 2, conveys most, if not all, of the information that is represented by the flowchart in Figure 1, but in a more concise manner. From the entity-relationship diagram in Figure 2, one can get the following information about CPS data items:

- An EMPLOYED ADULT is uniquely identified by their qstnum and occurnum.
- Number_of_jobs is an attribute of EMPLOYED ADULT.
- A single EMPLOYED ADULT has one and only one MAIN EMPLOYMENT and a single instance of MAIN EMPLOYMENT is had by one and only employed adult.
- An EMPLOYED ADULT has zero or one OTHER EMPLOYMENT and a single occurrence of OTHER EMPLOYMENT is had by one and only one EMPLOYED ADULT.
- A signle instance of MAIN EMPLOYMENT is uniquely identified by qstnum and occurnum.
- Usual_hrs_main, usual_hrs_total, and usual_ftpt_status are attributes of MAIN EMPLOYMENT.
- AT WORK LAST WEEK is a subtype of MAIN EMPLOYMENT.
- AT WORK LAST WEEK [main employment] is uniquely identified by qstnum and occurnum.
- Actual_hrs_main and actual_hrs_total are attributes of AT WORK LAST WEEK [main employment].
- ABSENT LAST WEEK is a subtype of MAIN EMPLOYMENT.
- ABSENT LAST WEEK [main employment] is uniquely identified by qstnum and occurnum.
- Reason is an attribute of ABSENT LAST WEEK [main employment].

Example 2

Example 2 is taken from the Demographics section of the Current Population Survey Questionnaire. Because of the complexity and length of the CPS demographic edit only an excerpt is shown below in Figure 3. The excerpt shown performs only one of many functions within the complete edit, but the example provides a feel for what kind of logic is necessary to edit the data under the current paradigm. What makes this section of the survey worth including as an example is the many relationships that exist among data items. They are more complex than those that inherently exist in the CPS Labor Force data shown in EXAMPLE 1.

Figure 2.--New Paradigm: Entity-Relationship Diagram of CPS Labor Force Information

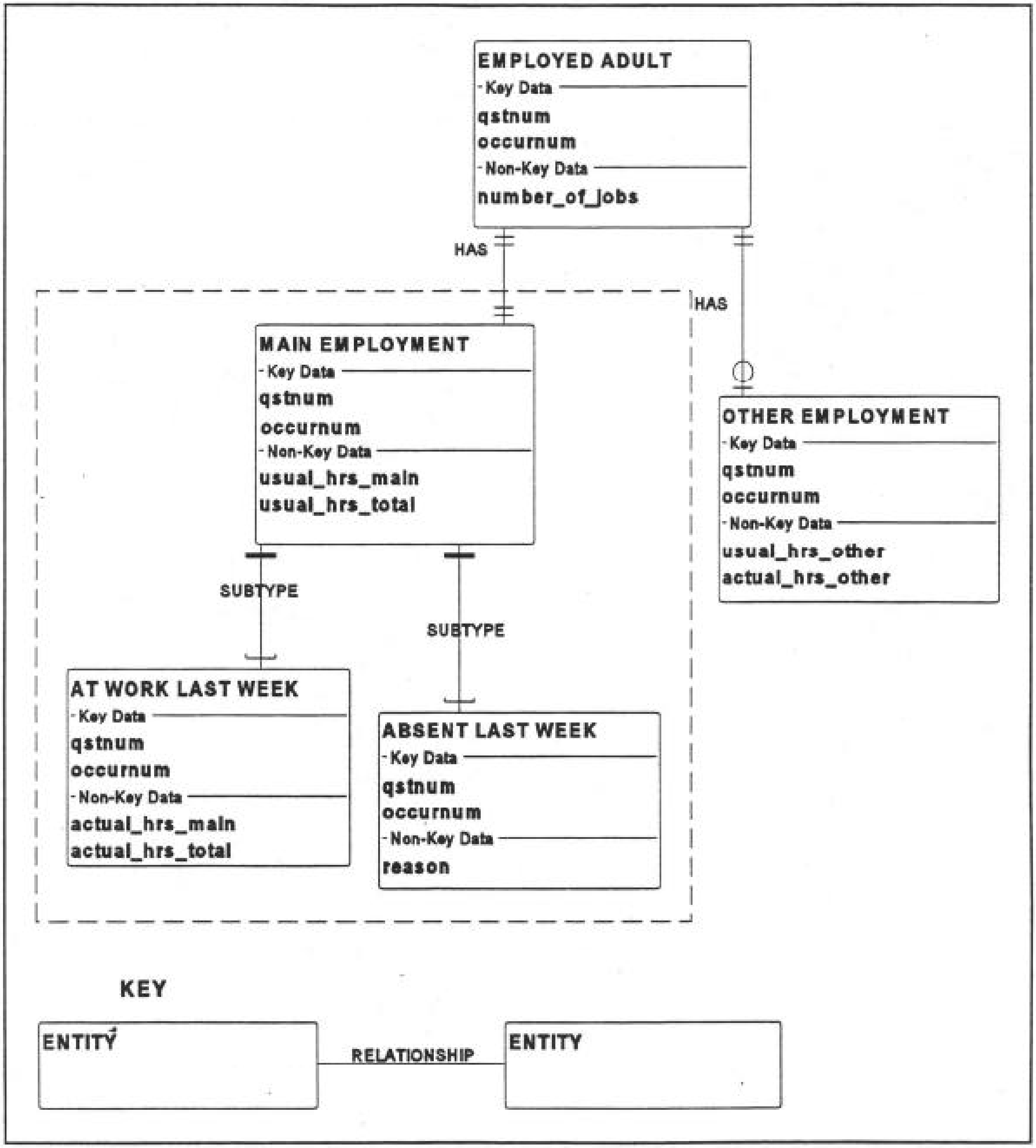


Table 2.--Current Paradigm: List of CPS Demographic Data Items

Data Item	Description	Question Text (if applicable)	Values
QSTNUM	Unique identifier for a questionnaire	Not applicable	1-*n* where n = the number of questionnaires in the survey.
OCCURNUM	Unique identifier for a person about which the interviewer collects information	Not applicable	1-*n* where n = the number of persons interviewed at a particular address (usually 16 or less)
AGE	Derived from date of birth	Not applicable	1-99
RRP	Relationship to Reference Person: Relationship to the first household member mentioned by the respondent, who is the owner or renter of the sample unit	How are you related to (reference person)?	1=Reference Person With Other Relatives in Household 2=Reference Person With No Other Relatives in Household 3=Spouse 4=Child 5=Grandchild 6=Parent 7=Brother/Sister 8=Other Relative 9=Foster Child 10=Nonrelative of Reference Person With Own Relatives in Household 11=Partner/Roommate 12=Nonrelative of Reference Person with No Own Relatives in Household 13=Nonrelative of Reference Person-Unknown Own Relatives
SPOUSE	Spouse Line Number: Line number of the person's spouse for household members whose spouse is a household member	Enter line number of spouse of [fill name] ASK IF NECESSARY	1-99=Line number 0=No one in household --
PARENT	Parent Line Number: Line number of the person's parent for household members whose parent is a household member	Enter line number of parent of [fill name] -- ASK IF NECESSARY	1-99=Line number 0=No one in household
MARITL	Marital Status	Are you now married, widowed, divorced, separated or never married?	1=Married, spouse present 2=Married, spouse absent 3=Widowed 4=Divorced 5=Separated 6=Never married
FAMNUM	Family Number: Each family unit within the household is assigned a sequential number	Not applicable	1-99
FAMREL	Family Relationship: Each family unit within the household has a reference person. Others in the family unit are assigned a code indicating their relationship to the family reference person	Not applicable	0=Not a family member 1=Reference person 2=Spouse 3=Child 4=Other relative

Figure 3.--Current Paradigm: Edit Pseudocode Excerpt

Description: If the reference person is married with their spouse present in the household, then this should be reflected consistently in the following items: RRP (Relationship to reference person); SPOUSE (Line number of spouse); MARITL (Marital status)

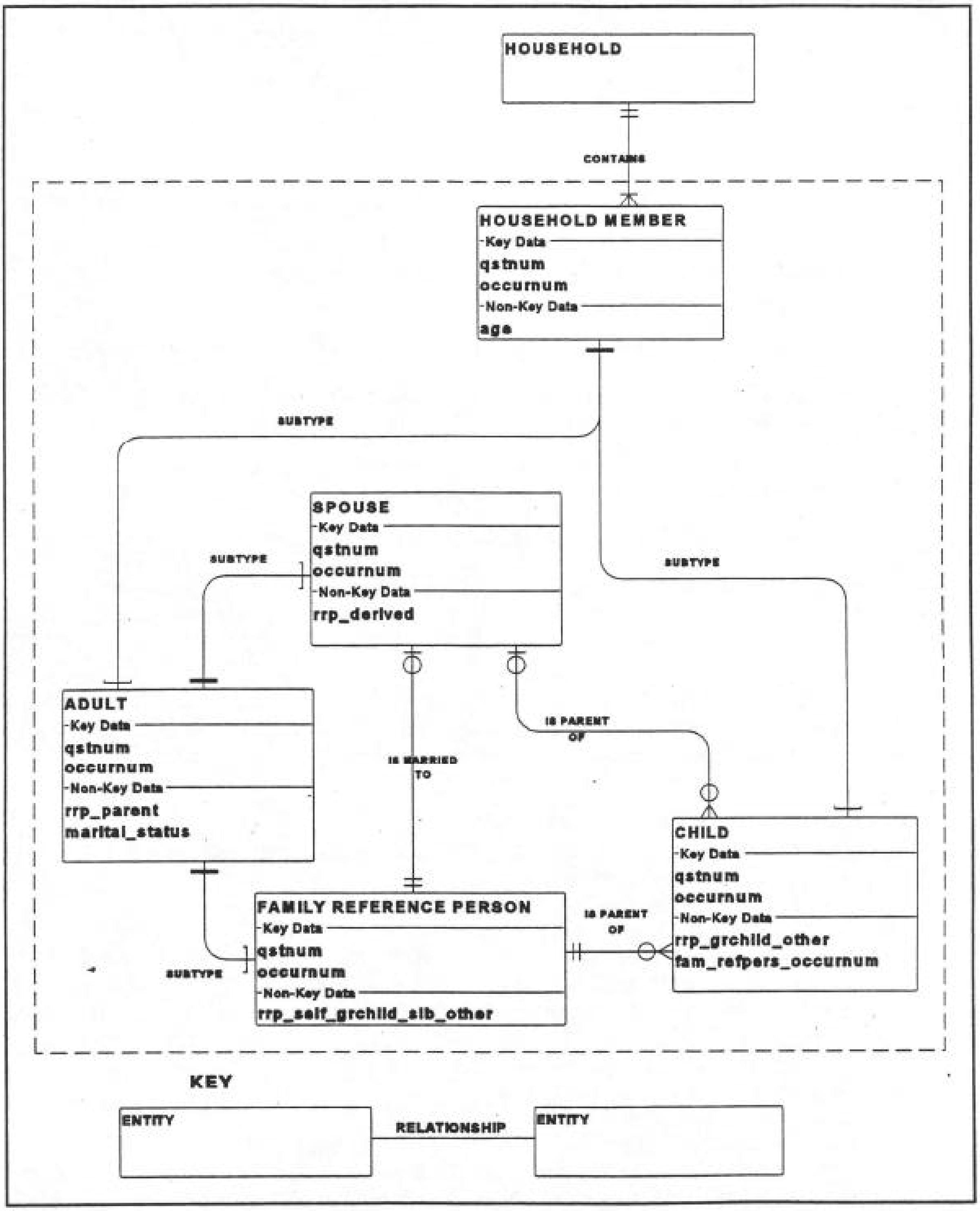
Pseudocode:

```
Do for each household:
If (person with RRP = "reference person with relatives") has SPOUSE > 0
Then Do:
    ◆ If ((person with SPOUSE = LINENO of (person with RRP = "reference person with relatives")) has RRP =
      "spouse of reference person"
    Then:
      • Set SPOUSE of (person with RRP = "spouse of reference person") = LINENO of (person with RRP =
        "reference person with relatives")
    Else:
        If SPOUSE of (person with LINENO = SPOUSE of (person with RRP = "reference person with
        relatives")) then
      • Set RRP of (person with LINENO = SPOUSE of (person with RRP = "reference person with relatives) =
        "spouse of reference person"
    Else:
      • Set SPOUSE of (person with RRP = "reference person with relatives") = blank
      • If MARITL of (person with RRP = "reference person with relatives) = "married, spouse present"
    Then:
      • allocate a value for MARITL that is one of the "unmarried" categories.
Endif
    ◆ Set MARITL of (person with RRP = "spouse of reference person") = "married, spouse present"
    ◆ Set MARITL of (person with RRP = "reference person with relatives") = "married, spouse present"
```

Again, as in EXAMPLE 1, under the current paradigm the data structure has minimal complexity and the edits have a high degree of complexity, but under the new paradigm, the logical data structure is more complex and informative. The following information is expressed in the entitity-relationship diagram in Figure 4:

- A HOUSEHOLD is uniquely identified by qstnum.
- A HOUSEHOLD contains one or many HOUSEHOLD MEMBERS, and a HOUSEHOLD MEMBER is contained by one and only one HOUSEHOLD.
- A HOUSEHOLD MEMBER is uniquely identified by qstnum and occurnum.
- Age is an attribute of HOUSEHOLD MEMBER.
- ADULT and CHILD are subtypes of HOUSEHOLD MEMBER.
- SPOUSE and FAMILY REFERENCE PERSON are subtypes of ADULT.
- ADULT, CHILD, SPOUSE, AND FAMILY REFERENCE PERSON are each uniquely identified by qstnum and occurnum.
- Rrp_parent and marital_status are attributes of ADULT.
- Rrp_grchild/other is an attribute of CHILD.
- Rrp_derived is an attribute of SPOUSE.
- Rrp_self/grchild/sib/other is an attribute of FAMILY REFERENCE PERSON.
- A SPOUSE is married to one and only one FAMILY REFERENCE PERSON, which is uniquely identified by qstnum, occurnum, and fam_refpers_occurnum where SPOUSE.fam_refpers_occurnum = FAMILY REFERENCE PERSON.occurnum; and a FAMILY REFERERNCE PERSON is married to zero or one SPOUSE.

Figure 4.--New Paradigm: Entity-Relationship Diagram of CPS Demographic



- A CHILD is the child of one and only one FAMILY REFERENCE PERSON which is uniquely identified by qstnum, occurnum, and fam_refpers_occurnum where CHILD.fam_refpers_occurnum = FAMILY REFERENCE PERSON.occurnum; and a FAMILY REFERENCE PERSON is parent of zero, one, or many CHILDREN.
- A CHILD is the child of zero or one SPOUSE which is uniquely identified by qstnum, occurnum, and parent2s_occurnum where CHILD.parent2s_occurnum = SPOUSE.occurnum; and a SPOUSE is the parent of zero, one, or many CHILDREN.

Problems With the Current Paradigm

If the central theme of the new paradigm is expressed as data edit = data relationship, then the central theme of the current paradigm would have to be expressed as:

data edit = logical process for changing a data item.

From an operational perspective, there are two distinct types of edits.

Type 1 allocates or imputes missing data and outliers and inconsistencies that the edit authors know about prior to data collection because of their knowledge of what needed to be edited in previous survey iterations, i.e. previous iterations of either the survey in question or another similar survey.

Type 2 allocates or imputes new values for unexpected results. These are outliers or inconsistencies that the edit authors do not know about until they or someone else examines the edited data from the survey. The real world socio-economic concepts about which we collect information, are more complex than the assumptions conveyed by a questionnaire (especially in the paper questionnaire environment). It is no surprise then that we get unexpected results. For example, until recently CPS did not allow same-sex married couples in its data. Our assumptions told us that if we found this in the dataset it was probably a data collection or keying error. The edit checked for this and edited the data to disallow it. Recently it was decided to allow same-sex married couples in CPS data. Our mental model changed based on our information about society, and this change was reflected eventually in the CPS questionnaire and edits. A difficulty with post-data-collection edits is that we don't always know whether we are changing a true outlier or correcting an error that occurred in any of the survey's operational processes leading up to the edits.

It is the Type 2 edits that consume the most resources during the time when survey operations staff are trying to meet deadlines for delivery of data to the survey sponsor. These unexpected results can originate from any point in any of the processes from questionnaire design through reformatting of data.

If we write edits that ensure all outliers are eliminated, we may miss getting feedback on the problems with the survey design or various operational procedures. If we take a more hands off approach, users complain that the data set has errors and is difficult to summarize and analyze.

The current paradigm leads to excess complexity in the edit process. To get a glimpse of the true complexity of the subject matter of a survey, one should study the edits of a survey that has been in operation for a long time. Because the longer a survey has been in operation, the more of the true complexity of the real world has been incorporated into the edits. For the same reason, questionnaires tend to become more complex over time. CATI/CAPI has allowed us to climb to a new level of possible questionnaire complexity, and we immediately took advantage of it because we always knew that a paper questionnaire could not be designed to handle the true complexity of the subject matter of most surveys.

To illustrate this tendency toward complexity, compare the CPS paper questionnaire that was in use prior to 1994 to the electronic version used since 1994. The paper questionnaire plus control card filled about 15-20 pages of condensed print. The electronic version fills hundreds of pages and has many more logical paths than the paper questionnaire.

Data-edits-as-procedures work well when they are attached to a data collection process that has been repeated a number of times without change. However, if a survey is annual or every 5 years and the survey design changes significantly each time, software maintenance side effects dominate and the edits can be unmanageable, or at least very expensive to manage (Pressman, 1992).

In other words there are two competing forces acting upon the manageability of edits. One tends to increase manageability over time, the other tends to decrease manageability over time. In the long run, even in surveys that don't change for a while, software maintenance side effects eventually take over because in the long run changes to a survey design always become necessary. These side effects are:

- Edits are often complex and difficult to understand and document.
- Any change to the questionnaire causes changes in the edits because of the very high degree of dependence between them.
- Changes to the edits are often poorly documented because they are developed in a hurry after data collection has ended.
- The accumulation of additions to the original edit design over time causes added complexity in the logic.
- People who understand the edits leave and newly hired people have a long learning curve and poor documentation to follow.
- Changes to one part of the edits may cause another part to work incorrectly.

Under the current paradigm, there is a tradeoff between usability of data and timeliness of data. If you accept the premise that a survey will yield unexpected results every time you implement a new or revised questionnaire or operational procedure, then you must conclude that data must be inspected and adjusted after data collection in order to provide the data usability that sponsors and end users require. Therefore you might conclude that you can have some edits prepared in advance but need an interactive data analysis and editing process after data collection to deal with unexpected results. Currently, there is a great desire for increased timeliness, but there is also a minimum standard for usability of data files that cannot be sacrificed.

A New Paradigm: Implications of the Central Theme

- ***Decreased Time Between Data Collection and Data Delivery***.--This could be achieved by

    - capturing data relationships as data is entered during the interview, or immediately after the interview, while the interviewer still has access to it; and

    - giving other participants in the survey process, no matter where they are physically located, immediate access (with appropriate confidentiality constraints) to that data and stored data relationships (Hammer and Champy, 1993) .

- ***Increased Maintainability.***--Separate technology maintenance from data administration. It is common wisdom within the software engineering field that the information in a system tends to be more stable over time than the processes in a system. Processes tend to be more technology dependent than information. Shifting some of the complexity of a survey from its component processes to its component information structures should make the operation as a whole inherently more maintainable.

- ***More Accurate and Up-to-date Documentation for Users.***--A logical data model, once developed could be used not only by the survey developers, but also by data users.

Difficulties

In addition to the beneficial implications listed above, it must be stated that there would most likely be difficulties in implementing a survey based on this new paradigm. Firstly, it would most certainly take more lead development time the first time it is tried for any given survey. Survey content experts would have to come together on a logical model of the data collected by the survey.

Secondly, there may be organizational problems involving the role of interviewers, the technology skills of survey staff, and the necessity of many organizational units coming together on a strategy for survey operations that emphasizes the joint development and sharing of complex datasets.

References

Allen, C. Paul, (1991). *Effective Structured Techniques from Strategy to Case*, New York, NY: Prentice Hall.

Hammer, M. and Champy, James (1993). Chapter 5 of *Reengineering the Corporation: The Enabling Role of Information Technology*, New York, NY: Harper Collins.

Pressman, Roger S. (1992) *Software Engineering: A Practitioner's Approach*, New York, NY: McGraw-Hill, 1992

Bibliography

Bureau of the Census, Demographic Surveys Division, *Edit Specifications for the 1994 Current Population Survey CATI/CAPI Overlap Test.*

Bureau of the Census, Demographic Surveys Division, *Data Documentation for the 1994 Current Population Survey CATI/CAPI Overlap Test.* ■

1

Chapter

The New View on Editing

Leopold Granquist, Statistics Sweden

Abstract

An international new view on editing has grown during the last fiveten years out of the results of research on editing carried out by eminent statistical agencies. The research shows that editing is expensive (20-40 percent of the budget cost), inefficient (the impact on quality negligible), hiding data collection problems, but that new strategies and methods may lower the cost substantially for the producer as well as for the respondent and in the long run increase the quality of data. The main point is gradually moving from cleaning up the data to identifying and collecting data on error sources, problem areas and error causes to get a basis for measures to prevent errors to arise in the data collection and processing. Thus, the editing process should produce data on the collection and processing, so called paradata, for a continuous improvement of the whole survey vehicle. This Total Quality Management (TQM) view on editing should imply lower cost and increased quality when checks are co-ordinated with the response and collection process and adapted to the respondent ability to provide data.

High quality cannot be accomplished by introducing as many and tight checks as possible and augmenting the number of follow ups with the repondents, but through careful design and testing of the set of edits, and fitting the checks currently to the data to be scrutinised. New types of edits and strategies should be used to focus the editing to those serious errors, which can be identified by editing. An important feature is to classify edits into critical and query edits. The critical edits shall be used to detect and remove fatal errors, that is those errors which the editing process has to remove from data. The query edits should be concentrated on those suspicious data, which when containing errors, may have a substantial impact on the estimates. The re-contacts to respondents have to be limited as much as possible, but when considered necessary, the contact should be used not only to find better data but to get intelligence of causes of errors, error sources and respondent problems of providing accurate data. The new technology with more and more powerful personal computers plays an important role for using new more efficient editing methods, as graphical editing; new strategies, as data entry editing and moving the editing closer to the data source in CAI and CASI modes of data collection. It is stressed that it is not an issue of translating old methods to a new technology, but to re-engineer the whole editing process under a new view on editing.

The New View on Editing

Leopold Granquist, Statistics Sweden

Introduction

It may be claimed that a new international common view on editing has grown the last five - ten years and become established through papers presented at the ISI conferences in Cairo 1991 (Linacre, 1991) and Florence 1993 (Lepp and Linacre, 1993; Silva and Bianchini, 1993; among others), the International Conference on Establishment Surveys at Buffalo 1993 (Granquist, 1995 and Pierzchala, 1995), the International Conference on Survey Measurement and Process Quality at Bristol 1995 (Granquist and Kovar, 1996), and at the annual Work Sessions on Statistical Data Editing organized by United Nations Statistical Commission and Economic Commission for Europe (ECE, 1994 and ECE, 1996).

The emphasis of the editing task is moving from just cleaning up the data, though still a necessary operation, to identifying and collecting data on errors, problem areas, and error causes to provide a basis for a continuous improvement of the whole survey vehicle. This Total Quality Management (TQM) view on editing implies lower cost for the producer, less respondent burden, and increased quality as checks are integrated and co-ordinated with the response and collection process and adapted to the respondent ability to provide accurate data.

This change in the editing process has been embraced by some statistical agencies when it was recognized that the heavy cost of editing cannot be justified by quality improvements as measured by numerous evaluation and other studies on editing. Some facts:

❒ Editing accounts for a substantial part of the total survey budgets. The monetary costs (hardware, software, salary and field costs) amount to 20-40 percent of the survey budget (Granquist,1984; Federal Committee on Statistical Methodology, 1990; and Gagnon et al, 1994, among others). Furthermore, there are costs related to lost opportunities, response burden and associated bad will, costs related to losses in timeliness (e.g., the machine editing of the World Fertility Survey delayed the publication of the results by about one year, Pullum et al., 1986), and indirect costs related to undue confidence in data quality and respondent reporting capacity and to using employees in inappropriate tasks (Granquist and Kovar, 1996).

❒ The ever ongoing rationalization of the editing process has not yet caused the process to be less costly or more efficient. The gains have been invested in attempts to raise the quality by applying more checks and to selecting more forms for manual review and follow-up (Pierzchala, 1995), resulting in only marginal improvement or more likely in overediting. There are even some examples that editing can be counter productive (Linacre and Trewin, 1989; Corby, 1984; Corby, 1986). Furthermore, some important error types cannot even be touched by editing (Pullum et al., 1986; Christianson and Tortora, 1995) or are impossible to identify by traditional error checks (Werking et al., 1988). Thus, editing is far less important for quality than survey managers believe, and high quality cannot be guaranteed by just adding more checks and augmenting the number of recontacts. On the contrary, such an approach may be counter productive and, worse of all, impose an undue confidence in the quality of the survey in the survey managers, in particular if editing is hiding problem areas instead of highlighting them.

- ❒ During the last five years new kinds of editing methods have been developed.They are termed macro-editing, selective editing or significant editing. Numerous studies and experiences of these methods indicate that the manual verifying work can be reduced by 50 percent or more without affecting the estimates (Granquist and Kovar, 1996). By removing unnecessary edits, relaxing the bounds of the remaining query edits by applying a method suggested in Hidiroglou-Berthelot (1986), and using a score function for identifying records for manual review developed by Latouche and Berthelot (1992), Engström (1995) succeeded in decreasing the manual review of a well-edited survey by 86 percent without significant consequences on the quality. The success of this research illustrates how important it is to design the set of query edits meticulously.

- ❒ The technological development has made it possible to move the editing closer to the data source, preferable while the respondent is still available. It opens the possibilities of getting more accurate data from the respondent, intelligence of what data are received and of the problems the respondent experiencies in delivering the requested data, comments concerning answers, and opportunities of conducting experiments. In general, CAI and CASI modes of data collection offer excellent possibilitties of getting data on error sources, problem areas and the reporting capacity among the respondents.

The Role of Editing

Definitions of data editing vary widely. Here editing is defined as the procedure for identifying, by means of edit rules, and for adjusting, manually or automatically, errors resulting from data collection or data processing (Granquist and Kovar, 1996).

An absolute and fundamental requirement on editing has always been, and should be, to identify outliers and those errors in individual data, which are recognizable as such to a user with access to individual data records, but without knowledge of the particular unit. Edits aimed at identifying such data , that is data which certainly are erroneous we call fatal edits and the process to ensure validity and consistency of individual data records is termed micro-editing (Granquist and Kovar, 1996). However, in surveys with quantitative data there might be errors, although not fatal, which significantly affect the estimates. Hence, query edits, that is edits pointing to suspicious data items, have to be added to the editing process.

As early as in the sixties, it was recognized that removing errors was not the most important aspect of editing, Pritzker et al. (1965) claim that it is more important to identify error sources or problem areas of the survey. Granquist (1984) agrees and says that the goals of editing should be threefold: To provide information about the quality of the data, to provide the basics for the (future) improvement of the survey, and to tidy up the data.

Editing -- A Historic Review

The low efficiency of editing processes in general is basically due to the survey managers ambition to allocate almost all resources to the third objective, cleaning up the data, especially to identifying errors by query edits. An explanation may be found in the following brief historical background.

Before the advent of computers, editing of individual forms was performed by large groups of clerks, often without any secondary school education. Though ingenious practices sometimes were developed, only simple checks could be undertaken. Editing was inconsistent not only between individual clerks but also over time for the same person. Only a small fraction of all errors could be detected. The advent of computers was recognized by survey designers and managers as a means of reviewing all records by consistently applying even sophisticated checks requiring computational power to detect most of the errors in data that could not be found by means of manual review. The focus of both the methodological work and in particular the applications was on the possibilities of enhancing the checks and of applying automated imputation rules in order to rationalize the process, Naus (1975). Nordbotten (1963) was the theoretical basis for the checks and imputation rules used in the cumbersome main frame automated systems developed in the late sixties and the seventies. Implicitly the process was governed by the paradigm: The more checks and recontacts with the respondents the better the resulting quality. The early systems produced thousands of error messages that had to be manuall examined, by referring back to the original forms and in more complicated cases to the respondents themselves. Changes were entered in batch and edited once again by the computer. Many records passed the computer three or four times and sometimes records were reviewed by different persons each time they were flagged. Occasionally cycles of 18 could occur (Boucher, 1991).

The research and development work became focused on rationalizing the EDP departments work of providing the survey managers with application programs and on developing generalized software.A methodology for generalized editing and imputation systems was developed by Fellegi and Holt (1976), implemented for editing of categorical data (e.g., CAN-EDIT, AERO, DIA), and for quantitative data (e.g., SPEER and GEIS) ECE (1994). These systems are well suited for removing fatal errors, defined as data that do not meet certain requirements.

The great break in rationalizing the process came as a direct consequence of the PC revolution in the eighties, when editing could be performed on-line on personal computers, at the data entry stage -- data entry heads-up -- during the interview, and by the respondent in CASI modes of data collection. Bethlehem et al. (1989) describe in detail the gains in on-line data entry editing on micros or minis as compared to main frame processing systems, and on the basis of their findings Statistics Netherlands developed the world-renowned system BLAISE. The process was substantially streamlined, but the gains were often used to allow the editors to process more records and to make more contacts with respondents to resolve encountered problems (Pierzchala,1995). The paradigm -- the more checks and contacts the better the quality -- was still considered valid. However, some evaluations carried out around 1990 said something else, that later on was corroborated in numerous evaluation and other studies on editing methods: 10 to 15 percent of the biggest changes made in the editing of economic items contribute to around 90 percent of the total change; 5 to 10 percent of the biggest changes bring the estimate within 1 percent of the final global estimate; only 20 to 30 percent of the recontacts result in changed values; the quality improvements are marginal, none or even negative; many types of serious systematic errors cannot be identified by editing (Granquist and Kovar, 1996; Werking and Clayton, 1988; Christianson and Tortora, 1995; Linacre and Trewin, 1989; among others).

Why Over-Editing Occurred

The main reason why editing processes remained inefficient is that processes seldom were evaluated, nor were indicators or other performance measures produced. Initially the approach seemed to be successful. Errors, even serious errors were detected. Ambitious survey managers and designers thus continued to see editing as an outstanding tool to achieve high-quality data, and relied upon that editing to fix any mistakes committed in earlier phases of the data collection and processing (including the survey and in particular the questionnaire design). Results from evaluations of editing processes contradicting this view were not considered applicable to their surveys. They still believed that investing in more checks and recontacts would be beneficial for the quality. But editing can be harmful to the quality, e.g., delaying the publishing of the results; causing bias when only certain types of errors are detected, for example those which move the estimates in a specific direction; inserting errors, when the reviewer manipulate data to pass the edits, so called creative editing Granquist (1995), which -also may give a wrong impression about the reporting capacity of the respondents; introducing errors by mistakes occurring in the recontact process.

Corner Stone of the New View

We have already established, that the primary and basic requirement on editing is to identify outliers and to remove fatal errors from individual data for which traditional editing is well suited. However, it is the second type of edits, the query edits, that are responsible for the high costs of editing, and accordingly the subject of this paper.

Careful Design and Evaluation of the Set of Query Edits

The design of the entire set of query edits is of particular importance in getting an acceptable cost/ benefit outcome of editing (Granquist, 1995). The edits have to be co-ordinated for related items and adapted to the data to be edited. Probable measures for improving current processes are relaxing bounds by replacing subjectively set limits by bounds based on statistics from the data to be edited, and removing edits which in general only produce unnecessary flags. It should be noted that there is a substantial dependence between edits and errors for related items. Furthermore, the edits have be targeted on the specific error types of the survey, not on possible errors.

The properties of the edits have to be controlled continuously for example by examining data of the outcome of edits. It is an absolute requirement on any editing system to produce statistics on error flags, changes related to edits and reasons for imputation, etc.

Focus on Influential Item Values and Records

The edits and/or the system should have built-in prioritizing rules to focus the review on the suspicious data items or records that have most influence on the estimates A leading principle in most of the methods suggested in Granquist (1991) is: begin with the most deviating values and stop verifying when estimates no longer are changed.

Another way of prioritizing the recontacts is to utilize score functions, as suggested by Latouche and Berthelot (1992) or Lawrence and McDavitt (1994). The idea is: to run the edits as decided by the subject matter experts, assign a score to every record with at least one flagged item according to the weight of the record, the potential impact of the suspicious item value, and the importance of the flagged item; and review only those records, which get a score exceeding a threshold value, determined in advance on basis of historical experience. When fatal edits are included, then records containing fatal errors have to be handled automatically in cases where the score of that edit does not guarantee that the threshold value is exceeded.

Numerous studies indicate that applying any of these methods will yield a reduction of the manual review work by 50 percent or more without any significant impact on quality (Granquist and Kovar, 1996). In an experimental study, Engström (1995) uses all the methods and eliminates 84 percent of the current review work.

Focus on Response Problems and Error Causes

The main aim of recontacting respondents should be to collect intelligence of respondent problems, error causes and reporting capacity. It is essential to look upstream to reduce errors in survey data, rather than to attempt to clean up single cases at the end. It is fundamental for the data quality that respondent data are of high quality. It means that the respondent in business surveys has to: understand exactly the question and underlying definitions; have data available in his information system; understand differences in definitions between the survey and his information system, and in case of large differences be able to give an acceptable estimate. Accordingly, to provide accurate data the survey design has to be adapted to the respondent's possibilities and conditions. Editing has to highlight respondent problems and reporting capacity, not hide them. This can be accomplished by changing the focus of the recontact process from ascertaining whether a suspicious value is wrong and finding a more accurate value, to acquiring knowledge of respondent problems and causes of errors. Thus, editing can be used to advantage in sharpening survey concepts and definitions and in improving the survey vehicle design.

CASI modes of data collection offer an excellent tool for furnishing intelligence about the response process and the accuracy of delivered data, provided that: edits and error messages are designed to help the respondent in understanding what data we want from him or her; possible error causes, definitions and other information needed for answering each particular item are prompted; the respondent can easily give comments to answers or explain why he or she cannot answer the question; warnings like did you include or exclude this and that component. Furthermore, experiments can be built in to get statistics on respondent behaviour, Weeks (1992) gives a detailed description of the possibilities.

The New View on Editing

Editing should be integrated with, but subordinated to, collection, processing and estimation. A main task is to provide a basis for designing measures to prevent errors.

Editing should be considered a part of the total quality improvement process, not the whole quality process. Editing alone cannot detect all errors, and definitely not correct all mistakes committed in survey design, data collection, and processing.

The paradigm -- the more (and tighter) checks and recontacts, the better the quality -- is not valid.

The entire set of the query edits should be designed meticulously, be focused on errors influencing the estimates, and be targeted on existing error types which can be identified by edits. The effects of the edits should be continuously evaluated by analysis of performance measures and other diagnostics, which the process shoud be designed to produce.

Editing has the following roles in priority order:

❒ Idẽntify and collect data on problem areas, and error causes in data collection and processing, producing the basics for the (future) improvement of the survey vehicle

❒ Provide information about the quality of the data

❒ Identify and handle concrete important errors and outliers in individual data.

Besides its basic role to eliminate fatal errors in data, editing should highlight, not conceal, serious problems in the survey vehicle. The focus should be on the cause of an error, not on the particular error per se.

References

Bethlehem, J.G.; A.J., Hundepool; M.H., Schuerhoff; and L.F.M., Vermeulen (1989). BLAISE 2.0 An Introduction, Vorburg, The Netherlands: Central Bureau of Statistics.

Boucher, L. (1991). Micro-Editing for the Annual Survey of Manufactures: What is the Value Added? *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 765-781.

Corby, C. (1984). *Content Evaluation of the 1977 Economic Censuses*, Statistical Research Division Report Series No. CENSUS/SRD/RR-84-29, Washington, DC: U.S. Bureau of the Census.

Corby, C. (1986). Content Evaluation of the 1982 Economic Censuses: Petroleum Distributors, *1982 Economic Censuses and Census of Governments, Evaluation Studies*, Washington DC: U. S. Department of Commerce, pp. 27-60.

Christianson, A. and R. Tortora (1995). Issues in Surveying Business: An International Survey, in B.G. Cox; D.A. Binder; N. Chinnappa; A. Christianson; M.J. Colledge; and P.S. Kott (eds.). *Business Survey Methods*, New York: Wiley, pp. 237 - 256.

Economic Commission for Europe, (1994). Statistical Data Editing: Methods and Techniques, Volume No. 1, United Nations New York and Geneva.

Economic Commission for Europe, (1996). Statistical Data Editing: Methods and Techniques, Volume No. 2, United Nations New York and Geneva (to appear).

Engström, P. (1995). A Study on Using Selective Editing in the Swedish Survey on Wages and Employment in Industry, Room paper No. 11, presented at the Conference of European Statisticians, Work Session on Statistical Data Editing, Athens, Greece, November 6-9, 1995.

Federal Committee on Statistical Methodology (1990). *Data Editing in Federal Statistical Agencies*, Statistical Policy Office, Working Paper 18, Washington, DC: U.S. Office of Management and Budget.

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, 71, pp. 17-35.

Gagnon, F.; Gough, H.; and Yeo, D. (1994). Survey of Editing Practices in Statistics Canada, unpublished report, Ottawa: Statistics Canada.

Granquist, L. (1984). On the Role of Editing, *Statistisk Tidskrift*, 2, pp. 105-118.

Granquist, L. (1991). A Review of Some Macroediting Methods for Rationalizing the Editing Process, *Proceedings of Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada, pp. 225-234

Granquist, L. (1995). Improving the Traditional Editing Process, in B.G.Cox; D.A. Binder; N.Chinnappa; A. Christianson; M.J. Colledge; and P.S.Kott (eds.) *Business Survey Methods*, New York: Wiley, pp. 385-401.

Granquist, L. and Kovar, J.G. (1996). Editing of Survey Data: How Much is Enough? in L. Lyberg et al. (eds.), *Survey Measurement and Process Quality*, New York: Wiley (to appear).

Hidiroglou, M. A. and Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, 12, pp. 73-84.

Latouche, M. and Berthelot, J.-M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys, *Journal of Official Statistics*, 8, pp. 389-440.

Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings, *Journal of Official Statistics*, 10, pp. 437-447

Lepp, H, and Linacre, S. (1993). Improving the Efficiency and Effectiveness of Editing in a Statistical Agency, *Bulletin of the International Statistical Institute: Proceedings of the 49th Session*, Florence, Italy, Contributed Papers Book 2, pp. 111-112.

Linacre, S. J. (1991). Approaches to Quality Assurance in the Australian Bureau of Statistics Business Surveys, *Bulletin of the International Statistical Institute: Proceedings of the 48th Session*, Cairo, Egypt, Book 2, pp. 297-321.

Linacre, S. J. and Trewin, D. J. (1989). Evaluation of Errors and Appropriate Resource Allocation in Economic Collections, *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 197-209.

Naus, J. I. (1975). *Data Quality Control and Editing*, New York: Marcel Dekker, Inc.

Nordbotten, S. (1963). Automatic Editing of Individual Statistical Observations, Conference of European Statisticians, Statistical Standards and Studies, No. 2, New York: United Nations.

Pritzker, L, J. Ogus and Hansen, M.H. (1965). Computer Editing Methods -- Some Applications and Results, *Proceedings of the International Statistical Institute Meetings* in Belgrade Yugoslavia, 1965, pp. 442-465.

Pierzchala, M. (1995). Editing Systems and Software, in B.G. Cox; D.A. Binder; N. Chinnappa; A. Christianson; M.J. Colledge; and P.S. Kott (eds.) *Business Survey Methods*, New York: Wiley, pp. 425-441.

Pullum, T.W.; Harpham, T.; and Ozsever , N. (1986). The Machine Editing of Large-Sample Surveys: The Experience of the World Fertility Survey, *International Statistical Review*, 54, pp. 311-326.

Silva, P. L. and Bianchini , Z. (1993). Data Editing Issues and Strategies at the Brazilian Central Statistical Office, *Bulletin of the International Statistical Institute: Proceedings of the 49th Session*, Florence, Italy, Contributed Papers Book 1, pp. 377-378.

Weeks, M.F. (1992). Computer-Assisted Survey Information Collection: A Review of CASIC Methods and Their Implications for Survey Operations, *Journal of Official Statistics*, 8, pp. 445-465.

Werking, G.; Tupek, A.; and Clayton, R. (1988). CATI and Touchtone Self-Response Applications for Establishment Surveys, *Journal of Official Statistics*, 4, pp. 349-362. ■

1

Chapter

Data Editing at the National Center for Health Statistics

Kenneth W. Harris, National Center for Health Statistics

Abstract

Data editing can be defined as the procedures designed and used for detecting erroneous and/or questionable survey data with the goal of correcting as much of the erroneous data as possible, usually prior to data imputation and summary procedures. This paper describes many of the data editing procedures used for selected data systems at the National Center for Health Statistics (NCHS), the Federal agency responsible for the collection and dissemination of the nation's vital and health statistics.

Data Editing at the National Center for Health Statistics

Kenneth W. Harris, National Center for Health Statistics

Background

The National Center for Health Statistics (NCHS) is the Federal agency responsible for the collection and dissemination of the nation's vital and health statistics. To carry out its mission, NCHS conducts a wide range of annual, periodic, and longitudinal sample surveys and administers the national vital statistics registration systems. These sample surveys and registration systems form four families of data systems: vital event registration systems, population based surveys, provider based surveys, and followup/followback surveys.

Much of what happens to the data covered by these data systems, from collection through publication, depends on the family to which they belong. At most steps along the way, various activities and operations are implemented with the goal of making the data as accurate as possible. These activities and operations are generally categorized under the rubric, "data editing." In the 1990 Statistical Policy Working Paper 18: *Data Editing in Federal Statistical Agencies*, [1] data editing is defined as:

Procedure(s) designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much of the erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures. [However, this report includes data imputation procedures.]

As will be shown in this report, data editing procedures vary greatly between NCHS data systems.

Twenty-four data systems are included in this report (see Table 1). For each data system, summary descriptions of NCHS data editing practices are provided in the following 11 areas:

Environment in Which Survey Takes Place
Data Processing Environment and Dispersion of the Work
Audit Trail
Micro-, Macro-, and Statistical Editing
Prioritizing of Edits
Imputation Procedures
Editing and Imputation Standards
Costs of Editing
Role of Subject Matter Specialists
Measures of Variation
Current and Future Research.

Table 1.--NCHS Data Systems Included in this Report

Registration Systems (8)

Mortality (MRS)
Fetal Mortality (FMRS)
Abortion (ARS)
Natality (NRS)
Marriage (MRG)
Divorce (DRS)
Current Mortality Sample (CMS)
Linked Birth and Infant Death Data Set (LBIDDS)

Population Based Surveys (3)

National Health Interview Survey (NHIS)
National Health and Nutrition Examination Survey (NHANES)
National Survey of Family Growth (NSFG)

Provider Based Surveys (7)

National Hospital Discharge Survey (NHDS)
National Survey of Ambulatory Surgery (NSAS)
National Ambulatory Medical Care Survey (NAMCS)
National Hospital Ambulatory Medical Care Survey (NHAMCS)
National Nursing Home Survey (NNHS)
National Home and Hospice Care Survey (NHHCS)
National Health Provider Inventory (NHPI)

Followup/Followback Surveys (6)

National Maternal and Infant Health Survey (NMIHS)
1991 Longitudinal Followup (LF) to the NMIHS
National Mortality Followback Survey (NMFS)
National Health and Nutrition Examination Survey (Cycle I) Epidemiologic Followup Study (NHEFS)
Longitudinal Study of Aging (LSOA)
National Nursing Home Survey Followup (NNHSF)

Within each of these areas, data editing practices are grouped according to the type of data system, i.e., vital event registration systems, population based surveys, provider based surveys, and followup/followback surveys.

Environment in Which Survey Takes Place

Registration Systems

The vital event registration systems cover six vital events: mortality, fetal mortality, induced termination of pregnancy (abortion), natality, marriage and divorce. For each of these systems, data are obtained from certificates and reports filed in state registration offices and registration offices of selected cities and other areas. Coverage for each registration system is limited to its prescribed registration area (RA). The oldest registration areas, mortality, fetal mortality, and natality, have been complete since 1933. These three are national data systems; i.e., they cover the entire United States. The marriage RA started in 1957 with 30 states and reached its current coverage of 42 states plus selected areas in 1986. The Divorce RA started in 1958 with 14 states and by 1986 had expanded to 31 states plus selected areas [2,3]. The Abortion RA started in 1977 with five states and reached its current coverage of 14 states in 1987.

Mortality (approximately 2,000,000 annual events), fetal mortality (60,000) and natality (4,000,000) registration are required by all states; registration completeness for the mortality and natality systems exceeds 99 percent. The Abortion RA collects information on approximately 300,000 abortions per year, about 22 percent of the annual U.S. total. (Because of budgetary constraints, NCHS has not processed abortion data since 1993). The Marriage RA, excluding Puerto Rico and the Virgin Islands, covers approximately 81 percent (785,000) of U.S. marriages. The Divorce RA, excluding the Virgin Islands, accounts for 49 percent (280,000) of the annual U.S. divorce count.

In addition to these six registration systems, two other data systems, the Current Mortality Sample (CMS) and the Linked Birth and Infant Death Data Set, are based on data obtained from the Mortality and Natality Registration Systems. The CMS is a 10 percent systematic sample taken from the regular mortality file on a monthly (month of death) basis. The CMS covers the 50 states, the District of Columbia and New York City; it includes 17,000-20,000 deaths per month. The Linked Birth and Infant Death Data Set, which also covers the 50 states, the District of Columbia and New York City, links the more detailed information from the birth certificate with the information from the death certificate for each of the approximately 40,000 infants who dies before his/her first birthday.

Population Based Surveys

Three of the Center's data systems are classified as population based surveys. They are the National Health Interview Survey, National Health and Nutrition Examination Survey, and the National Survey of Family Growth. The designs of these surveys are based on stratified multistage samples of households, where the household is defined as the basic sample unit. Based on established criteria, a person (one or more) in the sample household is selected as the ultimate sample unit, i.e., the unit of analysis.

National Health Interview Survey (NHIS)

The ***NHIS*** is a continuing nationwide sample survey in which data are collected on the incidence of acute illness and injuries, the prevalence of chronic conditions and impairments, the extent of disability, the utilization of health care services, and other health related topics. Generally, personal interviews are completed in 47,000 households for about 123,000 sample persons.

National Health and Nutrition Examination Survey (NHANES)

The ***NHANES*** obtains nationally representative information on the health and nutritional status of the American population through a combination of personal interviews (mostly in the respondent's home) and detailed physical examinations. These examinations are conducted in specially equipped mobile examination centers (MEC) that travel around the country. The last survey, NHANES III, the sixth in the cycle of health examination surveys conducted since 1960 [4], collected data on topics such as high blood pressure, blood cholesterol, infectious diseases, diabetes, HIV infection, blood lead levels, allergies, osteoporosis, and other nutritional status measures.

The NHANES III [5], conducted over two 3 year phases, 1988-91 and 1991-94, covered the U.S. civilian, noninstitutional population aged 2 months and older. Each phase constituted a national sample of about 20,000 persons, with an expected interview completion rate of 85-90 percent and a response rate of about 75-80 percent for the medical examination. More than 78 percent of the persons selected for the 1988-91 phase participated in the medical examination. Selected subpopulations, children (< 5 years), older persons (60+), Black Americans and Mexican Americans, were oversampled.

National Survey of Family Growth (NSFG)

The Center's third population based survey, the ***NSFG***, is a periodic nationally representative household survey of women of reproductive age (15-44 years). The survey, first conducted in 1973 [6], collects data on fertility and infertility, family planning, and related aspects of maternal and infant health. The 1988 survey, the fourth in the cycle [7], selected 10,000 eligible sample households from the frame of households that participated in the NHIS between 1985 through 1987. A total of 8,450 women were interviewed in person, in their own homes, by trained female interviewers.

Provider Based Surveys

Seven NCHS data systems form the family of provider based surveys, collectively called the National Health Care Survey (NHCS). Included here are the National Hospital Discharge Survey(NHDS), National Survey of Ambulatory Surgery (NSAS), National Ambulatory Medical Care Survey (NAMCS), National Hospital Ambulatory Medical Care Survey (NHAMCS), National Nursing Home Survey (NNHS), National Home and Hospice Care Survey (NHHCS), and the National Health Provider Inventory (NHPI). Whereas population based surveys use the household as the basic sample unit, provider based surveys use the medical provider (physician, hospital, nursing home, etc.) as the basic sample unit. The provider furnishes information on samples of provider/patient contacts, e.g., office visits, hospital stays, nursing home stays, etc.

Samples for these surveys range in size from the approximately 475 emergency rooms in the NHAMCS to the 87,000 facilities covered by the NHPI.

Followup/Followback Surveys

Six of the NCHS data systems included in this report are classified as Followup/Followback surveys. They are:

National Maternal and Infant Health Survey (NMIHS)
1991 Longitudinal Followup (LF) to the National Maternal and Infant Health Survey
National Mortality Followback Survey (NMFS)
National Health and Nutrition Examination Survey Epidemiologic Followup Study (NHEFS)
Longitudinal Study of Aging (LSOA)
National Nursing Home Survey Followup (NNHSF).

Sample sizes range from 7,500 to 26,000 persons.

Data Processing Environment and Dispersion of the Work

There are many similarities in the data processing activities employed by NCHS offices for their respective data systems. This is especially true for data systems within the same "family of surveys." For example, registration areas provide NCHS with coded and edited computer tapes or microfilm copies of vital event certificates which are converted to uniform codes and subjected to machine edits. The other surveys use CAPI (Computer Assisted Personal Interview), preliminary hand edits, machine edits, etc. There are, however, a number of procedures that cross "family survey" lines that are gaining greater usage with the rapid advances made in survey technology. Of particular interest to NCHS is "source point data editing" (SPDE). This refers to editing survey data by any means of access to either the interviewer (or other data collector), the respondent, or records within a limited time following the original interview or data collection. The time limit reflects the period within which the persons involved can reasonably be expected to remember details of the specific interview or, in the case of data collected from records, a time within which there is reasonable expectation that there has been no change to the records which would affect the data collected. Thus, data completion and accuracy are much more likely to result when source point data editing is used.

Audit Trail - This term refers to a process of maintaining, either by paper or electronically, an accounting of all changes of sample or survey data item values and the reasons for those changes. The level of effort varies by data systems; some are manual, while others are automated.

Micro-, Macro-, and Statistical Editing

This section describes three types of editing processes. The following definitions are used in this section.

❒ **Micro-editing**.--Editing done at the record or questionnaire level.

❒ **Macro-editing**.--Editing to detect individual errors by checking on aggregated data or by applying checks to the complete set of records.

❒ **Statistical editing**.--Editing based on statistical analysis of respondent data. It may incorporate cross-record checks, as well as historical data.

Micro-, macro, and statistical editing for the eight registration systems are all very similar. Automated edits are designed to (1) assure code validity for each variable and (2) verify codes or code combinations which are considered either impossible or unlikely occurrences.

For each of the other three types of data systems, most or all of the following procedures are used:

- ❒ Extensive machine micro-editing.
- ❒ Where appropriate, comparison of current estimates with previous years.
- ❒ Assuring reasonableness of record counts, sampling rates, etc.
- ❒ Checking ranges, skip patterns, consistency of data from different sources.
- ❒ Checking medical data for compatibility with age and/or sex.

Priority of Edits

None of the registration systems gives special priority to any item in the editing procedures. The other data systems prioritize their edits based on:

- ❒ Identifiers needed to link data files.
- ❒ Questionnaire items used to weight sample data to national estimates.
- ❒ Medical data incompatible with demographic data.

Imputation Procedures

Imputation is defined as a process for entering a value for a specific data item where the response is missing or unusable.

Registration Systems

Except for Abortion Registration, which does not impute for missing items, imputation procedures among registration systems apply primarily to demographic items. In Mortality registration, imputation procedures are done by machine, which checks for invalid codes. The following variables are subject to imputation procedures: age, sex, date of death, marital status of decedent, race of decedent, and education of decedent.

Missing natality data that are imputed include child's race, sex, date of birth, and plurality. Data imputed for the mother include race, age, marital status and residence. Imputation is done by machine, either on the basis of a previous record with similar information for other items on the record (e.g., mother's age imputed on the basis of a previous record with the same race and total-birth order), or on the basis of other information on the certificate (e.g., marital status on the basis of mother's and father's names, or lack of name). The tape documentation includes flags to indicate when imputation was performed.

Finally, marriage and divorce data imputation are limited to month of marriage (or divorce) and age of bride and/or groom (marriage only). Hot deck and cold deck imputation procedures are used. In hot deck imputation, a missing data item is assigned the value from a preceding record in the same survey having similar (defined) characteristics. In cold deck imputation, a missing data item is assigned the value from a similar record in a previous similar survey.

Population Based Surveys

Imputation procedures for the Center's other surveys differ from those used by the Registration Systems. In the case of the NHIS, unit nonresponse (missing sample cases) is imputed by inflating the sample case weight by the reciprocal of the response rate at the final stage of sample selection, or by a poststratification adjustment based on independent estimates of the population size in 60 age-race-sex categories.

Item non-response (missing question answers) is imputed, where possible, by inferring a certain or probable value from existing information for the respondent. For example, in the NHIS, a missing "husband's age" (or "date of birth") is assigned the value of "wife's age + 2 years."

In the NHANES, the calculation of sample weights addresses the unit nonresponse aspects of the survey except for special cases.

In the NSFG, the sample weights adjust for unit nonresponse. Imputation of missing items in the NSFG was carried out by the contractor. For the most part, a hot-deck procedure was used to impute missing values.

Provider Based Surveys

The provider based surveys have established imputation procedures for three types of nonresponse: unit nonresponse, record nonresponse, and item nonresponse. Unit nonresponse is imputed by inflating the sample weight of similar responding units. Record nonresponse is imputed by inflating the sample weight of similar responding cases to account for the missing cases. Item nonresponse is imputed by inferring a certain or probable value from existing respondent information.

Followup/Followback Surveys

Four of the followback surveys, the LSOA, the NMFS, NHEFS, and the NNHSF did not impute any data, although missing data items were filled in by using logical relationships as described in the above example. Unknown or inconsistent data were coded as "unknown." The other two used "hot deck" procedures.

Editing and Imputation Standards

For each of its registration systems, NCHS monitors the quality of demographic and medical data on tapes received from the states by independent verification of a sample of records of data entry errors. In addition, there is verification of coding at the state level before NCHS receives the data. All other systems employ error tolerance standards established for interviewer performance (if applicable), and enforced by editing and telephone reinterviews. Error tolerance standards are also established for coding and keying of data, and are enforced by sample verification.

Costs of Editing

The costs of editing are very difficult to determine, though some surveys and data systems appear to have a better handle on this than others.

None of the eight registration systems could provide an estimate of their editing costs. All other systems estimated their data editing costs between 10-30 percent of total survey costs.

Role of Subject Matter Specialists

For all surveys and data systems, the primary role of subject matter specialists is to write edit specifications, from which edit programs are prepared; to review results of edit runs and to adjudicate failures in collaboration with programmers. Their secondary role is to compare standard sets of estimates with historical series to identify anomalies. In addition, they also consult with survey design staff on field edits.

Measures of Variation

- No sampling error for 100 percent registration systems; however estimates of variation are computed for vital events <20.
- Marriage and divorce data tables list sampling errors by area expressed as a percent of the area total.
- Other surveys produce estimates of sampling (but not non-sampling) errors.
    - Selected surveys present estimates based on assumptions regarding the probability distribution of the sampling error.

Current and Future Research

There are several ongoing research activities and a number of others are being considered for the future. Resource constraints, both money and personnel, are the major limiting factors. The following represent programmatic changes in data collection, data processing/editing, data analysis, etc., that will occur or be investigated in future years. Aside from these specifics, however, perhaps the biggest change, one which is well underway now, and cuts across all surveys, is the shift from paper and pencil data collection to computerized data collection. This shift makes it more difficult to omit data items, to enter inconsistent or impossible data, etc.

- Implementation of electronic birth and death certificates by the states.

❐ Implementation of the Super ***MICAR*** (**M**ortality **M**edical **I**ndexing, **C**lassification, and **R**etrieval) system by all states. The intent of Super MICAR is to allow data entry operators to enter cause-of-death information as it is literally reported on the death certificate. Under the current MICAR system, cause of death information must be entered using abbreviations or standardized nonmeclature (Harris et al., 1993). Implementation of Super MICAR is essential to a successful electronic death certificate system.

❐ Determination of optimum imputation techniques (single and multiple procedures) and their applicability to NHANES.

❐ Evaluation of automated data collection methodology for the NHDS.

❐ Feasibility of developing an automated system for coding and classifying medical entities using the ICD-10.

❐ Feasibility of developing an automated system for data correction and creation of an audit trail (Followback surveys).

Summary

Data editing practices at NCHS are quite extensive. Unfortunately, detailed descriptions of these practices for this report were precluded because of space limitations. However, some summary findings on NCHS data editing practices are provided below and in Table 2.

❐ Among NCHS data systems, the cost of data editing is the least documented variable. Only five data systems provided dollar and other resource costs of their data editing procedures. Most of the other data systems offered "guestimates" of 10-20 percent of total survey costs, with a few "guestimating" as high as 30 percent of total survey costs.

❐ About sixty percent of the Center's data systems collect data on an on-going basis throughout the year and publish data on an annual basis.

❐ Two-thirds of the Center's data systems report item non-response rates under 5 percent.

❐ One third of the Center's data systems use Computer-assisted telephone interviewing (CATI) as their primary or secondary data collection method.

❐ One-half of the Center's data systems release micro-data with identifiable imputed data items; another one-quarter release micro-data without identifying imputed data items.

❐ Virtually all NCHS data systems have rules establishing minimum standards of reliability that must be met in order to disseminate data.

❐ Slightly more than one-third of the Center's data systems monitor analysts/clerks in their data editing procedures; three-fourths monitor their automated editing procedures. However, only three data systems formally evaluate their data editing systems.

Table 2.--Frequency of Selected Data Editing Practices Among NCHS Data Systems

	Yes	No	NA/DK
1. Data dissemination limited by confidentiality (privacy) restrictions?	24		
2. Does data system release microdata (respondent level data)?	19	5	
3. Are imputed items identified?	13	11	
4. For aggregated data, is information provided on percentage of a particular item which has been imputed?	5	16	3
5. Are there minimum standards for reliability of disseminated data?	21	1	2
6. Is information available on the cost of data editing?	5	19	
7. Are there procedures for monitoring editors, clerks, analysts, etc.?	9	14	1
8. Are there procedures for monitoring automated editing procedures?	18	4	2
9. Is there an audit trail (i.e., a record kept) for some or all data editing transaction?	21	2	1
10. Are performance statistics maintained in order to evaluate the data editing system?	3	21	
11. Has any analysis been done on the effect of data editing on estimates produced?	5	19	
12. Is survey data editing information released?	15	9	
13. Is validation editing performed?	22	2	
14. Is macro-editing used?	17	7	
15. Are any other data editing techniques performed?	6	17	1

Reference

Harris, Kenneth W.; Rosenberg, Harry, et al. (1993). Evaluation of an Automated Multiple Cause of Death Coding System, *Proceedings of the American Statistical Association*, Social Statistics Section, Washington, DC.

Footnotes

[1] Statistical Policy Working Paper 18: *Data Editing in Federal Statistical Agencies*, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, May 1990.

[2] National Center for Health Statistics: Data Systems of the National Center for Health Statistics, *Vital and Health Statistics*, Series 1-No. 16, Hyattsville, MD, December 1981.

[3] National Center for Health Statistics: Data Systems of the National Center for Health Statistics, *Vital and Health Statistics*, Series 1-No. 23, Hyattsville, MD, March 1989.

[4] National Center for Health Statistics: Cycle I of the Health Examination Survey, Sample and Response, United States, 1960-62, *Vital and Health Statistics*, Series 11-No. 1, Rockville, MD, May 1965.

[5] National Center for Health Statistics: Sample Design: Third National Health and Nutrition Examination Survey, *Vital and Health Statistics*, Series 2-No. 113, Hyattsville, MD, September 1992.

[6] National Center for Health Statistics: National Survey of Family Growth, Cycle I, *Vital and Health Statistics*, Series 2-No. 76, Rockville, MD, June 1977.

[7] National Center for Health Statistics: National Survey of Family Growth, Cycle IV, Evaluation of Linked Design,*Vital and Health Statistics*, Series 2-No. 117, Hyattsville, MD, July 1993. ■