Paper 120-26

Integrating SAS and OMR Data Scanning Technology to Improve Efficiency with Database Management

Peter H. Coulson, Portland State University, Child Welfare Partnership – Research Unit

ABSTRACT

Administrators rely on management information systems when making important decisions. The management information system may not contain all the data elements necessary to make informed decisions. When the management information system does not contain the necessary data, research projects are conducted to complement existing data. These research projects can be labor-intensive and time consuming.

Combining Optical Mark Recognition (OMR) technology with SAS can expedite the process. OMR technology is the scanning of forms with darkening dots. OMR technology enables researchers to customize the information gathering forms, develop multi-page booklets, and program the scanner to recognize obvious errors. OMR and SAS can provide more accurate data without keypunching and with minimal time devoted to database creation. Automating database creation and expediting data collection can provide more accurate data with less labor.

INTRODUCTION

Data collection can be a time consuming and labor intensive process when important data is not provided through a management information system. Many social service agencies have management information systems containing financial data; most client data is not in the management information system but contained in case records. State social service agencies often conduct research to assess service effectiveness, generate client profiles, and recognize differences in clients served in numerous geographic locations. Data must be gleaned from case records to provide pertinent information to program managers and administrators.

The Oregon child welfare agency began researching family characteristics and their association with important outcomes in the late 1980's. Most data was only available in case records. The data collected from case records complemented the state's management information system and provided administrators with necessary information. Initially the study included 100-200 variables and 500 observations; half the process involved data collection and half the process involved steps necessary to create databases for subsequent analysis.

When the project was first conducted in the late 1980's, cases were reviewed, data was coded, forms were edited, data was key punched, flat files were compared, and eventually data sets were created. The research was labor intensive and time consuming; the agency administration desired more information with less time devoted to the data collection and data processing. Eventually more variables were requested, more information was requested, and more observations were necessary. The data collection system that required coding, editing, and keypunching followed by labor intensive data set development was inadequate.

The optical scanning technology and database management techniques discussed below created an efficient and reliable system of data collection and database development. Combining optical scanning with SAS efficiencies enabled researchers to incorporate more information and more observations into the data collection process. Without these efficiencies, the research would not be available to administrators and managers when needed.

There are five steps in the current process involving SAS and optical scanning technology. The five steps are: importing of external files, creating a statistical random sample, preparing forms for data collection, programming the optical scanner, and developing the SAS data sets. Since this research is conducted each biennium, development of the data collection device is not included in this process. Each step is described below.

Importing an external data file

The management information system from Oregon's child welfare agency provides identifying information for each family and used for a variety of purposes. From these identification variables, cases are identified for review by the case reviewer, random samples are generated, and are developed lists for branch offices.

The external file from the social service agency is imported and variables are labeled. Labels are necessary to recognize the coding system employed by the social service agency. The following are SAS input statements for importing external files, attaching SAS labels and SAS formats.

```
PROC DBLOAD DBMS=EXCEL
DATA=MASTERC5.C5SCMAST;
PATH='D:\Data\EXCEL File';
PUTNAMES YES;
LIMIT=0;
LOAD;
RUN;
```

Labeling Variables example:

```
Label
CASENUM = 'Case Number'
PL = 'Person Letter'
BRANCH = 'SCF Branch';
Run;
```

Formating variables example:

```
proc format library=library;
value Branch
1='Baker'
2='Benton'
3='Clackamas';
run;
```

CREATING A STRATIFIED RANDOM SAMPLE

The imported file includes a listing of families who are provided services through the social service agency. The data are used for numerous purposes including comparisons among families served by different branch offices. To ensure an adequate number of observations for all branch offices and all client populations of interest, a stratified random sample is necessary. The following table illustrates the disproportionate number of observations drawn from each branch office. These tables are created by statisticians to reflect the research objectives and the resources available.

	A	B	C
1	BRANCH	Possible	Want
2	1	8	8
3	20	230	41
4	7	9	9
5	8	8	8
6	42	200	34

A SAS data set is created from the Excel spreadsheet. The SAS SQL and the random number generator (RANUNI) generate the sample with the following SAS code.

```
proc SQL;
create table work.new as select *,count(*) as
count, ranuni(1) as RANDNUMB from work.temp3
group by branch, c5Cohort order by branch,
c5Cohort,RANDNUMB;

data RANDNUMB;
set work.new;
by branch c5Cohort;
if first.c5Cohort then do ;
ct=0;
retain ct;
end;
ct+1;
if round(ct,1) le round(want,1) then output;
run;
```

The previous two steps – importing an external file and generating a random sample – are necessary for case reading projects conducted by social service agencies. Although SAS labels, SAS formats, SAS SQL, and SAS RANUNI improve efficiency, the greatest improvements occur with the next three steps.

INTEGRATING OMR TECHNOLOGY

Previous research relied on archaic technology. Case reviewers were provided forms containing columns of numbers. Each column would be a category that profiled a particular family characteristic (e.g. problems exhibited by the caretakers, services provided to the families). Each number within a column provided more specific information within a category – for caretaker problems, number "1" might indicate an unemployed caretaker, a "2" might indicate a recent birth, and so on. The family profile was generated from a group of numbers that were later keypunched and analyzed. The process was difficult and especially time consuming with larger research projects.

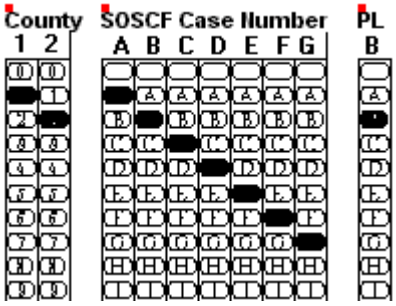
Incorporating Optical Mark Recognition (OMR) technology into the data collection process has improved both efficiency and accuracy. Case readers now darken dots on an optical scanning form to recognize family characteristics. OMR scanners, also known as data scanners, easily can scan 750-1000 pages per hour; data collection devices can be 1 page or many pages. Our data collection booklet, containing over 20 pages and 2000 variables, requires less than one minute to scan.

The errors associated with completing forms and keypunching have been minimized with OMR technology. Data sets involving over 2000 variables and over 5000 observations can be created in less than one-quarter the time previously required. OMR technology is fast, accurate, and cost effective.

OMR technologies can certainly improve data collection and the data set creation process. The OMR technology can also be combined with other software to further improve research processes. Creating custom data collection devices and "preslugging" can also improve efficiency and accuracy. Forms can be created by individuals collecting data or by professionals accustomed to the OMR software (Bubble Publishing® Suite). The format, the variable's location on the form, single or double-sided pages can all be determined by individuals reviewing case records. The flexibility of the OMR software enables researchers to customize forms and improve data collection.

Most labor-intensive research complements data collected in the management information system. Identifying variables (e.g. case number and county) from the management information system are often included on the information gathering form. Individuals collecting data essentially transcribe information from a list to the data collection form. Preslugging can eliminate the transcription process. The printed forms created using OMR software darkens the dots that would normally be transcribed by data collectors. The preslugging eliminates the unnecessary transcribing, eliminates errors, and provides information to data collectors. With our current form, preslugging occurs on the first page – the other 20 pages in the data collection form are created at the university and the booklets are reproduced by professional printers. The preslugged information is not contained in the booklets but rather contained on face sheets.

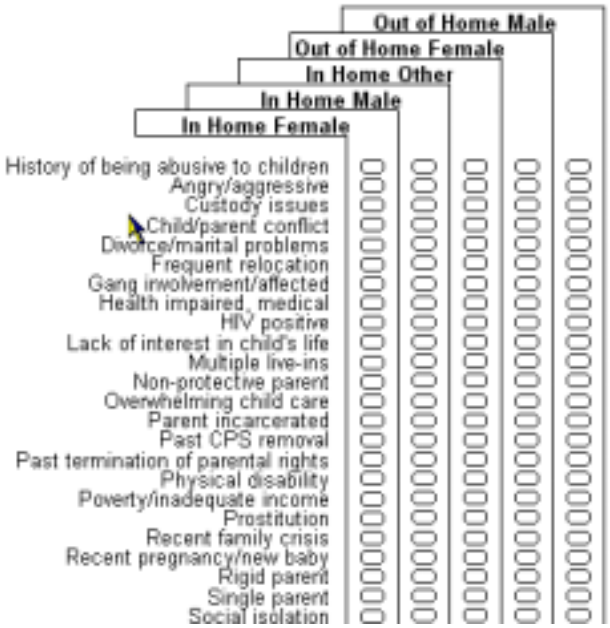
The example below illustrates preslugged variables:



THE OMR SCANNING PROCESS

OMR scanning is a simple process after the programming is complete. The data collection forms are stacked on the scanner, the scanner is interfaced with the personal computer and scanning begins. Data collection forms can be one page, multiple pages, single-sided pages, or double-sided pages. There are no limits to the number of variables per page. The completed data collection forms are converted into a flat file using the same software used to create the scanning forms (Bubble Publishing® Suite). Most programming associated with transferring data from the forms to the flat file occurs while the data is being collected. As forms are completed, data is scanned and files are created; as more forms are completed, the new data is scanned and added to the existing flat file. The speed of the scanner is impressive. A 20 page booklet with over 2000 variables requires less than one minute to scan. Organizing forms, removing perforated strips, and the loading of forms onto the scanner requires more time than the scanning itself. Although difficulties can arise, scanner problems are not common.

Below are two examples of variables collected using OMR technology. The first example contains 5 mutually exclusive choices; if more than one dot is darkened, the scanner will recognize the incompatibility and places a asterik (*) noting that that field was incorrectly bubbled. The second example illustrates the large volumes of data that can be easily and accurately collected using OMR technology. Data collectors become accustomed to the form and can efficiently complete forms.



DATASET CREATION

The typical research project involves formulating a question, designing an information gathering form, sampling, data collection, data set development, statistical analysis, and report writing. The data collection process generally requires more time than the other processes. When the data collection is completed, managers and administrators are eager for the results. Fortunately, converting OMR flat files into SAS data sets is no different than with other flat files. Often the SAS input statements are completed before the scanning is completed. Much of our SAS input code is generated using Excel spreadsheets. The scanning software creates flat files; these flat files can be converted directly into SAS data sets or converted into databases. The database capability incorporated into the scanning software contains numerous beneficial features. Transforming optically scanned forms into SAS data sets can often use these inherent capabilities; using Excel, in conjunction with the scanning software, provides numerous options for individuals creating SAS data sets. Preprogramming both the translation of OMR data into a flat file and converting the flat file into SAS data sets expedites the process. Since the data collection step is the most time consuming step in the process, OMR and SAS code can be generated while data is being collected.

CONCLUSIONS

The first three steps involved with integrating OMR technology into creating SAS data sets occur before data collection begins. Importing external files and creating a random sample are necessary for most research. Some minor efficiencies are realized by using DBLOAD, SQL, and the SAS random number generator. Major efficiencies are realized using OMR technology. The OMR technology enables custom designs of data collection forms and preslugging of identifier variables. These steps, all occurring before the data collection effort is initiated, can greatly improve the efficiency and accuracy involved with data collection. The final two steps – OMR scanning and SAS data set creation – provide the most notable efficiencies when integrating OMR and SAS technologies.

This system of integrating OMR and SAS technologies can expedite research processes. If managers and administrators can access needed information, the information will be utilized. Integrating OMR and SAS technologies minimizes time required between data collection and the reporting of results; this makes the information more accessible and more likely to be used by the administration.

ACKNOWLEDGMENTS

Scanning Dynamics Incorporated (800 493-9590) has modified their software to improve our data collection and database management system. The efforts of Ron Schlangen, Phil Jenkins, and Harold Lundberg are greatly appreciated.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Peter H. Coulson
Company: Child Welfare Partnership. PSU
Address: 4061 Winema Place NE
City state ZIP: Salem, Oregon 97305
Work Phone: 503-315-4266
Fax: 503-315-4280
Email: coup@chemeketa.edu
Web: www.cwp.pdx.edu