Benefits and Limitations of Using Only Administrative Data to Update Current Business Surveys with New Employer Births

James N. Burton, Carol S. King, James W. Hunt, U.S. Department of Commerce, Bureau of the Census, Washington, D.C. 20233. [1]

Keywords: administrative data; classification; business surveys; two-phase sampling

1. Introduction

In conjunction with each Economic Census the Census Bureau selects new samples for its current monthly and annual retail, wholesale, and service surveys. The frame for these samples is the Census Bureau's Business Register. On a regular basis, the Business Register is updated with administrative data, which includes information on new employer births. Konschnik, Monsour, and Detlefsen (1985) provide a discussion of how new samples are selected while both Walker (1997) and Konschnik and Walker (1999) discuss the details of how the Business Register is structured and updated.

To represent these new employer births, or simply births, in our current surveys we use a two-phase sample design. In the first phase, we identify those births on the Business Register not yet represented in our samples. We subject these births to sampling and mail the selected sample for more complete industry classification, measure of size, company affiliation, and survey specific information. In the second phase, we use the response data together with the most recent administrative information to select a subsample from the sample in the first phase. The newly selected births are then added to the current samples. This two-phase operation is done quarterly. Konschnik, Monsour, and Detlefsen (1985) provide details on both how births are identified from the Business Register and the two-phase sample design.

We use a two-phase design because historically administrative data provided only incomplete information on births. This was particularly true for industry classification. Recently, we have seen evidence that the amount of industry classification from administrative sources has increased. As a result, we decided to evaluate whether we could eliminate the first phase sample and subsequent data collection activities. Elimination of these data collection activities would provide a cost savings as well as allow us to represent births as much as three months, on average, earlier in our samples.

In this paper, we discuss the data required for birth sampling, their sources, and their use. Next we discuss the nature of the industry classification changes made by the Social Security Administration (SSA) and describe how these changes led us to reexamine our two-phase design. We then investigate the possibility of using the administrative data as a replacement for data collected in the first phase. We end with our conclusions and areas for further research.

1.1 Industry Classification

Throughout this paper, references will be made to the industry classification. Currently, this is on a 1997 North American Industry Classification System (NAICS) basis. The 1997 NAICS code is a six digit code, with the first two digits indicating a broad industry sector level and additional digits indicating more detailed industry classification.

Due to the need to maintain a previous sample on a Standard Industrial Classification (SIC) basis while introducing a new sample on a 1997 NAICS basis, the Census Bureau currently assigns industry classification codes to first phase selected births on a 1997 Standard Industrial Classification (SIC) bridge code basis.

1.2 The Employer Identification Number (EIN)

Both the administrative and sampling unit referred to in this paper is the Employer Identification Number (EIN). The EIN, which is assigned by the Internal Revenue Service (IRS), is the primary taxpayer identifier used by employer business firms. Under the Federal Insurance Contributions Act (FICA), every organization with paid employees must have an EIN. The Business Register receives new EINs and updates to existing EINs as described in Walker (1997). Because of this relationship, the EIN is the sampling unit for the first and second phase operations. The term 'births' will be used for these sampling units in the remainder of this paper. For

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

clarification, note that an employer firm may have as few as one EIN for all of its payroll operating establishments, or as many EINs as it has establishments, as the firm decides. However, 94% of all active EINs with paid employees have a one to one relationship between the EIN and the establishment (Walker, 1997).

Note that administrative and first phase industry classification codes are assigned to EINs, not to the individual establishments. Thus, for EINs having multiple establishments, the industry assigned to the EIN is the principal (in terms of receipts) industry. This is also true for any other qualitative information assigned to an EIN. Examples of this include participation in electronic commerce (e-commerce) or whether an EIN's service establishments are taxable or tax-exempt.

2. Reason for Current Two Phase Design

As stated in the introduction, the reason for the current two phase design is that certain information is needed to accurately represent births in the current surveys. Historically, either the required information was not available from administrative records or, when available, was either incomplete or not timely enough.

To accurately represent births in the current surveys we require:

- information for sampling such as industry classification, measure of size, whether a wholesale birth is merchant or non-merchant, and whether a service birth is taxable or tax exempt;

- information to prevent duplication such as company affiliation, and

- indication of whether a birth is participating in e-commerce or not for our e-commerce surveys.

In general, we consider a birth to be eligible for first phase if the industry classification of the birth is inscope to our current surveys (or has no classification and meets other criteria) and has an indication of operating with paid employees, either by having employees or by reporting non-zero quarterly payroll to the IRS. See Konschnik, Monsour, and Detlefsen (1985) for more details.

As examples of the information we collect in the first phase to meet these requirements for industry classification, we collect trade, description of firm's business or profession, principal lines of merchandise, and selling characteristics. For measure of size, we collect the latest two months of receipts as well as end of month inventories for wholesale.

The following table shows what administrative information is available for births that are considered eligible for first phase:

Table 1: Required Birth Items

Required Item	**Administrative Data Source Available at Time of Eligibility**
Industry Classification	SSA Industry Classification
Measure of Size	Employment as of March 12th and quarterly payroll from IRS. Beginning in 2002, estimated annual employment will be available from SSA.
Wholesale Inventories	Beginning of year and end of year inventories from the IRS.
Company Affiliation	None
Whether Merchant or Non Merchant	None
Whether Service Taxable or Tax Exempt	Legal Form of Organization from the IRS
Participating in E-commerce	None

For items where the administrative data were available, historically we have had the following problems. For industry classification from the SSA, this has been incomplete. The following table shows the percent of births with administrative classification at the time of first phase identification for the year 1998. Other years prior to 1998 show comparable rates.

Table 2: 1998 Classification Rates for Births Identified for First Phase

Quarter	**Percent Classified**
January 1998	17.5%
April 1998	26.4%
July 1998	22.4%
October 1998	38.6%

Source: Johnson, T. (1999)

For the measure of size, we have generally considered the administrative payroll or employment to be inferior to reported monthly receipts. The availability of administrative inventory is relatively recent, and work (see further research) remains to be done on this item.

The incompleteness of the industry classification, together with the unavailability of other data items and some concerns over the accuracy of the measures of size estimated using administrative payroll or employment are the reasons why the Census Bureau has maintained the two phase design. The next section discusses our decision to reconsider the use of a two phase design.

3. SSA Improvements

In the beginning of 1999, the SSA made two changes to its process for industry coding from form SS-4, "Application for Employer Identification Number". These changes were the result of a working arrangement between the SSA and the Census Bureau to attempt to improve both the timeliness and accuracy of the industry coding.

An SS-4 asks for a written description of the principal activity of the businesses operating under the EIN, the principal product and raw material used if the principal activity is manufacturing, and whether the purchaser of the goods or service sold is the public or a business.

The first change was the introduction of an automated coding system. In January of 1999, SSA began using an automated coding system known as the Coding Assist and Rulings Request System, or CARRS, to assign 1997 NAICS codes based on industry classification information available from the SS-4. CARRS was originally developed by Statistics Canada to aid their business classification efforts. It combines the information available in a classification manual with a search engine.

CARRS allows users to access a list of 1997 NAICS codes which match the description of a business' major activity. If the description leads to multiple possible classifications, the user is presented a list of possible codes and descriptions.

The second change was twofold. First, the Census Bureau and SSA agreed not to code a large backlog of SS-4s. Second, beginning with the first NAICS-based March 1999 SS-4 extract, the Census Bureau agreed to receive less information from the SS-4. The number of items received dropped to 6 from 12. See Custard (2000).

The following table shows the impact of these changes on rate of industry classification. These tables are based on quality assurance of incoming SS-4 information as it is posted to the Business Register.

Table 3: Amount of industry classification available from SS-4s posted to the Business Register, January of each year

Date	Percent Classified	Date	Percent Classified
January 1994	65.7%	January 1999	81.6%
January 1995	69.5%	January 2000	95.1%
January 1996	69.9%	January 2001	96.5%
January 1997	69.3%		

No data available for January, 1998

The increase in the classification rates in the past three years led us to question the need for two phases of sampling.

In the following section, we address issues related to the elimination of one phase of sampling. This includes the following issues: the quality of the administrative industry classification, the administrative data available for other items collected by the first phase, and the impact of not collecting the data for which we have no administrative sources.

4. Eliminating the First Phase

This section addresses how eliminating the first phase would effect the quality of the data for representing births in our samples. To do this, we first address those items where we have administrative information at the time of first phase - the industry classification, the measure of size, and whether a service birth is taxable or tax exempt. Only the receipt measure of size is addressed as no work has yet been done using the administrative inventories. For the items without administrative sources, a summary of the amount of response data lost as a result of eliminating the data collection is provided.

4.1 Industry Classification

Of all the data required to correctly sample and tabulate births in the current surveys, the most important item is the industry classification. Without accurate industry classification, births will be sampled with incorrect weights and tabulated in incorrect industries.

As one of the main motivators for this study was the increase in the amount of SSA industry classification, we begin with an investigation of those rates and follow with some comparisons between the administrative industry classification and the first phase response classification.

First, we investigated a difference we were seeing between the amount of industry classification available for births identified in the first phase and the SS-4 quality assurance reports. The following table shows the rates of industry classification for births identified for the first phase and SS-4s posted to the Business Register for the same period. The first phase rates are from Johnson (2000) and the SS-4 rates are from internal Census quality assurance reports. Other periods provide similar results.

Table 4: Comparison of Industry Classification Rates for 2000

	Births Identified for the First Phase	SS-4s Posted to Business Register
First Quarter	54%	94%
Second Quarter	33%	95%
Third Quarter	57%	95%
Fourth Quarter	41%	96%

There are two reasons why industry classification rates from the internal SS-4 quality assurance reports do not match the rates of those births that meet our requirements for first phase. First, we found that approximately 30% of all EINs with a unique establishment on the Business Register with current year payroll had not received an SS-4 update. Next, the lag between the time the Business Register receives a new birth EIN and the time that EIN receives an SS-4 update is from 2 to 3 months. Thus many births can be eligible for first phase based on the presence of administrative payroll but not have received industry classification from an SS-4. In essence, the two rates measure different things.

Second, we compared the administrative industry classification available at the time of first phase to the response industry classification from the first phase data collection. The results are show in table 5, below.

On table 5, the left column denotes the administrative classification available, for instance a 6-digit code provides a more detailed industry classification than a 2-digit code. The next column indicates the percent of administrative codes having that amount of industry classification. The remaining columns indicate the agreement between the administrative industry classification and the first phase response industry classification, with 'Different Inscope Trade Areas' indicating the worst possible result for us (based on only the administrative industry classification we would have selected a birth into our surveys that was not in our coverage) to 'Same Inscope Trade Area, 6-Digit Agreement' representing the best possible result (full agreement between the Administrative and the first phase response industry classification).

For us to reasonably use only the administrative industry classification, we would need a high level of agreement at the 4-, 5- and 6-digit levels, as these are the levels of our current sampling and tabulation. Here, we see that only 36.8% of the industry codes fall into this group.

While all industry classification is a difficult task, based on only the information provided by a respondent, we have several reasons to believe that the industry classification of the first phase births is more reliable than that done by the SSA. First is the amount of information collected in the first phase. The first phase data collection process asks several 1997 NAICS based industry classification and product line questions. Second, the first phase industry classification process makes use of detailed clerical edit specifications, multiple follow-ups, and analysts review. Due to the volume of industry classification codes being assigned and other priorities, SSA cannot conduct such extensive processing. Finally, Konschnik (1993) demonstrated the Census Bureau process for assigning industry classification to be of good quality.

Table 5: Comparison of Administrative to Response Industry Classification, for Births Selected in First Phase, Second Quarter, 2000
In percent of number of births with both administrative and response industry classification

				Same Inscope Trade						
	Percent of Administrative Codes	Response Out of Scope	Different Inscope Trade Areas	No Agreement	1 Digit Agreement	2 Digit Agreement	3 Digit Agreement	4 Digit Agreement	5 Digit Agreement	6 Digit Agreement
2 Digit Admin	3.6	0.0	1.4	0.2	0.5	1.5				
3 Digit Admin	9.6	0.1	1.4	0.9	0.8	1.0	5.4			
4 Digit Admin	13.5	0.6	2.5	1.4	1.3	1.3	1.3	5.1		
5 Digit Admin	53.7	3.3	9.9	5.8	3.7	3.5	5.0	2.0	20.5	
6 Digit Admin	19.6	0.8	1.5	1.6	1.4	4.1	1.0	0.3	1.8	7.1
Total	100.0	4.8	16.7	10.0	7.6	11.5	12.7	7.4	22.3	7.1

Source: King and Moore, 2000

So, while the amount of administrative industry classification at the time of first phase has increased, we still do not believe that either the amount available or the quality is enough for us to discontinue our own industry classification.

4.2. Measure of Size

A birth's measure of size is used for both its stratification and as a measure for imputation for second phase selected births that do not report in the current surveys. To determine if the administrative data can provide an adequate measure of size, it is enough to look at how the birth responded in the current survey of interest and determine if the measure of size for the birth calculated from first phase collected monthly data or the measure of size calculated from the administrative data agreed more closely with data reported in the current surveys.

To investigate this, we compared the measures of size for births selected in the second phase sampling operations in the second, third, and fourth quarters of 2000 with data reported for those births in the 2000 annual surveys. We then calculated estimated correlations between the birth measures of size and the annual survey reported data for second phase selected births. This was done for measures of size computed from first phase collected monthly receipts and from administrative payroll.

The results are included on the following table. Only results for the trade areas (Retail, Service, and Wholesale) and selected 1997 NAICS two digit sectors are shown.

Table 6: Estimated correlation between reported receipts and estimated receipts for Births Selected into the 2000 Annual Surveys

	Number estimated with sales	Correlation estimate	Number estimated with payroll	Correlation estimate
Retail	672	.767	804	.912
44	402	.820	489	.911
45	161	.641	183	.625
72	109	.981	132	.990
Service	704	.145	1115	.612
48	33	.844	34	.881
51	100	.848	346	.496
62	417	.197	481	.875
Wholesale	97	.687	55	.797

The results are preliminary as more research is still needed to determine the effect of outliers, if any, on the estimates as well as a detailed NAICS level analysis. But the preliminary results are interesting in that, in general, administrative payroll seems to estimate the reported annual sales better than the collected receipts. The relationship of the birth measures of size to the reported monthly data remains to be investigated.

4.3 Service Birth Taxable or Tax Exempt Status

The tax status for any service birth can be obtained from the administrative Legal Form of Organization (LFO). The LFO is available at the time a birth is identified on the Business Register and any changes that IRS receives to the LFO are provided to the Business Register as part of the regular updates.

For response information obtained for 2nd quarter, 2000, first phase births, the disagreement between the response information and the LFO was .10% and for 3rd quarter, 2000, .22%. This was based on the assumption that non-reporting service births are taxable. This assumption is based on the fact that the majority of service sampling units are taxable.

Based on these observations, it would appear that the administrative LFO can adequately replace the response data.

4.4 Items without Administrative Sources

For the items where we do not have administrative sources, the amount of response data provided by the first phase mailout is a good indication of the amount of data we would lose. The following items are covered: company affiliation, whether a birth is merchant or non-merchant, and indication of e-commerce.

The following table shows the response from the first phase data collection operations for the third and fourth quarters of 2000.

Table 7: Response for Items with No Administrative Sources

Data Item	Third Quarter 2000		Fourth Quarter 2000	
	%	Count	%	Count
Company Affiliation	5.1%	461	7.6%	1,187
Non-Merchant Wholesale	0.2%	23	0.4%	66
E-Commerce	5.4%	490	7.5%	1,167

Source: King and Moore (2000)

Some clarifications are in order. For the company

affiliation percentages, we collect both if the birth is owned by another firm and if the birth owns another firm. All respondents reporting 'Yes' to either of these questions are included in the percentages. For the 'Non-Merchant Wholesale' line, we considered only responses that indicated that the birth was indeed non-merchant, since, as for the service tax question, the majority of births are merchant. The percent is over all births, not just the wholesale births. For the indication of e-commerce, we included those that responded 'Yes' to the E-commerce question.

In summary, the net result of not collecting these items is that the burden of work would be moved from the data collection to the Census Bureau analysts. For example, analysts would have to carefully review selected births to determine if they were owned or affiliated with another company or if they participated in e-commerce. The number of non-merchant wholesale births identified is almost trivial.

5. Conclusions and Further Research

It appears that if the administrative industry classification could be improved, we would have a good argument for eliminating the first phase of our two phase methodology. All other information we collect would seem to have an adequate administrative replacement or could be replaced with work by our analysts.

While problems with the administrative industry classification would seem to prevent us from simply eliminating the first phase of our process, there are still opportunities to reduce respondent burden by eliminating items from the data collection where we have a viable administrative replacement. For example, the LFO can be used in place of collecting the service tax status question. Further investigation of using the administrative payroll and employment may lead to the removal of the monthly receipts questions. Additionally, further investigation of the administrative beginning of year and end of year inventories may indicate that collection of this item is unnecessary. Furthermore, investigation of optional designs for adding births to current surveys could lead to better uses of our administrative data and data collection resources.

References

1. Custard, R. (1999), "Processing Changes for Social Security Administration (SSA) Business Birth Files," unpublished memorandum, Washington DC: U.S. Census Bureau, EPCD.

2. Johnson, T. (1999), "Quarterly Interface with SSEL," unpublished memorandum, Washington DC: U.S. Census Bureau, Service Sector Statistics Division. Supplements for each quarter through 2001.

3. King, C. and Moore, L. (2000), "Characteristic of Birth Canvassing Operations for Cases Subjected to Second Stage Sampling in Second Quarter 2000," unpublished memorandum, Washington DC: U.S. Census Bureau, Service Sector Statistics Division. Additional supplements covered third quarter 2000 and fourth quarter 2000.

4. Konschnik, C. (1993), "Classification of B-625 Birth Sampling Forms," unpublished memorandum, Washington DC: U.S. Census Bureau, Service Sector Statistics Division.

5. Konschnik, C., Monsour, N., and Detlefsen, R. (1985), "Constructing and Maintaining Frames and Samples for Business Surveys", *Proceeding of the Survey Research Methods Section*, Alexandria VA: American Statistical Association.

6. Konschnik, C., Walker, E., et al. (1999), "2002 Redesign Report of Administrative Records Team," unpublished memorandum, Washington DC: U.S. Census Bureau, Economic Processing and Coordination Division.

7. Walker, E. (1997), "The Census Bureau's Business Register: Basic Features and Quality Issues," *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association.