

Using Australian Business Activity Statement data to improve survey design, methodology and processes at the Australian Bureau of Statistics

Gemma Van Halderen, Statistical Services Branch, Australian Bureau of Statistics, P.O Box 10, Belconnen, ACT 2616, Australia, g.vanhalderen@abs.gov.au

The views expressed in this paper are those of the author, and do not necessarily reflect those of the ABS

ABSTRACT

The Australian Bureau of Statistics is exploring statistical uses of administrative data for improving survey design, estimation and processes. The explorations have challenged survey methodologists and official statisticians, not least because of the timeliness of the administrative data. Use of administrative data to improve survey design and methodologies is possible with less timely administrative data, but not so for uses in substitution or supplementation.

1. INTRODUCTION

The Australian Bureau of Statistics (ABS) assists and encourages informed decision making, research and discussion by providing a high quality, objective and responsive national statistical service. The use of administrative data sources enables the ABS to provide cost effective statistics, increase the availability and use of non-ABS data in statistical products, and reduce reporting load through substitution as well as efficient survey designs.

A major Australian government initiative commenced in July 2000 with the introduction of The New Taxation System. Businesses registered on an Australian Business Register report financial activities and pay various taxes on a monthly, quarterly or annual basis. Financial details of wages and salaries paid to employees, turnover, goods and services tax collected and paid, fringe benefits tax, company tax, and cost of capital acquisitions are among a core set of data collected on a business activity statement. The Australian Business Register and the data collected on the business activity statement form a rich source of data to improve our survey design, methodology and processes.

Coinciding with the introduction of The New Tax System, the ABS started to explore strategies for using the new tax data. The ABS had for many years been using business income tax data to improve our annual surveys, but the new data provided an opportunity to improve our subannual surveys and statistical outputs. Early potential uses of the new tax data included

- *improving stratification* through the availability of new auxiliary variables
- *using alternative estimation methodologies* made possible by the availability of a range of auxiliary variables
- *data substitution* for existing statistical products
- *data supplementation* to enhance our current range of statistical products, especially at finer levels of disaggregation
- *improved survey processes*, such as editing, imputation and data confrontation
- *expanding* our range of statistical products and data items
- *in sampling frames maintenance*, for example to trigger removal of units that are no longer operating.

Assessment of the quality of the new tax data was a key element of early investigations. Over the last few years, the ABS has been embracing a quality framework into its survey processes and statistical outputs (Brackstone 1999; Lee and Allen, 2001). At the heart of the framework are six attributes - relevance, coherence, accuracy, timeliness, accessibility, and interpretability. Using this framework, a 'fitness for use' assessment of new tax data can be undertaken.

This paper has been prepared for the International Conference on Improving Surveys 2002. The paper presents details of early investigations by the ABS into two uses of the new tax data. Improving stratification of subannual and annual business surveys by updating our existing size benchmark demonstrated sampling efficiencies could be achieved. An evaluation of generalised regression estimation methodology for ABS business surveys demonstrated sample size reductions could be achieved when the auxiliary variable was correlated with the variable of interest being estimated, but more importantly, for multipurpose surveys the use of more than one auxiliary variable could achieve sample size gains even when estimating for different variables of interest. Before discussing these evaluations, the paper will use the quality framework to comment on the quality of the new tax data. When discussing accessibility, the paper will describe the development of an input data warehouse for storing and processing administrative data as well as ABS business survey data. The paper concludes with ABS intentions for further investigation and change.

2. DATA QUALITY ASSESSMENT

Data quality frameworks have been discussed in the context of managing a National Statistical Office (Brackstone, 1999) and for improving the information about quality which accompanies data products (Lee and Allen, 2001). In this section, the data quality framework is used to discuss the new tax data being used by the ABS to improve survey designs, methodologies and processes.

2.1 RELEVANCE

Relevance refers to the degree to which the data meets the real needs of clients (Brackstone, 1999). The main source of new tax data investigated by ABS was the business activity statement data. This data is considered very relevant to the ABS and as indicated in the introduction, the ABS has identified many areas in which the business activity statement data can be used to improve statistical processes, survey designs, and outputs. The relevance of the business activity statement data to the ABS was expected to some extent, as the ABS helped to develop the business activity statement data items to ensure some key data items, such as wages and salaries, were collected for statistical purposes.

2.2 COHERENCE

To be of greatest use to the ABS, business activity statement data needs to be coherent i.e. it needs to be consistent, logical and the various components need to be well aligned. This means, for example:

- that there is consistency in what is reported to the Australian Taxation Office with the annual business income tax return, and other company annual reports (such as those to the ABS)
- the data is complete i.e. there are no 'holes' in the data where data is missing for certain parts of the business population or for certain data items
- the aggregates and trends in the data are consistent with that published from other sources
- that relationships in the data are sustained over time e.g. data is of the same order of magnitude in each corresponding reference period and that there are no third month 'blips' where businesses report for a quarter instead of a month, and
- that similar relationships in the data are observed within individual data items and over time.

Coherence is a major factor for determining whether the tax data **will** be used to improve, substitute, supplement or expand ABS's statistical products. Early investigations of the new tax data indicated a degree of coherence existed, although timeliness issues have meant the ABS is yet to thoroughly evaluate this quality attribute.

2.3 ACCURACY

Accuracy refers to the degree to which the information correctly describes what it was designed to measure (Brackstone, 1999). For statistical products, accuracy is usually characterised by a relative standard error whilst for data itself, measurement errors can be used to describe accuracy.

Factors identified as affecting the accuracy of the new tax data fall into two broad categories.

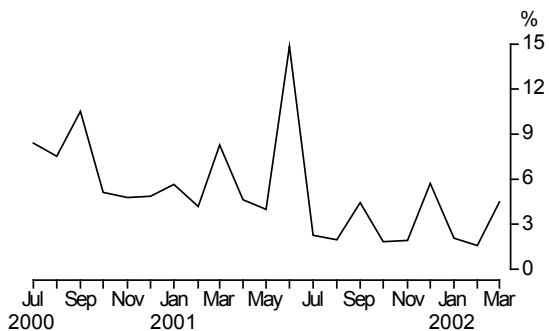
1. Incorrect or inaccurate data supplied by client. This may include providing correct data but presenting it at the wrong label, making arithmetic errors, failing to complete the required labels, compulsory fields are incomplete, role obligations are incorrectly selected or dates are invalid or missing.
2. Incorrect transfer of client supplied data to the system, including scanning errors and information incorrectly interpreted e.g. 'misread' handwriting or incorrect numerical data not picked up via key edit.

Editing the Business Activity Statement Data

From early investigations, an editing strategy for the business activity statement data was developed to identify and correct as many inconsistencies as possible. The edits were developed and applied either as the data was being loaded to ABS systems, or after the data was loaded and historical records or other values could be used to determine if an edit was required. Missing wage values for large businesses were a common inaccuracy detected by the editing process. Other inaccuracies detected were the wage value being too high or too low compared to other variables on the business activity statement return; or where the wages value inconsistently equals the amount of income tax withheld from employees by the business.

Graph 1 gives an indication of the overall percentage of business activity statement data records requiring edits for the wages variable, showing a noticeable improvement since the introduction of The New Taxation System in July 2000. The number of business activity statement returns increases fourfold in the third month of a quarter. This puts additional stress on editing systems. Furthermore, because wages and salaries is primarily collected for statistical purposes, corrective action 'at source' is not the highest priority. The ABS therefore expects to detect inaccuracies in wages and salaries and that the percentage of edits to be larger in the last month of each quarter. This can be seen in graph 1 for June 2001.

Graph 1: Records requiring editing of wages on finalised business activity statement records



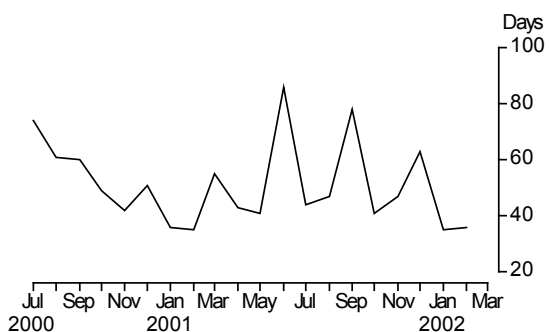
Other improvement strategies

Aside from an explicit editing strategy, the ABS and Australian Taxation Office are working closely to improve the accuracy of the data 'at source'. Regular liaison meetings are held where data quality concerns are raised and strategies for improvement are discussed. Discussion of investigations undertaken further improves the understanding of what quality issues are arising and the importance of them to each agency. This partnership is contributing to the improvement in data accuracy seen in graph 1.

2.4 TIMELINESS

Timeliness refers to the delay between the end of the reference period to which the data pertains, and the date when the data becomes available (Brackstone, 1999). Timeliness of the new tax data is a major impediment to using the new tax data widely within the ABS. Graph 2 shows changes in timeliness of finalised business activity statement data since July 2000.

Graph 2 : Days to achieve 70% of business activity statement returns finalised



Improvements are being seen in the time between when the data pertains and when it is finalised eg. 70% of February 2002 data were finalised 36 days after the end of the period, compared with 74 days for the July 2000 data. However, these improvements are not constant and the results still compare unfavourably with current ABS publication requirements eg. cut off for monthly Retail publication is 13 working days after the end of the reference period, and for the Quarterly Economic Activity Survey cutoff is 5 weeks after the end of the reference period. It needs to be recognised though, that the cutoffs for the business activity statement data are driven by administrative requirements, not statistical requirements. Cutoffs can vary between three and eight weeks, depending on whether a business remits monthly, quarterly, or through a tax agent. Taking into account these cutoffs, time to follow up late returns and time to process, it is unlikely that the business activity statement data will be able to be used for substitution in existing ABS statistical outputs with tight turnaround between the end of the reference period and the date when data becomes available.

Timeliness was a key concern when investigating the use of the new tax data for substitution, supplementation and new statistical products. Timeliness was not of such a concern, though, for investigating the use of the new tax data to assess generalised regression estimation or to improve stratification. Where data values were not available or available but not finalised to sufficient accuracy, predicted values can be used as updated size benchmarks, say, by averaging values over previous quarters.

2.5 ACCESSIBILITY AND INTERPRETABILITY

Accessibility of the new tax data by the ABS is improving. In early periods, problems with the data occurred in each extract. The problems related to both content and format, have been largely one-off, and once resolved, have not recurred. Furthermore, the use of CD-ROMs for data delivery has improved the link between the ABS and the source of the new tax data, and developments in electronic reporting offer further potential in this area.

Input data warehouse

Once within the ABS, accessibility of the new tax data as well as other large administrative datasets is an ongoing challenge. The new tax dataset is large and expertise in its content and format is in short supply. To help improve accessibility of the new tax data as well as survey data more generally, the ABS has started exploring the development of an analysis oriented data warehouse. The warehouse will be a managed unit record data store that services collection activities, including editing and estimation, analysis, research and management needs from initial data capture up to movement of aggregated data to the ABS' managed output data store.

The ABS input data warehouse is being developed to support both real time operational uses as well as analytical uses. The incorporation of real time operational uses is an extension of traditional warehouse models, incorporating the benefits of a warehouse into application systems to enhance the functionality available in the statistical processes during the processing cycle.

In addition to exploring opportunities for business and survey improvements through technology and methodology, the ABS is exploring organisational opportunities to harness efficiencies in its processing of economic data. The input data warehouse is seen as a key technological enabler in the development of new organisational arrangements. It supports a number of business outcomes such as reduced provider load, organisational efficiencies and improved data quality that have been difficult to achieve from having data and processes in separate systems and platforms across the ABS.

Eight business outcomes have been identified from the use of an input data warehouse in production processes. These benefits include many opportunities for improving ABS business surveys. For example, better use of existing knowledge about a respondent will reduce the number of times they may be approached during a survey. Methodological improvements, such as scientifically-based significance editing and intensive follow up will also be more readily explored as a by product of data audit trails captured during regular survey processing. Automated confrontation procedures in real time will be explored, particularly across surveys and in conjunction with existing administrative data sources.

The input data warehouse is expected to substantially reduce survey processing costs. The ABS expects to be better positioned to respond rapidly to changing needs and to more efficiently build and support new systems. This includes easier and more cost effective evaluation of the costs and benefits of current survey practices including input editing and intensive follow up. Following evaluation, system implementation of methodological changes will be more rapid and managed through a single unit record store rather than disparate solutions.

Finally, the input data warehouse will be used to improve the ABS' ability to support new analytical outputs and new products through the availability of confronted, time series unit record data in a single secure warehouse.

3 USING THE NEW TAX DATA

Despite concerns about the timeliness of the new tax data in section two, the ABS was able to evaluate the potential for improvements in the areas of survey stratification and alternative estimation methodologies. This section describes preliminary evaluations undertaken by the ABS in these areas.

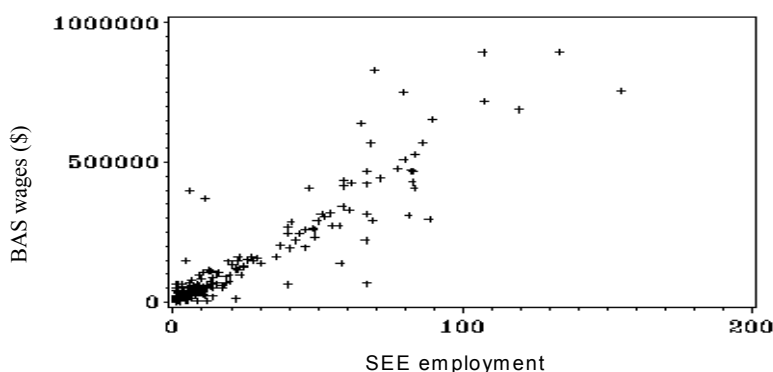
3.1 Survey Design - Stratification

One of the first opportunities for using the new tax data was to update benchmark values used in survey stratification. The ABS has maintained its own register of Australian businesses for many years. Stratification variables contained on the ABS business register include industry (based on the Australian New Zealand Standard Industry Classification, ANZSIC) and an employment size measure. The employment size measure is maintained for many large or significant businesses, but is not systematically maintained for the majority of small to medium size businesses. This can lead to sampling inefficiencies as well as classification errors when publishing estimates based on size groups.

Whilst timeliness and initially accuracy of the business activity statement data did not support changing stratification variables or stratification schemes, the new tax data did support the updating of our current size benchmark used in business survey stratification. Two employment-like size variables were available. When registering with the Australian Business Register, businesses are asked to provide their number of payees. The second source, obtained more frequently, is wages and salaries on the business activity statement. The ABS investigated the use of both the number of payees and the wages and salaries figures to derive an updated, yet comparable in magnitude, size measure. This was done by regressing average quarterly wages and salaries (or number of payees if wages was not available) against average quarterly employment collected in a recent ABS Survey of Employment and Earnings. No intercept was allowed in the regression model to ensure the derived size benchmark was zero for units with zero wages or payees. Using these associations, a scaling factor was applied to wages or payees to derive a size benchmark.

The scatterplot in graph 3 was typical of the association found between the new tax data and employment collected in the ABS Survey of Employment and Earnings. For the Health and Community Services Industry given in graph 3, there was a relatively strong association between the two data sources. To illustrate how the scaling factor works, if a unit had a tax data value of \$105,500 and for the Health and Community Services Industry the scaling factor was 7036, the units derived size benchmark became 14.99.

Graph 3 : Scatterplot of employment from ABS Survey of Employment and Earnings against business activity statement (BAS) wages for the Health and Community Services Industry



The objective of this investigation was to improve stratification efficiencies, not to find a perfect association between employment and the new tax size variables. Thus the objective was to derive a size measure that was suitable for use in stratification. After deriving and applying a suitable factor to wages and salaries (or payees if wages and salaries was not available), some adjustments were needed. It became apparent that the derived size measure was sometimes largely inconsistent, for the same business, with their existing employment measure. Further investigation of the wages and salaries or payees data indicated it was the more likely source of inaccuracy than the existing employment measure. A compromise between the derived size benchmark and existing employment measure was reached to avoid any adverse affect of businesses being incorrectly assigned to larger or smaller size strata.

Table 1 compares the number of businesses on our business register by their current employment benchmark and the derived size benchmark. Table 1 indicates an increase in the number of units in the larger size groups and considerable movement between size groups. Overall, 74.4% remain in the same size group, 14.2% move into a higher size group and 11.4% drop to a lower size group.

Table 1: Comparison of Derived Size Benchmarks and Current Employment Benchmarks

	Derived Size Benchmarks								
	Counts	0-4	5-9	10-19	20-49	50-99	100-199	200+	Total
Current Employment Benchmarks	Missing	57,166	8,381	5,032	2,696	1,448	0	0	74,723
	0-4	536,203	53,420	19,883	7,156	1,523	0	0	618,185
	5-9	57,076	34,003	14,120	4,720	705	0	0	110,624
	10-19	9,484	12,391	14,179	6,330	1,308	0	0	43,692
	20-49	2,150	2,136	5,279	8,259	2,315	613	48	20,800
	50-99	343	192	388	1,485	1,707	696	175	4,986
	100-199	9	6	18	45	63	86	43	270
	200+	39	14	23	37	36	30	147	326
	Total	662,470	110,543	58,922	30,728	9,105	1,425	413	873,606

The updated size measure has been embraced as the new size measure for business surveys from July 2002. At the time of writing, sample size changes as a result of using the updated size measure are not available. They are expected to be modest. By far the largest improvement from using the new size benchmark is not by reducing respondent load, but through systematically updating the size measure on a regular basis and thus improving sampling efficiencies.

3.2 Survey Methodologies - Generalised Regression Estimation

In late 2000, the ABS commenced investigations into sample size gains that could be achieved from using Generalised Regression Estimation, or GREG for short. (Bishop, 2000). ABS business surveys currently use two methods of estimation; number-raised estimation and ratio estimation. While ratio estimation allows the use of one auxiliary variable to improve the precision of the estimates, GREG estimation allows the use of more than one auxiliary variable and hence has the potential to be more efficient than number-raised and ratio estimation.

The business activity statement data items used to evaluate GREG were total salary, wages and other payments and total sales, income and other supplies.

Two ABS surveys were used as case studies to assess whether decreases in variances and hence reductions in sample sizes could be achieved by using GREG. The ABS Monthly Retail Business Survey collects one of the ABS' main economic indicators, Retail Turnover. It uses ratio estimation with frame employment as its auxiliary variable. The Quarterly Economic Activity Survey collects, from private businesses, income from the sale of goods, services, wages and salaries, company profits, and inventories in selected industries in Australia. It uses number-raised estimation.

Our investigation compared the current method of estimation used in each survey with

- ratio estimation using the most appropriate business activity statement variable as an auxiliary; and
- GREG estimation using a combination of business activity statement variables as auxiliaries.

To make our evaluation across surveys comparable, variances of estimates at the stratum level were used to determine optimum sample sizes. Relative standard errors were then calculated and used to obtain samples sizes to yield similar RSEs using the other estimation methods under investigation.

The results of our investigation are presented in table 2. Based on the sample sizes shown, RSEs for turnover from the Retail Business Survey, and turnover and wages from the Quarterly Economic Activity survey were calculated for each estimation method. Estimating turnover from the Retail Business Survey using its current estimation method of ratio estimation realised an RSE of 1.52%. If we used sales as the benchmark, the RSE reduces to 1.05% but if we use wages as the benchmark the RSE increases to 1.79%. However, if we use GREG estimation the RSE is nearly halved to 0.82%.

Similar results are obtained for estimating turnover or wages from the Quarterly Economic Activity Survey. RSEs reduce from 1.59% to 0.78% for estimating turnover, or reduce from 1.41% to 0.71% for estimating wages. Notice that ratio estimation will also reduce RSEs in the Quarterly Economic Activity survey when the auxiliary variable is correlated with the variable of interest. However, the benefit of GREG can be clearly demonstrated for the Quarterly Economic Activity survey which has many key variables of interest (turnover and wages are just two) and a different auxiliary variable for each variable of interest is not always possible.

Table 2 : Comparison of estimator efficiencies based on optimal allocation - RSEs

Comparison of estimator efficiencies based on optimal allocation	Australia level RSE Retail Business Survey Sample size 3770	Australia level RSE Quarterly Economic Activity Survey - turnover Sample size 12,373	Australia level RSE Quarterly Economic Activity Survey - wages Sample size 12,373
Number raised estimation	not used	1.59%	1.41%
Ratio estimation - stratum, with frame employment as benchmark	1.52%	not used	not used
Ratio estimation - stratum, with BAS sales as benchmark	1.05%	0.82%	1.54%
Ratio estimation - stratum, with BAS wages as benchmark	1.79%	1.54%	0.82%
GREG estimation - stratum	0.82%	0.78%	0.71%

The GREG evaluation aimed to demonstrate the likely gains in sample sizes that could be achieved. The results in table 2 indicate that to achieve the same RSE in the Retail Business Survey using GREG instead of ratio estimation, sample sizes can be reduced from 3,770 to 1,600 units in the sampled sector. For the Quarterly Economic Activity Survey, sample sizes can be reduced from 12,373 to 5,322 in the sampled sector by using GREG estimation rather than number raised estimation.

The evaluation indicated that using GREG with business activity statement data may afford substantial savings in sample sizes. The evaluation was preliminary and many caveats were attached to it. Matching survey data with the business activity statement data was difficult due to the lack of a hard match key, thus poor match rates were obtained. Methods for calculating variances and hence RSEs were also rudimentary; and an adhoc treatment of outliers was applied. Finally, imputation of missing business activity statement data for units on the sampling frame was required to obtain population totals for use in estimation.

Despite these caveats, the evaluation showed the potential gains that could be made and the ABS continues to invest in further development of the GREG methodology for ABS business surveys. Work has progressed in two main areas - developing a suitable variance estimation methodology and outlier methodology; as well as developing a tool to producing the generalised regression estimates and variances. (Preston and Chipperfield, 2002). The prototype tool will be used during 2002/03 to investigate more fully the gains that can be achieved from using GREG estimation. These investigations will include evaluation of the costs associated with implementing a new estimation methodology and system into business surveys.

4 CONCLUSIONS AND FUTURE OPPORTUNITIES

The availability of the new tax data has enabled the ABS to extend strategies for using non-ABS data in our statistical methodologies, processes and designs. Business income tax data has been used by the ABS for many years to improve our Annual Economic Activity Survey and the new tax data is enabling the ABS to explore possibilities of improving subannual business surveys, survey processes, and develop new or extended statistical products.

Timeliness of the business activity statement data has been the major deterrent to exploring substitution, supplementation and development of new or extended statistical products. The new tax data is considered relevant and coherent, editing strategies for improving accuracy are in place and working, accessibility is improving, and an understanding of what the new tax data means is growing.

The availability of the new tax data, despite concerns with its timeliness, still provides the opportunity to improve sample design and estimation methodologies. The ABS has successfully updated its stratification size benchmark to improve sampling efficiencies and these will be realised during 2002/03. Investigations into generalised regression estimation are continuing and early indications are that sample size reductions can be achieved, thus reducing reporting load on the Australian business community. The development of an input data warehouse to improve the accessibility of administrative data as well as ABS business survey data more generally will make these investigations, and other investigations into the use of administrative data, more rapid and efficient.

To conclude, the overall strategy for exploiting the availability of the new tax data will be as follows.

1. Continue to work with the Australian Taxation Office to improve the quality of the business activity statement data, particularly its timeliness, while recognising the limited potential in this area due to cutoffs, processing requirements, etc
2. Continue to explore robust methods for improving survey design and methodologies such as scientifically-based statistical significance editing and intensive follow up methods.
3. If timeliness improves, expand investigations into the use of the new tax data into areas of supplementation, substitution and new statistical products. The availability of the new tax data, as well as other administrative data sources, in an input data warehouse from July 2003 will help facilitate these investigations.

REFERENCES

BISHOP, G. (2000). Methodological Issues Associated with using BAS data for improving sample design and estimation for business surveys. *ABS Methodology Advisory Committee paper*, November.

BRACKSTONE, G. (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology*, Vol. 25, No. 2, pp. 139-149.

LEE, G. and ALLEN, B. (2001). Educated Use of Information about Data Quality. *International Statistical Conference*, Fifty third Session, Seoul

PRESTON, J and CHIPPERFIELD, J. (2002). Using a Generalised Estimation Methodology for ABS Business Surveys. *ABS Methodology Advisory Committee paper*, June.