Revised FINAL DRAFT December 2004

Business Longitudinal Database (BLD) - Proposed Design

Issues

1. At the July 2004 ESSCC, two issues were raised on the BLD paper that require addressing in a further ESSCC paper. They are:

> The BLD team to set out the main user needs in relation to the Business Longitudinal Database; this should include a full recommendation as to what data items should be included in the BLD and at what time intervals they should be collected.
>
> Mr Sutcliffe and the BLD team to develop a prototype of the Business Longitudinal Database design, based on the current Innovation Survey in the first instance (with possibilities in relation to the Integrated Business Characteristics Survey to be also examined); the proposal in relation to this prototype should include a costing and timing proposal, updated sample design figures, revisiting the statistical unit issue, future design issues, and revisiting the scope and coverage issues.

2. The relationship with the proposed Integrated Business Characteristics Survey (IBCS) is discussed in a separate paper but is an essential issue to be resolved in the short term. In particular, it is important to obtain clarification on the requirement for a BLD to provide the capacity to deliver cross-sectional estimates.

Recommendations

Recommendation 1: *The BLD should be designed to meet longitudinal modelling requirements and not to produce cross-sectional estimates.*

Recommendation 2: *That the basic BLD design methodology be accepted.*

Recommendation 3: *The TAU is adopted as the unit of reference for the BLD.*

Recommendation 4: *The scope and coverage for the BLD will be all TAUs on the ABS Business Register with at least one active tax role and will:*

include all business sizes; non-employing, small, medium and large;
include SISCA sectors 1, 2 and 4, and exclude sectors 3 and 5;
include Government Business Entities with SISCA sector 1;
exclude ANZSIC Division A (Agriculture, Forestry and Fishing); and
exclude Trusts, Not for Profit Institutions, Churches..

Recommendation 5: *That ESSCC endorse reference period of 2004/05 (dispatch in Feb 2006) as a the most appropriate date for the new business longitudinal survey.*

Desired Outcomes from ESSCC

The desired outcomes of this ESSCC discussion is that the above recommendations are endorsed.

Business Longitudinal Database (BLD) - Proposed Design

Introduction

1 The aim of this paper is to broadly discuss the status of conceptual, measurement and survey design issues surrounding the construction of the survey component of the Business Longitudinal Database (BLD) and to progress with some direction setting.

2 There has been recognition within our user community, and in most overseas statistical agencies, of the requirements for data linking; leveraging off administrative datasets and the development of longitudinal datasets especially business longitudinal datasets. The BLD encapsulates all aspects of these complex issues.

3 The main outcome of the initial feasibility work into the business longitudinal database acknowledged that there is no generally recognised or standardised methodology to collect and disseminate longitudinal data. The prime function of the BLD will be to facilitate the analysis of a range of business policy issues. In order for this analysis to occur it is essential that the BLD contains not only financial variables but also, explanatory business characteristic variables. Without these variables it will not be possible to accurately assess the underlying causes of business growth and performance. To obtain these variables, we will need to survey businesses.

4 This paper develops collection options and adopts pragmatic solutions wherever applicable. This paper addresses the following aspects:

the aims and objectives of a BLD including the evaluation of user requirements;

basic longitudinal survey design issues;

the unit of reference;

scope and coverage;

resource issues; and

timing.

Aims and objectives of the BLD

5 The prime function of the BLD will be to facilitate the analysis of a range of business policy issues based around business growth and performance with emphasis on business structures. Strong demand from a wide range of government and private sector users of the previous Business Longitudinal Survey (BLS) data has resulted in the ABS initiating a project to develop a Business Longitudinal Database. This database will use the ABR as the population frame, and a combination of administrative and ABS survey data. The ABN has provided the basis for hard-linking a wide range of datasets, including those referred to above, something that was not possible with previous longitudinal database development work.

Evaluation and clarification of user needs and uses.

6 The ABS has consulted with a wide range of users and obtained information about the types of analyses that are likely to be performed on the BLD. These users have also provided advice on the trade-offs they would make under different budgeting scenarios. For example, they suggest it would be preferable to target particular industries rather than having sparse coverage in all industries. The main issue at the moment is that the ABS does not have an existing survey which collects information on business characteristics. Current thinking is that if we want to maximise the amount of data available for each business included in the BLD then we have four main options:

financial data from administrative sources such as BIT and BAS data;
financial data from ABS surveys such as EAS and related surveys;
business characteristics data existing surveys such as the Innovation Survey and Business Use of IT Survey; and
business characteristics data from a separate longitudinal survey.

The design and conduct of an Integrated Business Characteristics Survey to cover the third and fourth requirements is being considered.

7 In September 2004, members of the BLD External Advisory Group (EAG) were asked to consider the detailed list of topics specified by users during a consultation process late in 2003. These requirements are summarised in Appendix 1. The discussion at the EAG meeting was useful in developing a better idea of the relative priorities but a significant amount of additional work is required before the requirements can be considered final.

8 When further considering these requirements, we will need to be cognisant of the following issues:

Many of the specified requirements can be addressed using existing cross-sectional data. They are often questions about the nature of the profile an industry which can be handled by existing datasets. An example would be the percentage of exporters which have female managers.

Longitudinal modelling is about cause and effect involving time lags (ie dynamic modelling). This form of modelling is almost essential if two factors can be both cause and outcome of each other. For example, does higher R&D cause higher profits or are more profitable companies more likely to spend more on R&D since they have the money to do so? Cross-sectional datasets will never allow us to uncover the true relationship. Hence preference in data collection should be given to data relevant to hypotheses that have this known dynamic. We need to determine if something is this going to be a causal factor or an outcome that we are interested in? This said, some factors are useful as controls, but in most cases the controls are other possible causal (& therefore complicating) factors.

Most models can only handle a limited number of variables (questions), otherwise the estimation is poor (too few degrees of freedom). So there is no point having too many closely related (and therefore correlated) questions. Analysts will usually seek to reduce the number of questions using some data reduction method such as factor analysis. Related to this is the problem of asking questions with very low numbers of responses. If there are too few positive responses (so the frequency distribution of the questions, is 95% zeros and 5% positives) we may not get a significant estimated coefficient. It is better to make the question broader so more people respond positively. This also makes the question less time specific and more temporally endurable.

List questions, are good for cross-sectional analysis but can not usually be used for modelling. A list question is: what are the main forms of barrier to further expenditure on R&D? (with listed responses). What difficulties do you have hiring skilled labour? etc. In some cases, these questions can be used in modelling, but the relevant information can be more efficiently extracted in the survey by simply asking respondents on a scale 1 to 10 for how difficult they have found hiring skilled labour over the last 12 months.

Questions about 'what would you do if...' are not of general interest unless one is doing a study on the realisation of business intentions (these are not common). Usually you get stronger information from analysing what people have actually done. When you ask then about hypothetical behaviour you have a large an unknown error about how seriously they answer the question and how committed they are to doing what they thought they might.

9 Based on discussions with external and ABS analysts, it appears that the preference is for a smaller-denser dataset (a dataset which is smaller in terms of sample size yet denser in terms of data items) to augment a larger-sparser dataset.

Sample Design Options

Background

10 Since the design of the BLD was last presented to ESSCC in July 2004, further work has been completed. This has involved assessing how well current ABS surveys can provide units to populate the BLD initially and over time. That is, progressing Option 2 presented in the previous ESSCC paper, using real rather than example sample sizes.

11 Option 2 essentially added a new sample of units to the BLD each year - called a panel - and followed these units for five years - referred to in longitudinal literature as waves. This work has assumed that the requested exemption from the current provider load requirements will be granted, allowing units to be retained in the BLD for five years rather than three.

12 This design was presented to the BLD External Advisory Group on the 23rd of September, where it received general support. Incorporating feedback from this meeting into the design presented (essentially that a sample of large businesses be included and that Industry is a more important dimension than state), has resulted in the current proposed BLD design presented in this paper.

13 Note in the following that response is assumed to be complete (that is, there are no adjustments made to sample size to allow for non-response) and detailed work on attrition rates (that is, businesses that cease operating between one year and the next) has not been undertaken and a fixed rate has been assumed over the five years.

14 It is also important to note that the optimal sample for a longitudinal dataset created specifically for modelling/analysis purposes is likely to be quite different to that of an optimal sample created to produce point-in-time estimates. In usual business surveys where the latter is the aim, the allocation tends to be disproportionate since the population is skewed (the small number of large businesses contribute the majority of the final estimates). Such a disproportionate allocation is unlikely to be optimal for estimating parameters of statistical models, where a sample spread over the domain of study would be most useful. Thus the BLD should not be designed to produce accurate cross-sectional estimates as this would compromise its ability to meet the longitudinal modelling requirements.

Recommendation 1: The BLD should be designed to meet longitudinal modelling requirements and not to produce cross-sectional estimates.

Design of the BLD

15 In order to develop a viable design over a long period of time we have split the task into two main themes:

> Firstly, to maximise the degree of overlap between this BLD and existing ABS surveys by looking to see if an appropriate sub-sample of the existing samples can be developed. The optimal situation would be for the first wave of the first panel to consist completely of businesses currently selected in an ABS survey.

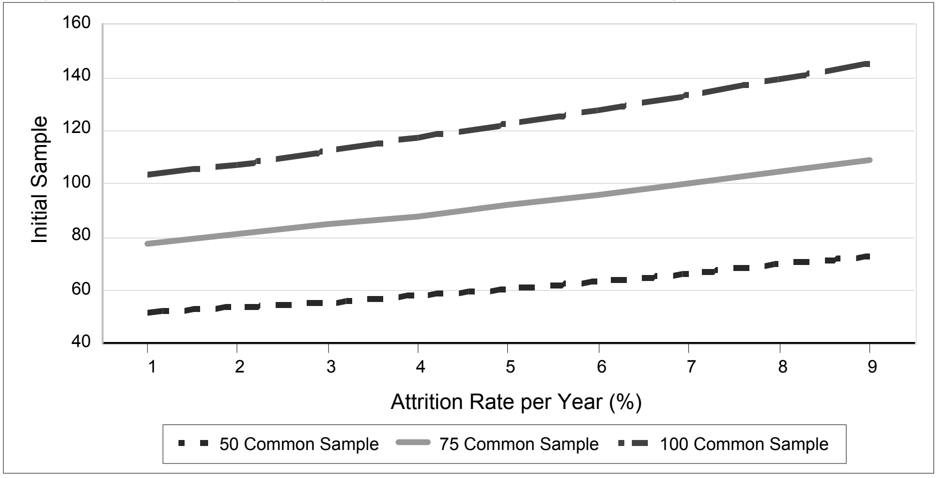
Secondly, to view the creation of the BLD as an ongoing process, where after 5 years the total sample size would reach a 'steady state'. In this framework new sample is added every year, so a new five year panel is commenced. Therefore, in the steady state situation, there are 5 panels operating, each at different ages. The aim is then to determine what is an appropriate common sample over five year periods. The problem can be viewed in two ways:

determining the steady state sample size (S) required to produce a common sample (C) which meets the analysts needs;

determining the common sample (C) produced given a fixed steady state sample size (S).

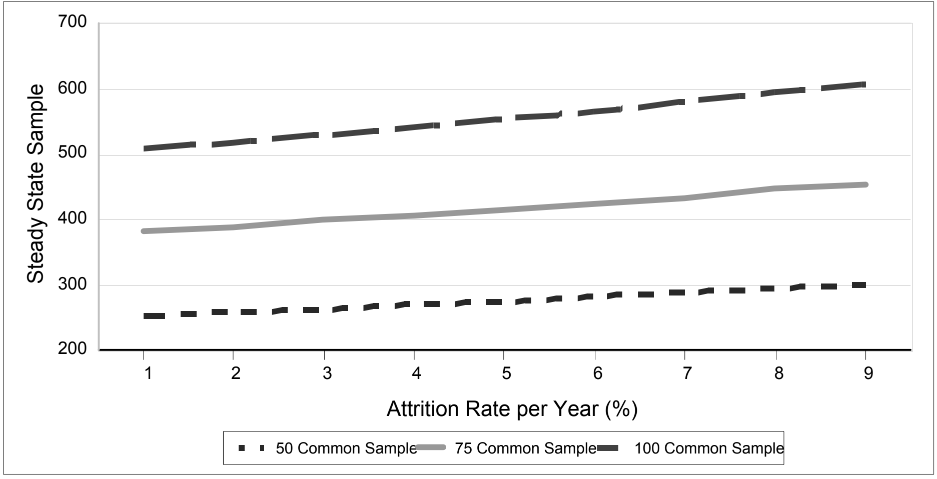
16 It is possible to construct the relationship between the common sample C over a five year period and the steady state sample size S. To achieve a common overlapping sample of size C after five years, the initial sample taken needs to be larger to allow for attrition over time. Naturally, the larger the attrition rate, the larger the initial sample will need to be. Similarly, the larger the common sample required, the larger the steady state sample will be in any one wave. The first graph below shows the initial sample size required to achieve either 50, 75 or 100 common units after five years, based on different attrition rates. As expected, for a given attrition rate, the larger the common sample size required, the larger the initial sample needed. Further, as the attrition rate increases, so does the size of the initial sample needed for the same common sample (as more businesses are expected to cease over the five years).

Graph 1 : Initial Sample Required to achieve common sample



17 Graph 2 depicts the relationship between the steady state sample size (ie total sample included in a given year/wave) for a given common sample size after five years, for various attrition rates. Again, for a particular attrition rate, the steady state sample increases as the common sample required increases. And for a given common sample size requirement, the steady state sample increases as the attrition rate increases.

Graph 2 : Steady State sample required to achieve common sample



18 After comparing a number of combinations of common and steady state sample sizes with a fixed 8% attrition rate, it appears the initial sample required is approximately 1.5 times the common sample, and the steady state sample size approximately 6 times the common sample size.

<u>*Initial BLD Sample using current ABS surveys as a base*</u>

19 Given the above design, a prototype was constructed using current Innovation and BUIT selections as a basis. While some adjustments were made to account for the differences in scope between these and the proposed BLD, the available data was still missing some key groups of units. Namely non-employers, and those in ANZSIC Division N (Education). Note that this does not mean these are out-of-scope of the BLD, just that these two surveys are unable to provide such units.

20 Table 1 gives the distribution of the population of businesses in the matched scope between the BLD and the Innovation and BUIT surveys. Only businesses in the Private sector are included. Note that in this table and those following, the unit used is the TAU as currently used in Innovation and BUIT surveys. For information on how this relates to the EG units for an example BLD sample see Appendix 2.

Table 1 : Distribution of available businesses, by size and industry, Private sector

	B	C	D	E	F	G	H	I	J	K	L	O	P	Q	Total
0-4	1707	35779	381	86088	27467	77973	21262	27511	6781	18934	108192	28812	11846	21648	474381
5-19	441	11493	63	9588	8844	20403	7958	3579	296	2770	23308	8495	2690	3832	103760
20-199	219	4474	37	1614	2812	3722	2355	1007	102	758	6223	1547	743	727	26340
200+	67	592	19	75	200	173	101	110	14	129	325	67	54	32	1958
Total	2434	52338	500	97365	39323	102271	31676	32207	7193	22591	138048	38921	15333	26239	606439

21 For the prototype presented in this paper we have assumed that we require 30 businesses in each ANZSIC Division by Size Group. Setting the attrition rate at 8% per year, the required steady state and initial sample sizes at the ANZSIC by size group and overall are as shown in Table 2. This reveals that for any given ANZSIC by size group, 42 new businesses need to be added each year, and the steady state sample size will be 179. In Appendix 3 similar tables are presented where the requirement is 20 common businesses in each ANZSIC by Size Group.

Table 2 : BLD structure for 30 common units in each ANZSIC by size group

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
Panel 1 ->	42	39	35	33	30					
Panel 2 ->		42	39	35	33	30				
Panel 3 ->			42	39	35	33	30			
Panel 4 ->				42	39	35	33	30		
Panel 5 ->					42	39	35	33	30	
Panel 6 ->						42	39	35	33	30
Panel 7 ->							42	39	35	33
Panel 8 ->								42	39	35
Panel 9 ->									42	39
Panel 10 ->										42
Total Sample Size	**42**	**81**	**116**	**149**	**179**	**179**	**179**	**179**	**179**	**179**

22 If we were aiming to ensure that once a given business has completed its five years in sample it has at least five years out of sample a larger population than the steady state values given above are needed. For example, in Year 9 in Table 2 above, there are 179 businesses currently included. In the following year a further 42 new businesses will be included, and we need to ensure that these aren't selected from the 30 which rotated out in any of the previous 5 years. Ignoring for a moment the fact that of the 30 that rotate out each year a number will die in each subsequent year, in Year 10 we cannot select from those units which rotated out in Years 6, 7, 8 or 9 or 10. That is, there will be 150 businesses that cannot be included. Along with the 179 which will be in the sample, this gives a minimum population requirement of 329 businesses.

23 Comparing this requirement to Table 1 which gives available population, we can see that the 20-199 group in ANZSIC B (Mining), the two larger size groups in ANZSIC Division D (Electricity, Gas and Water), and the two larger size groups in ANZSIC J (Communication) will be unable to accommodate the 329 required businesses. These have been excluded from further analysis. The 200+ groups with less than 329 businesses are retained, as rotation is not as important an issue for these larger units. However, where there are less than the initially required 42 units, these groups are excluded, so 200+ by ANZSIC Division D (Electricity, Gas and Water), ANZSIC Division J (Communication) and ANZSIC Division Q (Personal and Other Services). Since ANZSIC Division N (Education) is out of scope for both BUIT and Innovation population counts are not presented in Table 1. Frame information from other sources reveals that there are sufficient in-scope private sector units in size groups 0-4 and 5-19.

24 The total common sample required over these remaining 50 ANZSIC by size groups (ie 50*30 = 1,500) is given in Table 3 below. As this table shows, for each wave a new panel of 2,100 units will be added to produce the 1,500 common units. By Wave 5 when the steady state situation will be reached, a total of 8,950 units will be in the BLD in each wave/year.

Table 3 : BLD sample for 30 common units in each ANZSIC by size group

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
Panel 1 ->	2100	1950	1750	1650	1500					
Panel 2 ->		2100	1950	1750	1650	1500				
Panel 3 ->			2100	1950	1750	1650	1500			
Panel 4 ->				2100	1950	1750	1650	1500		
Panel 5 ->					2100	1950	1750	1650	1500	
Panel 6 ->						2100	1950	1750	1650	1500
Panel 7 ->							2100	1950	1750	1650
Panel 8 ->								2100	1950	1750
Panel 9 ->									2100	1950
Panel 10 ->										2100
Total Sample Size	**2100**	**4050**	**5800**	**7450**	**8950**	**8950**	**8950**	**8950**	**8950**	**8950**

25 To determine whether the ABS will need to source data from units not selected in either Innovation or BUIT, we have mapped the basic BLD prototype design given above against the current selections from these surveys. Table 4 shows the distribution of these samples (after scoping to meet as close as possible the BLD scope, and removing those ANZSIC by size groups where there is insufficient population to meet the provider load requirements as discussed in paragraph 23 above). Since ANZSIC Division N is out of scope of both BUIT and Innovation no sample is available for this industry. Note that all the 0-4 units and those in Industries O and Q are taken from BUIT, whilst the remainder are from Innovation. This is not necessarily the sample which would be used for an Integrated Business Characteristics Survey (IBCS) as it would not be a simple merger of the two current designs.

Table 4 : Distribution of current samples, by size and industry, Private sector

	B	C	D	E	F	G	H	I	J	K	L	N	O	P	Q	Total
0-4	173	172	86	753	112	262	236	296	397	133	201	0	288	143	245	3497
5-19	90	1128	-	144	147	110	104	130	-	81	226	0	116	129	54	2459
20-199	-	1052	-	123	136	111	109	134	-	75	231	-	46	97	43	2153
200+	67	592	-	75	200	173	101	110	-	129	325	-	67	54	-	1893
Total	**330**	**2944**	**86**	**1095**	**595**	**656**	**550**	**670**	**397**	**418**	**983**	**0**	**517**	**423**	**338**	**10002**

26 Given that all ANZSIC by size groups (except those in Education) in Table 4 above have at least 42 businesses, it would be possible for the BLD sample to be taken (almost) fully from current ABS survey samples. Further discussion with users needs to occur to determine whether all industry by size groups are needed on the BLD, for example ANZSIC Divisions D and J may be excluded given the relatively small sample sizes.

27 The above sample would be selected to ensure that there is a certain number of units in each size category and ANZSIC Division. However, it is of interest to see how such a sample is distributed over other characteristics, such as State and reported data such as 'innovator'. The distribution across states for an example sample (excluding Education) is given in Table 5, which shows that it is relatively similar to that available. Reported data has not been assessed as the Innovation survey results have not yet been released.

Table 5 : Distribution of Panel 1, Wave 1 Sample by State

State	BLD Sample	%	Inscope (BUIT and Innovation) Sample	%
NSW	693	34.38	3343	33.4
Vic	416	20.63	2365	23.6
Qld	295	14.63	1497	14.7
SA	214	10.61	992	9.9
WA	228	11.31	1028	10.3
Tas	69	3.42	314	3.1
NT	48	2.38	211	2.1
ACT	53	2.63	252	2.5

<u>*BLD Sample over time*</u>

28 Maintenance of the sample over time was not presented to the External Advisory Group on the 23rd of September. This issue of whether the units remain in the Innovation and BUIT surveys or rotate out and so need to be followed up separately was felt to be an internal ABS concern.

29 Given the current provider load policy the majority of sampled units initially selected in either Innovation or BUIT will not be included in these samples after the third year has been completed, and some will rotate out even sooner. For Panel 1 only, this is depicted in Table 6 below, where the initial sample has been selected so as to maximise the length of time each unit will be included in the Innovation or BUIT sample. After the first year, of the 1,950 continuing live units (from the original 2,100) 1,872 will still be selected in Innovation or BUIT, and 78 will be in Education. After the second year, 29 (non-education) units will have rotated out bringing the total of live rotated out/Education units to 99. For the fourth and fifth years of the BLD a large proportion of the remaining live units (666 in Year 4 then 872 in Year 5) will need to be sourced separately. Table 6 also gives the breakdown by size group. As this shows, continuing after Year 3 are the large CEd units, and some of the sampled unit (these are taken from strata which are non-rotating). This situation is also shown diagrammatically in Figure 1 below Table 6.

Table 6 : BLD Panel 1 - units remaining in Innovation or BUIT survey over time

		Year 1	Year 2	Year 3	Year 4	Year 5
0-4		588	546	487	0	0
5-19		504	468	410	323	65
20-199		462	429	369	298	233
200+		462	429	385	363	330
Continuing	*Total*	*2016*	*1872*	*1651*	*984*	*628*
0-4			45	105	135	180
5-19			39	91	117	156
20-199			33	77	99	132
200+			33	77	99	132
Died	*Total*		*150*	*350*	*450*	*600*
0-4		42	39	38	495	450
5-19		42	39	45	106	325
20-199				16	65	97
200+				0	0	0
Rotated out and Education	*Total*	*84*	*78*	*99*	*666*	*872*
Total		**2100**	**2100**	**2100**	**2100**	**2100**

30 Figure 1 refers to the proposed Integrated Business Characteristics Survey (IBCS), which is the intended vehicle for collecting (as much as possible) BLD data. It is assumed that the sample size of the IBCS is 8,000.

Figure 1 : BLD Panel 1 over five years

Year 1	→	Year 2	→	Year 3	→	Year 4	→	Year 5
BLD1 only 84		BLD1 only 78		BLD1 only 99		BLD1 only 666		BLD1 only 872
BLD1/IBCS 2,016		BLD1/IBCS 1,872		BLD1/IBCS 1,651		BLD1/IBCS 984		BLD1/IBCS 628
IBCS only 5,984		IBCS only 6,128		IBCS only 6,349		IBCS only 7,016		IBCS only 7,372

31 It is difficult to accurately portray the situation in following years, as the sample available from which to select the Panel 2 BLD units is dependent on the frames used for the Innovation and BUIT selections, plus the amount of rotation that occurs. While it is possible to predict planned rotation, unforced rotation (due to frame birthing and deathing) cannot be anticipated with any accuracy. Table 7 belows shows for subsequent panels the amount of sample available from Innovation and BUIT, that is, after excluding those units selected in Panel 1 and any that have rotated out. Replacing the rotation outs are rotations in, which are available for selection in these future panels. Also available would be births, but as mentioned it is impossible to predict where these will be. Also impossible to predict is units in the Innovation and BUIT samples which will be deathed in subsequent years. As such the numbers presented below are the best-case scenario. Any frame birthing and deathing or unusual rotation patterns will reduce the size of the available Innovation and BUIT samples, and increase the size of the BLD sample which needs to be sourced elsewhere.

32 Table 7 shows that Panel 2 of the BLD could also be taken fully from the Innovation and BUIT samples, but the pattern across time is different to that in Panel 1 with more units rotating out sooner. This pattern continues for the subsequent panels.

Table 7 : BLD Sample split by source

		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
	Innovation/BUIT	2,016	1,872	1,651	984	628	
	Additional	84	78	99	666	872	
	Innovation/BUIT		2,016	1,767	1,489	733	429
	Additional		84	183	261	917	1,071
	Innovation/BUIT			2,009	1,782	1,526	800
	Additional			91	168	224	850
	Innovation/BUIT				1,980	1,782	1,540
	Additional				120	168	210
	Innovation/BUIT					1,873	1,763
	Additional					227	187
	Innovation/BUIT						1,743
	Additional						357
Innovation/BUIT		**2,016**	**3,888**	**5,427**	**6,235**	**6,542**	**6,275**
Total	**Additional**	**84**	**162**	**373**	**1,215**	**2,408**	**2,675**

33 A possible alteration that could be made to the situation depicted in Tables 6 and 7 above would be to take a larger sample for the first one or two panels to enable us to refine the estimates of year-to-year attrition for input into the design of the future panels. This would also result in a large amount of overlapping sample in the first few years. Up until Year 4 the steady state sample size has not yet been reached, so adding additional sample would not result in a larger first few panels than the on-going situation. However, it would impact on the amount of sample available for future panels of the BLD from Innovation and BUIT as more would be included in the first few panels.

Sample Design Conclusion

34 We have presented the total sample size required for the BLD to result in 30 common live units over a five year period in each of 50 ANZSIC Division by size groups. It appears that the initial sample (apart from that in ANZSIC Division N - Education) can be taken from the existing ABS surveys of Innovation and BUIT. However, for the fourth and fifth years of each panel, the majority of included units will have rotated out and will need to be contacted in a separate process to collect the BLD data items.

35 The above sample size can be reduced by either :

> rescoping - reducing the number of ANZSIC Division by size groups included in the BLD; or
>
> resizing - reducing the common sample available within each ANZSIC Division by size group (see Appendix 3 for an example of 20 common units over 5 years).

Recommendation 2: That the basic BLD design methodology be accepted.

Statistical Unit/Reference Unit

36 The aim of the BLD is to combine unit level data from different sources to provide data for analysis of the growth and performance of businesses. These data sources don't necessarily have the same units of reference. This raises the question of which unit should be the unit of reference in the BLD. The statistical unit is the base level unit at which information is sought. The choice of the statistical unit has been revisited and discussed with users at the recent EAG meeting in September. After much debate, there was agreement that the TAU should be recommended as the unit.

37 One of the main data sources proposed for the BLD is Australian Taxation Office (ATO) data for which the unit of reference is the Australian Business Number (ABN). The ABN is an ATO unique identifier and is available on the Australian Business Register (ABR) and Business Activity Statement (BAS) datasets, and is linked to the Business Income Tax (BIT) data upon loading to the ABS Input Data Warehouse (IDW). Another main data source is ABS survey data for which the unit of reference is usually a Type of Activity Unit (TAU) and less often is an Enterprise (EN) or an Enterprise Group (EG). As ATO and ABS data will be the main sources of data for the BLD, each of the ABN, TAU, EN and EG should be considered as potential units of reference.

38 For entities which are part of the ATO maintained population (ATOMP), the ABN, TAU, EN and EG are effectively equal. This means linking between different data sources containing any of these units of reference is not an issue. That is, linking issues do not influence the choice of BLD unit of reference for the ATOMP.

39 For entities which are part of the ABS maintained population (ABSMP), the relationship between an ABN and the other 'maintained' maintained, the TAU, EN and EG, is more complex. The alternatives are: use the ABN link to the EG or use the ABN link to a TAU. For the EG, the linkages are direct but

there is not linear relationship between ABN's and TAU's. Within the ABSMP, most TAU link one to one with ABN's however there are cases of many ABN's to one TAU, many TAU's to one ABN and many ABN's to many TAU's. While these latter scenarios are more unusual, the maintenance of any ABN selected via a TAU in the ABSMP will be time more consuming and complex. Data cannot always be confidently linked between these units. Nevertheless, the TAU is the main economic statistical unit and the most appropriate unit for industry analysis.

40 The sampling of the TAU's in the BLD and the relatively small population of these type of units in the BLD (see para xx) will mean that the maintenance is minimised and appropriate resources can be devoted to maintaining ABN/TAU links. The costings model allows for some extra resources in the mapping, editing and relationship maintenance of these large or complex TAU's selected in the sample. The choice of using an EG instead of a TAU in the ABSMP for the BLD are summarised in Appendix 2.

Recommendation 3: The TAU is adopted as the unit of reference for the BLD.

Scope and Coverage

41 As with most statistical collections, there are some groups excluded from the population frames. The population for selection of business units for the BLD is for all ABNs on the Australian Business Register with at least one active tax role and will;

include all business sizes; non-employing, small, medium and large;
include SISCA sectors 1, 2 and 4, and exclude sectors 3 and 5;
include Government Business Entities with SISCA sector 1;
exclude ANZSIC Division A (Agriculture, Forestry and Fishing); and
exclude Trusts, Not for Profit Institutions, Churches.

42 The scope and coverage would include all business sizes, small, medium and large. Analysts need a representative economy-wide panel, including coverage of differing business sizes in order to perform meaningful analyses, and compare performance. Tracking large businesses will be more complex and costly, and selections of these unit will be kept to a minimum. These large businesses traditionally have a high volume of activity, for example takeovers, mergers etc and larger financial activity.

43 The scope of the BLD should include units in all industries in the Australian economy, as taken from the Australian and New Zealand Standard Industrial Classification (ANZSIC) excluding Division A (Agriculture, Forestry and Fishing), and Division M (the Government sector).

A meeting was held with representatives from the agriculture area of the ABS Goods and Services National Statistical Centre on 16 February and conceptually, it was agreed that the Agriculture industry may be included in the BLD. However, practically, it was agreed that at present, due to the volatility of recent changes to the agricultural population frame that inclusion of the agriculture industry be deferred. Most business survey areas exclude Agriculture from their selections, and therefore any BLD would produce a panel that had only a few overlaps with Agriculture data, ie hypothetically, lots of "missingness".

Government industry units are not part of the businesses which are of interest to our users

Collection of data for Division K is also being considered for exclusion. The BLD development team will undertake further research into inclusion of the Finance Division. The Finance sector has the approximately 325,000 trusts, is very difficult to establish the linkages using ABN (from EAS experience); are not well defined on the ABR and ABSBR. BLD users are aware of the issues associated with this Division and would not be too disturb if excluded from any survey collection.

44 The scope of the survey would exclude all businesses in SISCA sector 3 (government sector). Government units are not typically market driven or oriented - they are not businesses per se. The exclusion of some of the public sector data was an issue for some of our users, for example, the export

of educational services from government units would be missing from the survey. However, we envisage that data can be collected for most aspects of government unit via other vehicles.

45 Government business entities (GBEs) should be included in the BLD, (SISCA sector 1) as they participate in the market and have 'profit maximizing' aims or charge economically significant prices.

46 The intention is also to exclude all businesses in SISCA sector 5 (non-profit institutions serving households). These "businesses" are not required to lodge Business Income Tax and do not operate under market conditions. Nevertheless, some may have substantial business activity and are often coded to SISCA 1 as a result of their corporate trading activities. Examples of these businesses include second hand clothing stores. Not for Profit organisations, and Churches are also exclude because their activities are not seen as relevant to a business longitudinal database analysis.

47 Non-employers remain an area where survey coverage is irregular and are presently still being investigated. Overlap with existing surveys has been discussed but for most ABS surveys, non-employers are not sampled. The BLD users have expressed a requirement for non-employers to be included. There are a number of government initiatives based around growing businesses, exempting from certain legislative processes (unfair dismissal etc) and analysis of financial and productivity performance. the issue that needs considering in the sample design aspect of this decision is the large number of non-employing businesses. This is discussed in Appendix 4.

48 The design effect of including non-employers will flow through to the sample sizes, namely:

Inclusion of non-employers - 11,635

If one size group were used for all non-employers and all 15 industries were used this would add 15 groups to the BLD. Aiming for 30 common after 5 years in each would add (179*15=) 2,685 units to the steady state.

Inclusion of non-employers by size - 19,690

If the non-employers were somehow split into size categories (say 4) this would add 60 groups to the BLD, and so add (179*60=) 10,740 units.

Recommendation 4: The scope and coverage for the BLD will be all TAUs on the ABS Business Register with at least one active tax role and will:

include all business sizes; non-employing, small, medium and large;
include SISCA sectors 1, 2 and 4, and exclude sectors 3 and 5;
include Government Business Entities with SISCA sector 1;
exclude ANZSIC Division A (Agriculture, Forestry and Fishing); and
exclude Trusts, Not for Profit Institutions, Churches.

BLD cost issues

48 The BLD costs will depend on the sample size, panel maintenance costs and survey processing costs. The cost model being developed takes account of:

fixed cost associated with producing a database each year;
the cost of adding administrative data;
the cost of adding data from other surveys;
the cost of keeping businesses in the BLD panel after they rotate out from ABS surveys;
the cost of conducting a specific survey to collect longitudinal business characteristics; and
processing costs (editing, validation); and

data access and dissemination and output costs.

49 Appendix 5 outlines the estimates of costs associated with samples of 12,000 and 6,500 up to 18 waves into the business longitudinal survey. The costing model is based on the user funded ABS corporate charging regime and would need to be adjusted if the underlying assumptions are altered. There have been adjustments made for a 3% pa increase in salaries and other costs. Most of the assumptions are discussed in the spreadsheet.

50 The estimated staffing costs within the NSC are based on the present team structure however, that may change over time, depending on a number of factors. One of the key determinants will be data access arrangements. There may be a requirement for a larger support team internally if the majority of analytical projects are done on a collaborative basis but such arrangements may normally be undertaken on a cost recovery basis, reducing the overall costs to the ABS.

Timing issues

51 Choosing a dispatch and reference period for the BLD survey that will give the greatest coherence of data is dependent on the availability and quality of suitable data for matching. These data will comprise; ABR, BAS, BIT, existing ABS collections as well data from the BLD survey.

52 Data are available for ABR, Business Activity Statement (BAS) and Business Income Tax (BIT) data for the 2000/01 reference period onwards. However, for BIT data there is an 18 month lag in data availability following the reference period. Data from the initial Innovation Survey (in respect of the 2003 calendar year) is scheduled for release in November 2004. Data for the 2001/02 and 2002/03 EAS and related surveys will be released in early 2005. However, the effort required to map the metadata to the CPCF framework is such that data for previous years may not be available on the IDW until early 2006 at the earliest.

53 The earliest a business longitudinal survey could be developed would be in respect of the 2004/5 financial reference period, with collection activity commencing in February 2006 (the despatch date for the IBCS). The 2004/5 year would be year 1, 2005/6 would be year 2 and 2006/7 would be year 3 etc.

54 Any output from the 2004/5 BLD survey can be matched back to the 2001 administrative data (BIT, BAS and ABR, as well as trade data). In effect this means that by linking the administrative to the BLD survey data we can construct a *virtual* longitudinal database after the first BLD survey data is available for release.

Recommendation 5: That ESSCC endorse reference period of 2004/05 (dispatch in Feb 2006) as a the most appropriate date for the new business longitudinal survey.

Business Survey Methodology
Methodology Division

Business Demographics and Frames Section
Integration, Coordination and Innovation Branch
Economic Statistics Group

October 2004

Appendix 1 - Key User Requirements

User Requirements.XLS

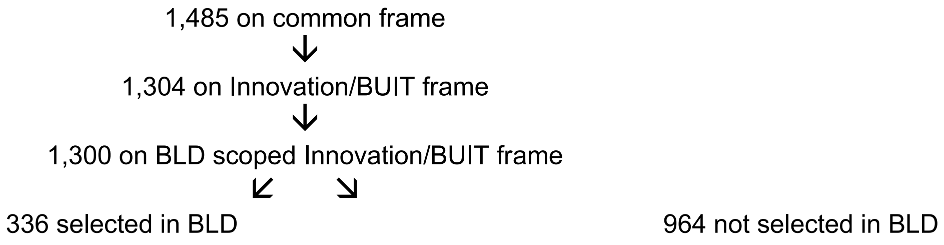
Appendix 2 - Relationship of example selections at the TAU level to the EG unit

Tables 5, 6 and 7 above use an example selection of units taken from Innovation and BUIT at the TAU level. It is of interest to determine how this relates to the EG level unit.

Of the 2,013 TAU units selected, 508 are in the ABSMP and the remaining 1,505 are in the ATOMP (and so by definition are an EG unit). Of the 508 ABSMP TAUs, 171 are one-TAU EG units (based on information from the Common Frame). The remaining 337 TAUs belong to 240 EG units.

Figure 1 shows all 1,485 TAUs on the common frame that are associated with these 240 EG units. This reveals that of the 1,300 TAUs in-scope of the BLD, 336 were selected and 964 were not. Of these 336 TAUs selected, 40 belong to EGs where all the in-scope TAUs were selected - 32 of which are EGs with only 1 in-scope TAU, and the remaining 4 of which are EGs with 2 in-scope TAUs. Details of the 204 EGs where not all in-scope TAUs are selected are given in Table A1_1 below.

Figure 1 : TAUs from multi-TAU EGs included in example BLD sample



1,485 on common frame

1,304 on Innovation/BUIT frame

1,300 on BLD scoped Innovation/BUIT frame

336 selected in BLD

964 not selected in BLD

Table A2_1 details the 204 EGs for which not all the in-scope TAUs are selected in the BLD sample. As this shows, the majority of these EGs have only 2 to 5 TAUs (128 of the 204 EGs). There are only 4 EGs with greater than 20 TAUs associated, the largest of which has 62 TAUs. The final row in the table shows the average proportion of the available TAUs included in the BLD sample, which generally decreases as the number of TAUs associated with the EG increases. (Note this is an average over EG level proportions).

Table A2_1 : EGs where not all TAUs are selected

	# of TAUs in the EG in scope of the BLD								
# of Selected TAUs	2-3	4-5	6-7	8-9	10-15	16-20	21-30	30+	Total
1	68	40	18	7	11	3	1	0	148
2	6	11	6	4	6	1	1	0	35
3	0	3	4	2	3	0	0	1	13
4	0	0	0	1	2	1	0	0	4
5	0	0	0	0	0	0	1	0	1
6	0	0	0	0	1	2	0	0	3
Total	**74**	**54**	**28**	**14**	**23**	**7**	**3**	**1**	**204**
Average Proportion	**0.46**	**0.30**	**0.23**	**0.21**	**0.16**	**0.17**	**0.09**	**0.05**	**0.32**

The largest size group used in this prototype is 200+. Given that these large units will be the most complicated to follow due to structural changes, dependent on the user requirements for large businesses one option may be to include only businesses up to employment of 500. Table A2_2 below shows the distribution of the large businesses by 200-499 and 500+. This reveals that 796 of the 1,958 are larger than 500 employment. Note this would only be a sensible option if it were the really large businesses that underwent the most frequent and complex changes.

Table A2_2 : Large businesses (200+) in-scope of the BLD by size

	B	C	D	E	F	G	H	I	J	K	L	O	P	Q	Total
200-499	32	351	10	38	136	88	73	63	6	76	207	37	27	18	1162
500+	35	241	9	37	64	85	28	47	8	53	118	30	27	14	796
Total	67	592	19	75	200	173	101	110	14	129	325	67	54	32	1958

Appendix 3 : 20 Common Units in each ANZSIC Division by Size Group

Note that the tables in this appendix are numbered to correspond with tables in the main text.

Before re-producing the tables in the main text for 20 common units after 5 years rather than 30, the following graph is presented to show the relationship between the common sample and steady state sample sizes with an 8% attrition rate assumed. There is a relatively straight-line relationship between the two, with the steady state sample approximately 6 times the common sample. To generate the total BLD sample size this steady state sample size must be multiplied by the number of ANZSIC Division by Size groups (48 above).

Graph 3 : Steady State sample required to achieve common sample
(8% attrition per year assumed)

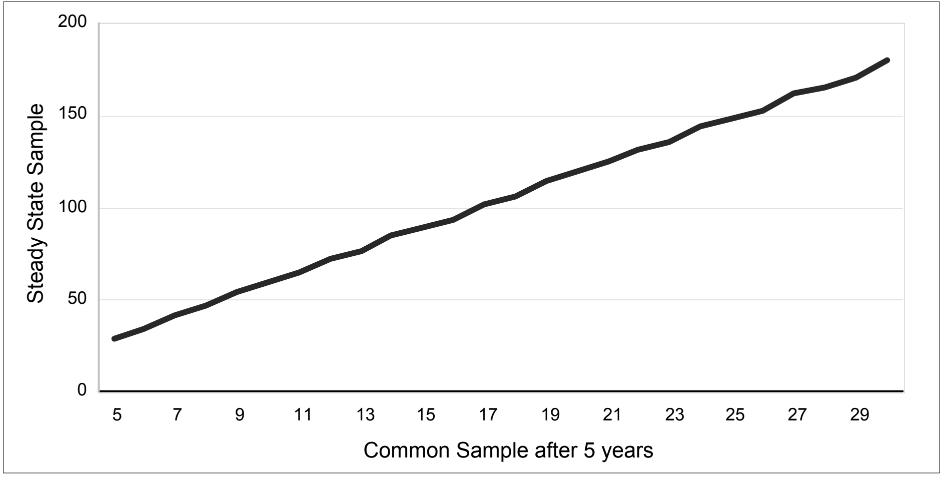


Table A3_2 shows that for 20 common units in a given ANZSIC by size group the initial sample required is 28 (compared to 42 for 30 common units) and the steady state sample size is 120 (compared to 179 for 30 common units).

Table A3_2 : BLD structure for 20 common units in each ANZSIC by size group

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
Panel 1 ->	28	26	24	22	20					
Panel 2 ->		28	26	24	22	20				
Panel 3 ->			28	26	24	22	20			
Panel 4 ->				28	26	24	22	20		
Panel 5 ->					28	26	24	22	20	
Panel 6 ->						28	26	24	22	20
Panel 7 ->							28	26	24	22
Panel 8 ->								28	26	24
Panel 9 ->									28	26
Panel 10 ->										28
Total Sample Size	**28**	**54**	**78**	**100**	**120**	**120**	**120**	**120**	**120**	**120**

Assuming the same 50 ANZSIC by size groups are included the total common sample size will be 1,000 (compared to 1,500 for 30 common units). Table A3_3 shows this over time, where each year a new panel of 1,400 units (2,100 for 30 common) would be added, and the steady state sample size is 5,760 (compared to 8,950 for 30 common units).

Table A3_3 : BLD sample for 20 common units in each ANZSIC by size group

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
Panel 1 ->	1400	1300	1200	1100	1000					
Panel 2 ->		1400	1300	1200	1100	1000				
Panel 3 ->			1400	1300	1200	1100	1000			
Panel 4 ->				1400	1300	1200	1100	1000		
Panel 5 ->					1400	1300	1200	1100	1000	
Panel 6 ->						1400	1300	1200	1100	1000
Panel 7 ->							1400	1300	1200	1100
Panel 8 ->								1400	1300	1200
Panel 9 ->									1400	1300
Panel 10 ->										1400
Total Sample Size	**1400**	**2700**	**3900**	**5000**	**6000**	**6000**	**6000**	**6000**	**6000**	**6000**

Table A3_6 shows the original 1,344 units over five years, and reveals that again, it is only the large CEd units which will still be in either Innovation or BUIT in Years 4 and 5.

Table A3_6 : BLD Panel 1 - units remaining in Innovation or BUIT survey over time (20 common per ANZSIC by size group)

		Year 1	Year 2	Year 3	Year 4	Year 5
0-4		392	364	333		
5-19		336	312	282		
20-199		308	286	241		
200+		308	286	264		
Continuing	*Total*	*1344*	*1248*	*1120*	*242*	*220*
0-4			30	60	90	120
5-19			26	52	78	104
20-199			22	44	66	88
200+			22	44	66	88
Died	*Total*		*100*	*200*	*300*	*400*
0-4		28	26	27	330	300
5-19		28	26	30	286	260
20-199				23	242	220
200+				0		
Rotated out and Education	*Total*	*56*	*52*	*80*	*858*	*780*
Total		**1400**	**1400**	**1400**	**1400**	**1400**

Appendix 4 : Non-employers

The definition of a non-employer used by the Economic Activity Survey (EAS) is a unit with a non-active ITW role but at least one of either a GST or ITI role active. Any unit with a SISCA of 3000 is excluded. The following table shows the population of non-employers in EAS by certain ANZSIC Divisions (core EAS industries) based on the June 2004 frame. In total there are 1,914,471 non-employers in these twelve ANZSIC Divisions. Note that ANZSIC Divisions B (Mining), C (Manufacturing) and D (Electricity, Gas & Water) are not included in Table A4_1, while ANZSIC Division N (Education) is.

Table A4_1 : EAS non-employer population

ANZSIC Division	E	F	G	H	I	J	K	L	N	O	P	Q
Population	348849	56853	147520	28964	114562	28443	366454	550395	27700	75142	76699	92890

If the non-employers were to be included in the BLD a common sample over 5 years of 30 could be taken from each ANZSIC Division to correspond with the employing segment of the population. If all industries were included (ie those in the above table plus B, C and D) this would add a further 450 units to the common sample in each panel. If interest were in only a few industries - say construction - then sample from only these would be added.

Appendix 6 : BLD Survey timing

BLD timeline-1.123