TECHNICAL PAPER

NCVER

Estimation of apprentice and trainee statistics

Brian Harvey
National Centre for Vocational Education Research (NCVER)

INFORMING POLICY AND PRACTICE
IN AUSTRALIA'S TRAINING SYSTEM

Publisher's note

The views and opinions expressed in this document are those of the author and do not necessarily reflect the views of the Australian Government or state and territory governments.

TD/TNC 100.15

Published by NCVER, ABN 87 007 967 311

Level 11, 33 King William Street, Adelaide SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

Phone +61 8 8230 8400 **Email** ncver@ncver.edu.au
Web <https://www.ncver.edu.au> <http://www.lsay.edu.au>

Contents

Tables and figures

Tables

Figures

Glossary of terms

Collection:	a set of data that is submitted quarterly by the state and territory training authorities to NCVER. The data are cumulative from a set point in time (currently data are submitted from 1 July 2002 to the current date). Collections are numbered sequentially from the first collection submitted to NCVER in 1994. The March 2009 quarterly publication uses data from the June 2009 collection, which is collection number 60.
Reporting lag:	the delay in time between an event (commencement, completion, etc.) occurring and the reporting of that event to NCVER via the state and territory training authority data submissions.
Lag ratio:	the ratio of the actual number of events of a given type that occurred in a particular quarter to the number of those events that have been reported at a given quarter.
Expired contracts:	contracts where the expected completion date has passed, but no outcome has been recorded as yet.
In-training:	the number of apprentices and trainees who have commenced, recommenced or suspended their contracts, where their contracts have not yet been completed, cancelled/withdrawn or expired.
Adjustment of estimates:	involves a modification of the estimate produced and occurs where the pattern of reporting lags shows a clear departure from the assumptions of the estimation model.

Estimation of apprentice and trainee statistics

Introduction

Apprentice and trainee data are reported by the state and territory training authorities to the National Centre for Vocational Education and Training (NCVER) on a quarterly basis, starting at the September quarter of 1994. The set of data submitted that quarter is referred to as collection 1. The sets of data submitted in subsequent quarters are referred to as collection 2, collection 3 and so on. At the time of writing, the set of data being submitted is for the June 2009 quarter and is referred to as collection 60. Of particular interest are the numbers of contracts of training that commence, complete, cancel/withdraw, recommence, expire or suspend, and the time at which these events occur (referred to as the 'date of effect'). From these events, the number of contracts 'in-training' at a given time can be calculated.

Between the occurrence of one of these events and the appearance of the corresponding datum in the national collection, there is a chain of administrative processes that must be followed. It is not unusual for some time to elapse before information about events appears in the national collection. These time gaps are referred to as 'reporting lags'. Thus, data about events occurring in a given quarter might require several collections to be completely reported. As a result, accurate counts take time to accumulate. However, waiting for all the data to be submitted reduces their usefulness. In order to get timely information that can be used for monitoring apprentice/trainee activity and formulating policy, a reliable estimate of the final counts is required as soon as possible after the quarter in which the events occur.

The current method of calculating estimates from the National Apprentice and Trainee Collection was endorsed in September 2004 by the National Training Statistics Committee. The purpose of this technical paper is to describe:

- the way data accumulate over many collections
- how items derived from the data change as the data accumulate
- the endorsed estimation method (developed from the analysis of the above)
- the formulae for calculating the required estimates
- potential weaknesses in the method.

Throughout the main text of this document, data for New South Wales commencements and expired contracts are used as examples to illustrate the concepts being discussed. Appendix A contains example data for other events (completions, cancellations etc.) and other states/territories. The reader can use this data to verify that the properties discussed are not peculiar to the examples.

The estimates produced by the endorsed methodology are subject to review. Estimates that are associated with high relative errors or are unusually high or low are examined and if possible adjusted. Documentation relating to the review for collection 60 (June quarter 2009) and subsequent collections can be found in the Adjustment notes for apprentice and trainee estimates document, which is provided as a supporting document for the current collection.

The accumulation of data over time

As stated above, the data for events that occur in a particular quarter are submitted over several collections. This means that, in any collection, data for the most recent quarters will undercount the amount of activity that occurred. Conversely, data for earlier quarters will be progressively closer to being a complete count of the activity that occurred.

Table 1 New South Wales—reported counts of commencements

	Quarter commencement occurred												
Collection reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
47 (Mar-06)	17179												
48 (Jun-06)	24645	9899											
49 (Sep-06)	25429	14444	10286										
50 (Dec-06)	25664	15325	16190	10408									
51 (Mar-07)	25783	15515	16955	16168	15684								
52 (Jun-07)	25834	15605	17260	17247	25641	11350							
53 (Sep-07)	25861	15631	17344	17482	26656	16336	13204						
54 (Dec-07)	25869	15646	17360	17555	26904	16813	18362	10237					
55 (Mar-08)	25871	15649	17363	17572	27027	16995	19168	16827	16384				
56 (Jun-08)	25879	15651	17364	17577	27068	17069	19466	17906	26677	12103			
57 (Sep-08)	25878	15655	17367	17578	27082	17088	19534	18126	27956	20684	11107		
58 (Dec-08)	25874	15652	17368	17580	27087	17094	19559	18192	28311	21775	19358	8831	
59 (Mar-09)	25871	15650	17367	17578	27084	17096	19562	18217	28406	**22030**	**20357**	**16803**	15454

Table 2 New South Wales—lag ratios (inflation factors) for commencements

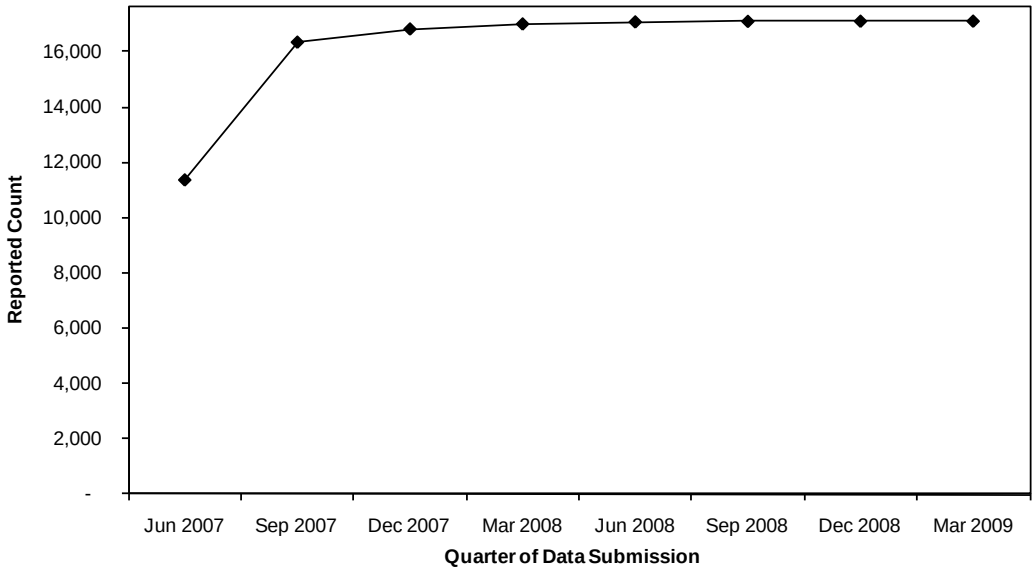
	Quarter commencement occurred												
Collection reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sept-08	Dec-08	Mar-09
47 (Mar-06)	1.5059666												
48 (Jun-06)	**1.0497464**	1.5809678											
49 (Sep-06)	**1.0173817**	**1.0834949**	1.6884114										
50 (Dec-06)	**1.0080658**	**1.0212072**	**1.0726992**	1.6888932									
51 (Mar-07)	1.0034131	**1.0087013**	**1.0242996**	**1.0872093**	1.7268554								
52 (Jun-07)	1.0014322	1.0028837	**1.0061993**	**1.0191917**	**1.0562771**	1.5062555							
53 (Sep-07)	1.0003867	1.0012155	1.0013261	**1.0054914**	**1.0160564**	**1.046523**	1.4815208						
54 (Dec-07)	1.0000773	1.0002557	1.0004032	1.0013102	**1.0066905**	**1.0168322**	**1.0653524**	1.7795253					
55 (Mar-08)	1	1.0000639	1.0002304	1.0003415	1.002109	**1.0059429**	**1.0205551**	**1.0826053**	1.7337646				
56 (Jun-08)	0.9996909	0.9999361	1.0001728	1.0000569	1.0005911	1.0015818	**1.0049317**	**1.0173685**	1.0648124	1.8202099			
57 (Sep-08)	0.9997295	0.9996806	1	1	1.0000738	1.0004682	1.0014334	**1.0050204**	1.0160967	1.0650745	1.8328081		
58 (Dec-08)	0.9998841	0.9998722	0.9999424	0.9998862	0.9998892	1.000117	1.0001534	1.0013742	1.0033556	1.0117107	1.0516066	1.902729	
59 (Mar-09)	1	1	1	1	1	1	1	1	1	1	1	1	1

The way the data accumulate over time displays some consistent patterns that recur (with some variation) from collection to collection. The existence of the consistent patterns allows the inference (i.e. estimation) of complete counts for the recent quarters (whose reported counts are too low). The amount of variation in the patterns is an indicator of the likely error in the estimate.

The following example uses real data (refer to table 1, the column circled in red) to illustrate how data accumulates over time. In the data submitted in collection 52 (June 2007), New South Wales reported that 11 350 commencements occurred in the June 2007 quarter. Over the following collections, more information about the June 2007 quarter was processed. In collection 53, New South Wales reported that 16 336 commencements had occurred in the June 2007 quarter (a 44% increase over the initial count). Reading down the column, it can be seen that the reported count continues to increase every quarter, although the size of the increase diminishes with each quarter.

Figure 1 displays how, after an initial period of growth, the count stabilises over time to a more or less constant value. There is about a 50% increase from the initial count of 11 350 to the count of 17 096 reported in collection 59. The other important feature to notice is that it is in the second year of submissions that the graph is (approximately) horizontal. Therefore, a close approximation to the actual number of commencements occurring in New South Wales during the June 2007 quarter can be obtained after about one year of updates.

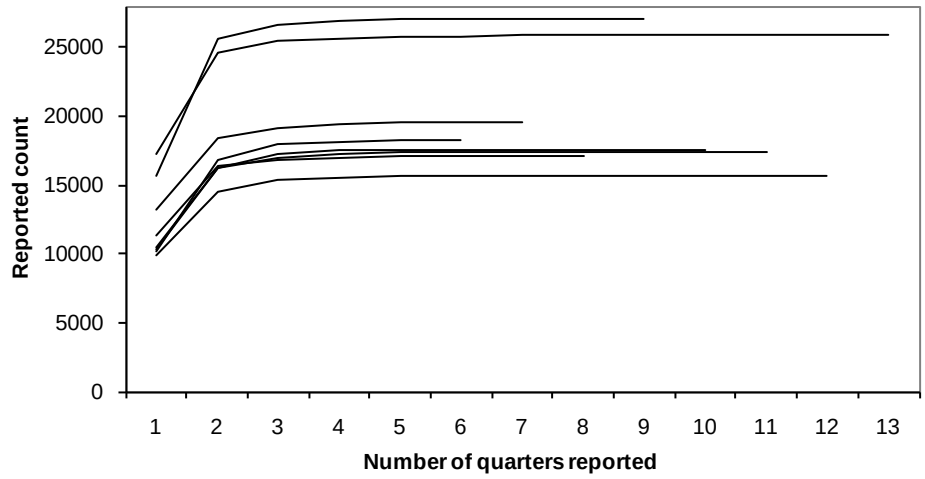
Figure 1 June quarter 2007 commencements in New South Wales—reported count by data submission



The general shape of the graph in figure 1 is typical of the graphs that would be produced for other states/territories, other events (completions, cancellations etc.) and for other quarters. To illustrate this, figure 2 shows the growth curves for the New South Wales commencements that occur in the quarters March 2006 to December 2007. Each line on the graph is derived from the corresponding column in table 1. Note that the purpose of the graph is to emphasise the spread of the curves, rather than identify which quarter gives rise to which curve. For this reason (and the fact that there are so many curves) no legend to identify quarters is included on the graph.

All the curves in figure 2 share the trait of having a steep slope initially but ultimately becoming flat. However, it is not yet clear how this information could be used to estimate actual counts from the initial reported counts. For example, in table 1, the column for the December 2008 quarter shows that the initial count of 8831 grew to 16 803 in the following collection. Figure 2 leads us to expect that the reported counts will eventually converge to a constant value as time passes, but to estimate that value is not yet obvious.

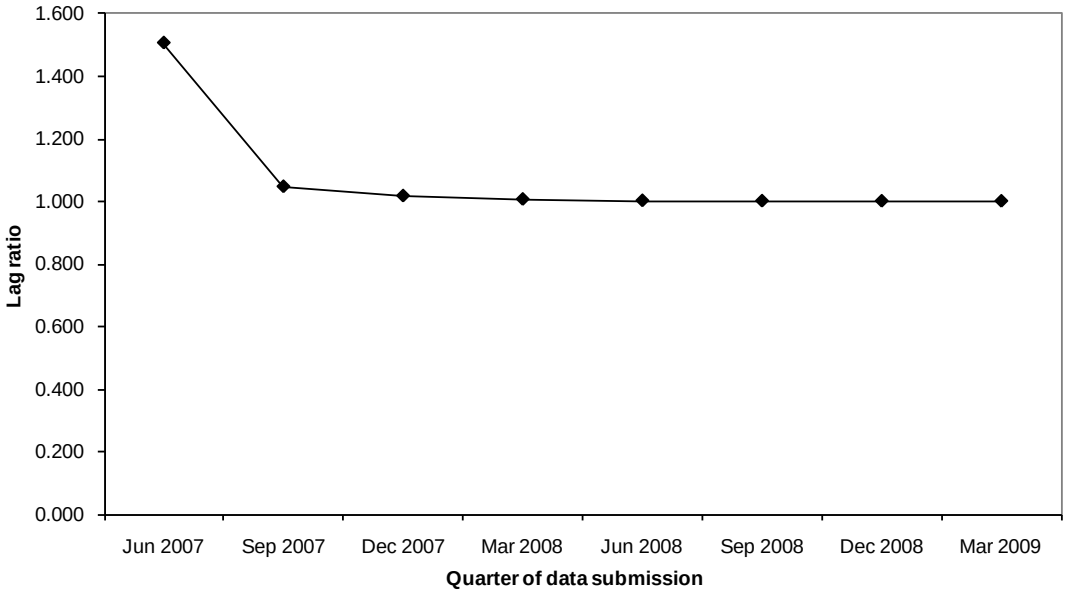
Figure 2 Commencements in New South Wales—reported count by number of quarters reported (March 2006 to December 2007)



In order to progress further, the data in table 1 can be looked at in a different way. Table 2 is derived from table 1 by a simple calculation. The entry in each cell of table 2 is the result of dividing the count in the corresponding cell in table 1 into the count in the cell at the bottom of the same column in table 1. For example, consider the cell in the June 2007 column and the collection 53 row in table 1. The count is 16 336. The count at the bottom of the column (i.e. the collection 59 row) is 17 096. The result of 17 096/16 336 is 1.046523, which becomes the entry for the June 2007 column and the collection 53 row in table 2. This ratio indicates the growth in the count from the September 2007 to March 2009 (i.e. about 4.65%).

There are eight rows of counts in the June 2007 column of table 1. The last row in that column corresponds to the last point of the growth curve in figure 1. Given how flat the graph is at this point, it is safe to assume that the ratio (1.046523) represents the growth in the reported count from collection 53 to the actual value of the count (or more correctly, a very close approximation to it).

Figure 3 June quarter 2007 commencement lags in New South Wales—lag ratios by data submission



Due to the previously discussed reporting lags, the counts in table 1 tend to increase from the top to the bottom of a column. Therefore, the ratios in table 2 tend to decrease from the top to the bottom of a column. The ratios are referred to as 'lag ratios' due to their relationship to reporting lags in the data. Figure 3 is the lag ratio equivalent of figure 1.

Since the calculation for every cell of table 2 involves the last row of table 1, the last row of table 2 contains only the value 1. Those columns (quarters) that have had over a year of updates can be considered close enough to reporting actual counts in the last row of table 1. Thus, in table 2, the ratios calculated for those columns can be considered as reliable indicators of how much data is still to be collected before the actual counts are achieved. Conversely, the columns that have had fewer updates cannot be considered close to reporting actual counts in the last row of table 1. Thus, in table 2, the ratios calculated for those columns will change noticeably over time and so are not useful for estimation purposes.

Figure 4 Commencement lags in New South Wales—lag ratios by number of quarters reported (March 2006 to December 2007)

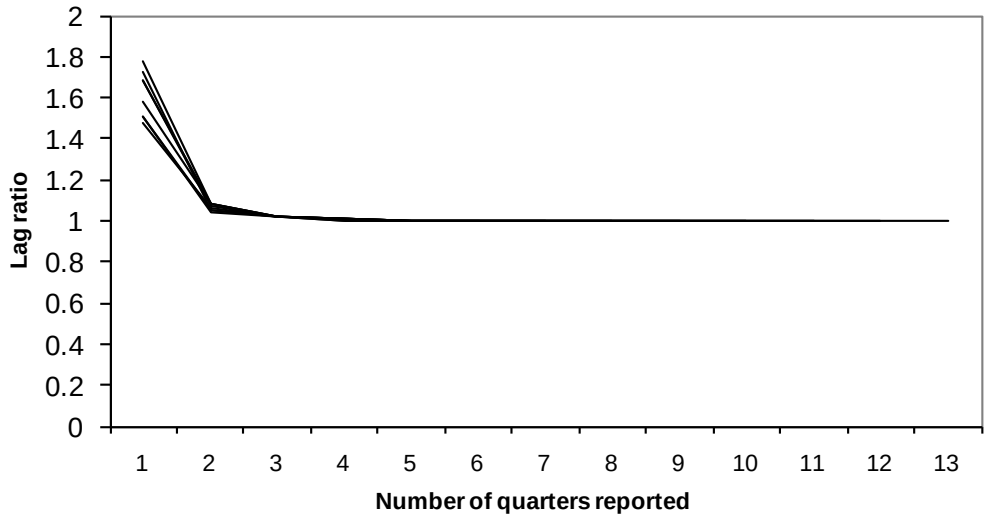


Figure 4 is the lag ratio equivalent of figure 2. Each line on the graph is derived from the corresponding column in table 2. As before, the key feature is the spread of the lines. There is quite a lot of variability in the lag ratios corresponding to the initial counts for the quarters. The graph implies that the initial count reported for a given quarter could eventually grow by about 50–80% before reaching the actual count.

The good news is that the lines rapidly converge so that by the second time a quarter is reported, the lines are already becoming difficult to distinguish on the graph. This implies that there is a lot less variability in the lag ratios at this stage. Examining table 2, it can be seen that these lag ratios (highlighted in red) vary from a low of 1.046523 to a high of 1.0872093. This implies that the second count reported for a given quarter could eventually grow by anything between 4.6 and 8.7% before reaching the actual count.

The news gets even better. The lines converge towards a value of one, showing less variation with each update. This pattern suggests a method for calculating useable estimates of the actual count for a quarter in a relatively quick time frame. The amount of variation associated with the lag ratios for the first quarter reported is too high to produce reliable estimates. However, the variation of lag ratios associated with the second quarter are much lower. By taking the second count reported for a quarter and multiplying it by the average of the second quarter lag ratios, an estimate of the actual count for that quarter can be calculated. The relative error in this estimate is the same as the relative error in the estimate of the average of the second quarter lag ratios.

Similarly, the variation of lag ratios associated with the third and fourth quarter reported counts are successively lower. Using the appropriate average lag ratio to multiply the third and fourth reported counts for a quarter will lead to successively better estimates of the actual count. Beyond the fourth update, the reported counts are close to final and it is doubtful that an estimate could be produced that would be better than just using whatever value the reported count has at that time. In addition to producing estimates of the actual counts, the variability in the lag ratios can be used to calculate the likely variability of those estimates.

These concepts will be discussed with more precision in the sections that discuss the estimation method.

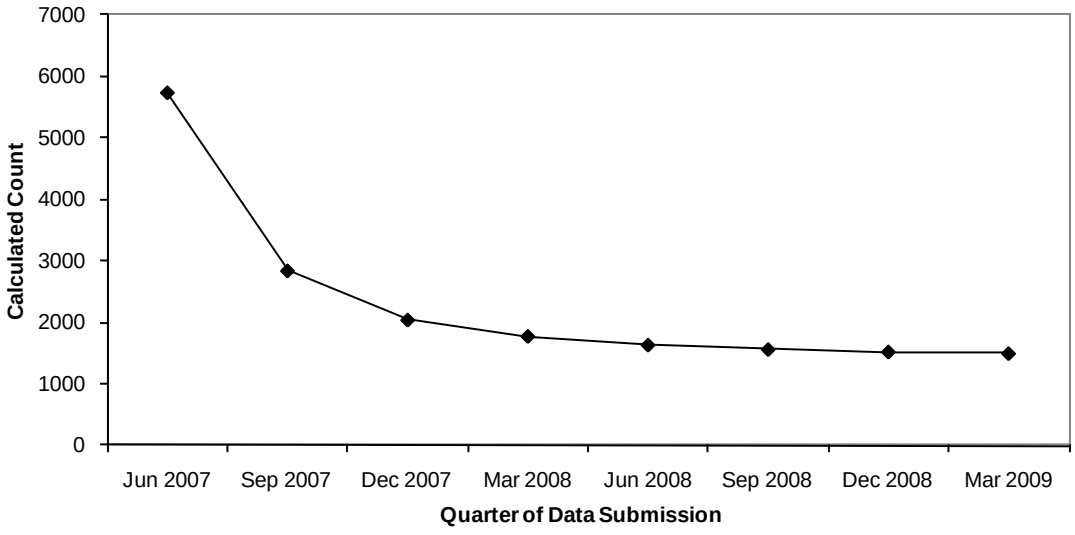
So far, New South Wales has been used to illustrate a possible way to estimate the number of commencements that occur in a quarter with only a one-quarter delay, rather than wait over a year for the reported counts to become reliable. Fortunately, the general features described above are common across states and hold true for data relating to completions, cancellations and withdrawals, differing in minor details only. The most significant difference relates to cancellations and withdrawals. It tends to take longer for the reported counts to approach the actual counts—about two years rather than one.

Derived items

Thus far we have examined the properties of the data that are submitted to NCVER. An important statistic required by researchers and policy-makers alike is the number of contracts which are active at the end of a given quarter; that is, the number in-training. This is not submitted to NCVER but is calculated from the data that are submitted. If any of the data items used in the calculation cannot be considered as actual counts, then the calculated in-training statistic cannot be considered as actual. The actual number in-training can be estimated simply by using estimates for the required data items in the calculation. This concept will be discussed with more precision in the sections that discuss the estimation method.

The calculation for in-training is very simple. The number in-training at the end of any quarter is the number in-training at the end of the previous quarter, plus the number that start training during the quarter, minus the number that cease training during the quarter. Commencements and recommencements give the number that start training in the quarter. Completions, cancellations, withdrawals and expiries give the number that cease training in the quarter.

Figure 5 June quarter 2007 expired contracts in New South Wales—calculated count by data submission

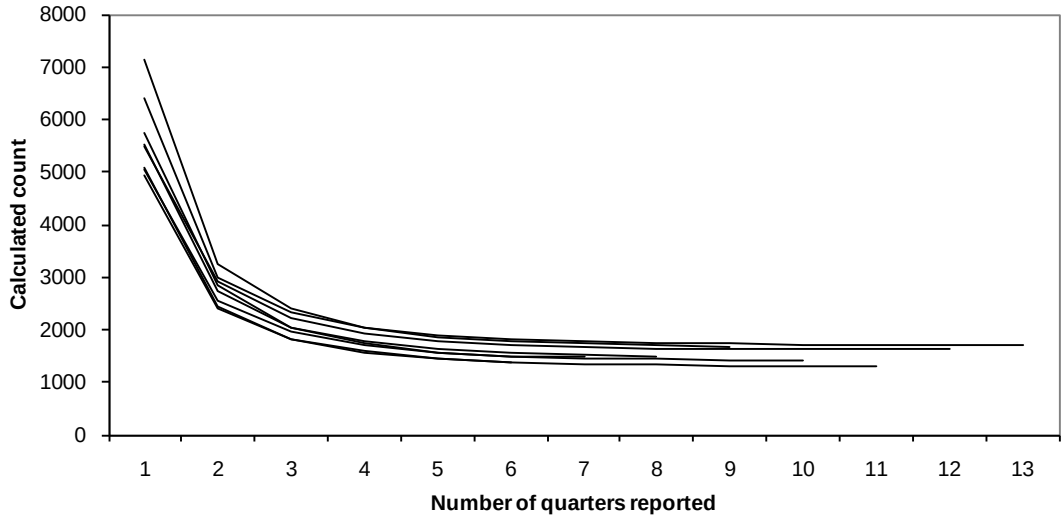


The number of expiries is also a derived item. A contract becomes expired when the expected completion date has passed but no information has been received to indicate the current status of the training. This item is automatically derived in the apprentice and trainee database. Expired contracts for a given quarter, like commencements, completions etc., also appear in the collection with an initial value, and with the passage of time approach a final value. However, unlike the submitted items, the numbers decrease with time rather than increase (see table 3). This is not surprising since, with the passage of time, more information about contracts becomes available and therefore the number of contracts with unknown outcomes becomes less.

Since expiries do not accumulate the same way that the other items do, it is instructive to step through an example as was done for commencements. Table 3 displays the counts for the number of contracts that are expired at the end of a given quarter by successive submissions of data. Taking June 2007 as an example (the column ringed in red), it can be seen that the number of expired contracts is initially calculated to be 5731. This number decreases every quarter until, at the bottom of the column, the number has reduced to 1499. Figure 5 displays how, after an initial period of steep decline, the count stabilises to a more or less constant value.

The general shape of the curve in figure 5 is typical for expired contracts in other states/territories and for other quarters. Figure 6 displays the curves for New South Wales contracts that expire in the quarters March 2006 to December 2007. As in figure 2, the spread of the curves is the key point of the graph, so once again the various quarters are not labelled. It can be seen that the curves all have a common shape.

Figure 6 Expired contracts in New South Wales—calculated count by number of quarters reported



(March 2006 to December 2007)

As was done with the commencements example, it is useful to look at the data in a different way. Table 4 is derived from table 3 by a simple calculation. Because expired contracts decrease over time, the calculation is different from the calculation discussed for commencements. Each cell in table 4 is the result of dividing the count in the corresponding cell in table 3 into the count at the top of the same column in table 3. The ratios, like the counts, decrease from the top to bottom of any column. As before, these ratios are referred to as lag ratios due to their relation to reporting lags.

Figure 7 is the lag ratio equivalent of figure 5. It displays the data from the column ringed in green in table 4. The graph shows that once enough time has passed, the downward trend in the lag ratios diminishes so that the right-hand end of the graph is approximately horizontal.

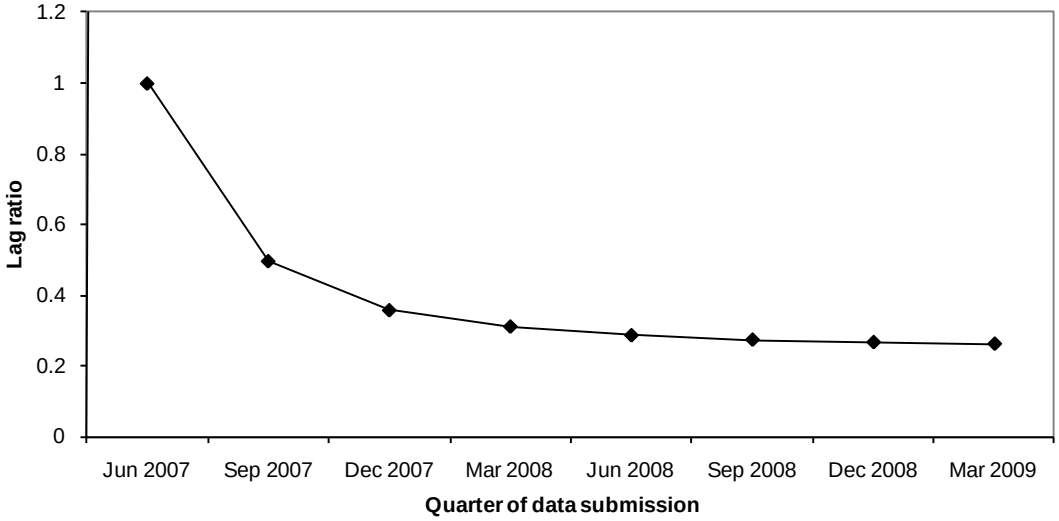
Table 3 New South Wales—calculated counts of expired contracts

	Quarter occurred												
Quarter reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
47 (Mar-06)	6386												
48 (Jun-06)	3004	5480											
49 (Sep-06)	2322	2916	5073										
50 (Dec-06)	2043	2237	2446	5507									
51 (Mar-07)	1899	1910	1814	2740	7142								
52 (Jun-07)	1827	1776	1561	2038	3252	5731							
53 (Sep-07)	1767	1700	1433	1741	2417	2841	5063						
54 (Dec-07)	1745	1670	1370	1578	2026	2043	2562	4924					
55 (Mar-08)	1737	1645	1342	1479	1855	1778	1968	2403	7114				
56 (Jun-08)	1723	1636	1327	1447	1778	1637	1715	1830	3867	6030			
57 (Sep-08)	1716	1628	1313	1439	1740	1565	1558	1587	2752	3190	5664		
58 (Dec-08)	1707	1620	1309	1420	1691	1525	1498	1436	2355	2452	3025	6019	
59 (Mar-09)	1703	1617	1306	1412	1679	1499	1471	1392	2131	2066	2295	3120	6602

Table 4 New South Wales—lag ratios (inflation factors) for expired contracts

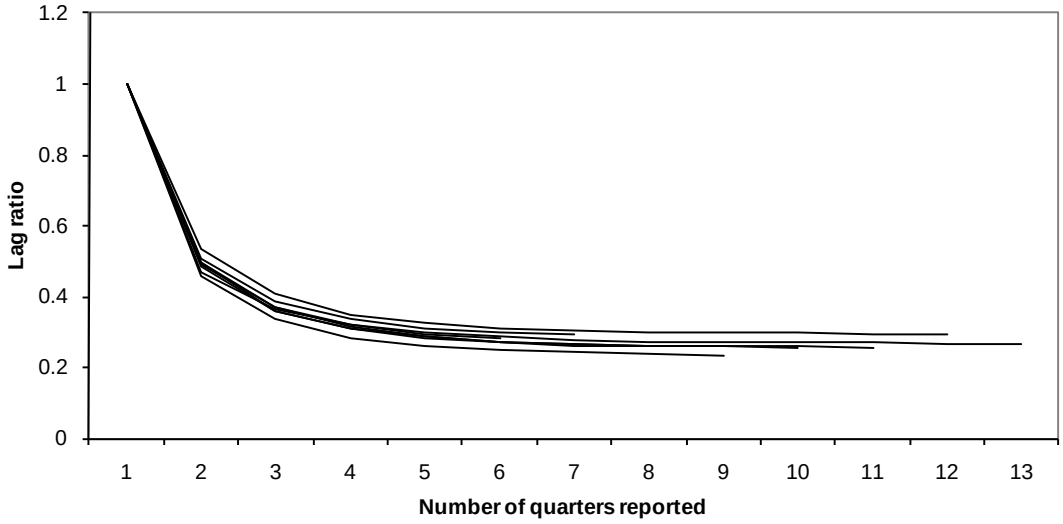
	Quarter occurred												
Quarter reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
47 (Mar-06)	1												
48 (Jun-06)	0.4704	1											
49 (Sep-06)	0.3636	0.5321	1										
50 (Dec-06)	0.3199	0.4082	0.4822	1									
51 (Mar-07)	0.2974	0.3485	0.3576	0.4975	1								
52 (Jun-07)	0.2861	0.3241	0.3077	0.3701	0.4553	1							
53 (Sep-07)	0.2767	0.3102	0.2825	0.3161	0.3384	0.4957	1						
54 (Dec-07)	0.2733	0.3047	0.2701	0.2865	0.2837	0.3565	0.506	1					
55 (Mar-08)	0.272	0.3002	0.2645	0.2686	0.2597	0.3102	0.3887	0.488	1				
56 (Jun-08)	0.2698	0.2985	0.2616	0.2628	0.2489	0.2856	0.3387	0.3716	0.5436	1			
57 (Sep-08)	0.2687	0.2971	0.2588	0.2613	0.2436	0.2731	0.3077	0.3223	0.3868	0.529	1		
58 (Dec-08)	0.2673	0.2956	0.258	0.2579	0.2368	0.2661	0.2959	0.2916	0.331	0.4066	0.5341	1	
59 (Mar-09)	0.2667	0.2951	0.2574	0.2564	0.2351	0.2616	0.2905	0.2827	0.2996	0.3426	0.4052	0.5184	1

Figure 7 June quarter 2007 expired contract lags in New South Wales—lag ratios by data submission



To illustrate that the general shape of the graph in figure 7 is typical, figure 8 displays the curves for several quarters for New South Wales. It can be seen that the general shape is the same for all the quarters represented on the graph. The curves indicate that number of expired contracts will eventually diminish to around 25–30% of the number originally calculated.

Figure 8 Expired contract lags in New South Wales—lag ratios by number of quarters reported (March 2006 to December 2007)



The method: Overview

The above discussion gives enough understanding of the National Apprentice and Trainee Collection to now describe the estimation methodology decided upon in 2004.

The quarterly and annual publications report on commencements, completions, cancellations/ withdrawals and in-training. Recommencements and expired contracts are not reported, but are included in the calculations for in-training.

As previously mentioned, in-training is derived by a simple calculation. For any given quarter, to get the number of contracts in-training at the end of the given quarter, simply use the previous quarter's calculated in-training and add the number of contracts that start training in the given quarter, and subtract the number of contracts that cease training in the given quarter.

After a thorough examination and discussion of the lag ratios, the decisions described below were made.

- For the quarter being collected for the first time, there is too much variability in the lag ratios to allow reliable estimation. Thus there is a one-quarter delay between a quarter first being in the collection and when information about that quarter is first reported in the quarterly publication.[1]
- Commencements take about a year to accumulate to a value which can be considered close enough to be treated as actual. Hence for a given quarter, the second, third and fourth times commencements are published they are estimated. After that, the reported counts are published.
- Recommencements are relatively small compared with commencements. It was decided that there was not enough to be gained from estimation relative to just using counts. Therefore, only the latest counts are used.
- Completions take about a year to accumulate to a value which can be considered final. Hence for a given quarter, the second, third and fourth times completions are published they are estimated. After that, the latest counts are reported.
- Cancellations/withdrawals take about two years to accumulate to a value which can be considered close enough to be treated as actual. Hence for a given quarter, the second to the eighth times completions are published they are estimated. After that, the reported counts are published.
- Expired contracts take about two years to diminish to a value which can be considered final. Hence for a given quarter, the second to the eighth times expired contracts are published they are estimates. After that, the published values are derived from reported data.

A direct consequence of these decisions is that calculations of in-training for quarters being published for the second to the eighth time are considered to be estimates of actual values. For quarters earlier than this, the calculations are taken to be actual.

The estimates for commencements, completions, cancellations/withdrawals and expired contracts are simple in concept. For any quantity that is to be estimated, just take the appropriate count from the latest data available and multiply it by a factor that represents the average lag ratio appropriate to the quarter being estimated.

As an example, refer back to tables 1 and 2, where the latest data presented is collection 59 (March 2009). December is reported for the second time. The latest count for December 2008 is 16 803 (highlighted in red in table 1). As previously discussed, history suggests this number will eventually grow by anywhere between 4–10%. The figures shaded in red in table 2 represent a sample of lag ratios for quarters being reported for the second time. The average of these lag ratios (1.068) represents a typical growth that could be expected for the December 2008 quarter. The estimate is therefore 16 803 x 1.068 = 17 945. The estimate is subject to a variance which is proportional to the variance in the lag ratios. Similar steps can be applied for the September 2008 (blue highlight) and June 2008 (yellow highlight) quarters.

The remaining issue to be discussed is how to determine which quarters to use when calculating an average lag ratio. The further back in the past a quarter is, the more likely it is that the corresponding lag ratios are a reliable representation of the way data accumulated over time. However, processes can change over time, which could mean lag ratios from quarters too far in the past might not reflect the way data will accumulate for recent quarters. The challenge is to select a set of quarters far enough in the past to be considered final but as recent as possible. This

1 With the passage of time, the demand for reporting at least some information about a quarter when it was first collected has increased. In response to this demand, NCVER has determined that some trend estimates at the Australian level of aggregation can be released for this quarter. The trends are produced by first using the method described in this paper to calculate a quarterly time series of estimates up to and including the quarter being reported for the first time. The trend estimate for this quarter is calculated by seasonally adjusting and smoothing this series of estimates.

also has an impact on how many quarters to include in the sample. The decisions made are described below.

- Completions and commencements each take about a year before approaching a final value. The time period from five to 12 quarters prior to the latest quarter is used for taking samples of lag ratios.
- Cancellations/withdrawals and expired contracts each take about two years before approaching a final value. The time period from nine to 16 quarters prior to the latest quarter is used for taking samples of lag ratios.

The method: Detail

The above discussion has covered the principles underlying the estimation methodology. This section details the formulae used to calculate the estimates and can be safely ignored by those readers who do not wish/need to deal with such things.

Calculate the lag ratios

For commencements and completions we have:

$$LR_{i,j} = \frac{C_{0,j}}{C_{i,j}}$$

where

$C_{x,y}$ is the count reported in x quarters ago, for events occurring y quarters ago,

$LR_{x,y}$ is the lag ratio corresponding to the count reported in x quarters ago, for events occurring y quarters ago,

i = 0, 1, 2, 3, … , j

j = 5, 6, … , 12.

For cancellations/withdrawals we have:

$$LR_{i,j} = \frac{C_{0,j}}{C_{i,j}}$$

where

$C_{x,y}$ is the count reported in x quarters ago, for events occurring y quarters ago,

$LR_{x,y}$ is the lag ratio corresponding to the count reported in x quarters ago, for events occurring y quarters ago,

i = 0, 1, 2, 3, … , j

j = 9, 10, … , 16.

For expired contracts we have:

$$LR_{i,j} = \frac{C_{i,j}}{C_{j,j}}$$

where

$C_{x,y}$ is the count calculated x quarters ago, for expiries occurring y quarters ago,

$LR_{x,y}$ is the lag ratio corresponding to the count calculated x quarters ago, for expiries occurring y quarters ago,

i = 0, 1, 2, 3, … , j

j = 9, 10, … , 16.

For example, in table 1 the count in the March 2006 column (j = -12) and the June 2006 row (i = -11) is 24 645. The count at the March 2009 row (i.e. the zero row) of that same column is 25 871. The lag ratio corresponding to i = -11 and j = 12 is therefore 25 871/24 645 = 1.0497, which is the number in the first column and second row of table 2.

Calculate the average of the lag ratios

For commencements and completions we have:

$$\overline{LR_i} = \frac{\sum_{j=5}^{12} LR_{j-i,j}}{8}$$

where

$\overline{LR_i}$ is the average of the lag ratios for quarters being reported for the ith time,

$LR_{x,y}$ is the lag ratio corresponding to the count reported in x quarters ago, for events occurring y quarters ago,

i = 1, 2, 3

j = 5, 6, … , 12.

For cancellations/withdrawals we have:

$$\overline{LR_i} = \frac{\sum_{j=9}^{16} LR_{j-i,j}}{8}$$

where

$\overline{LR_i}$ is the average of the lag ratios for quarters being reported for the ith time,

$LR_{x,y}$ is the lag ratio corresponding to the count reported in x quarters ago, for events occurring y quarters ago,

i = 1, 2, … ,7

j = 9, 10, … , 16.

For expired contracts we have:

$$\overline{LR}_i = \frac{\sum_{j=9}^{16} LR_{0,j}}{8}$$

where

$\overline{LR}_i$ is the average of the lag ratios for quarters being reported for the i[th] time. Note that this is the same value for all of the quarters being estimated.

(Recall that for expired contracts, we are only interested in the proportion of the original count that remains.)

$LR_{x,y}$ is the lag ratio corresponding to the count reported in x quarters ago, for events occurring y quarters ago (x = 0 implies the current quarter).

i = 1, 2, … ,7

j = 9, 10, … , 16.

As an example, in table 1, the September quarter is being reported for the third time in collection 59 (the number highlighted in blue). This is two quarters before March 2009 (the current quarter), so i = 2. The corresponding lag ratios in table 2 are therefore the rows that correspond to two quarters below the top entry in the columns for the months March 2006 (12 quarters prior to current) and December 2007 (5 quarters prior to current). Refer to the blue highlighted cells in table 2.

Calculate the variance of the lag ratios and the average of the lag ratios

The variances of the lag ratios and the average lag ratio are calculated using standard formulae,

$$\text{var}(LR_i) = (\sum_{j=5}^{12} (LR_{j,j-i} - \overline{LR}_i)^2) / 7$$

$$\text{var}(\overline{LR}_i) = \text{var}(LR_i) / 8$$

where

$\text{var}(LR_i)$ is the estimated variance of the lag ratios,

$\text{var}(\overline{LR}_i)$ is the estimated standard deviation of the average of the lag ratios.

Other symbols as before.

Note that

$$SD(LR_i) = \sqrt{\text{var}(LR_i)}$$

is the standard deviation of the lag ratios.

Calculate the estimates

For commencements, completions and cancellations/withdrawals we have:

$$E_i = \overline{LR}_i * C_{0,i}$$

where

E_i is the estimate of the 'final' count of activity occurring i quarters prior to the current quarter.

Other symbols as before.

For expired contracts we have:

$$E_i = \overline{LR}_i * C_{i,i}$$

where

E_i is the estimate of the 'final' count of activity occurring i quarters prior to the current quarter.

Other symbols as before.

Calculate the prediction errors of the estimates

In the apprentice and trainee collection the goal of estimation is to predict the final value of the count for a given quarter. We can consider the estimate to be equal to the (unknown) correct value plus some error term. The variance of the estimate is the variance of the error term, which can be re-expressed as the difference between the estimate and the final count. Thus we define:

$$\Pr Err(E_i) = \sqrt{\text{var}(F_i - E_i)}$$
$$= \sqrt{\text{var}(F_i) + \text{var}(E_i)}$$

where

F_i is the final count for events occurring i quarters before the current quarter,

E_i as before.

For commencements, completions and cancellations/withdrawals we continue as follows:

$$\sqrt{\text{var}(F_i) + \text{var}(E_i)}$$
$$= \sqrt{\text{var}(C_{0,i} R_i) + \text{var}(C_{0,i} \overline{LR}_i)}$$
$$= \sqrt{C_{0,i}^2 \text{var}(R_i) + C_{0,i}^2 \text{var}(\overline{LR}_i)}$$
$$= \sqrt{C_{0,i}^2 (\text{var}(R_i) + \text{var}(\overline{LR}_i))}$$
$$= C_{0,i} \sqrt{\text{var}(LR_i) + \frac{\text{var}(LR_i)}{8}} \quad *$$
$$= C_{0,i} \sqrt{\text{var}(LR_i)(1 + \frac{1}{8})}$$
$$= C_{0,i} SD(LR_i) \sqrt{1 + \frac{1}{8}}$$

where

R_i is the (unknown) final lag ratio.

* Note that the final lag ratio is assumed to come from the same population as the sample of known lag ratios. This means the variance of the former is the same as the latter, thus allowing the substitution in this line.

Other symbols as before.

Similarly, for expired contracts we have:

$$\sqrt{\text{var}(F_i) + \text{var}(E_i)}$$
$$= \sqrt{\text{var}(C_{i,i} R_i) + \text{var}(C_{i,i} \overline{LR}_i)}$$
$$= \sqrt{C_{i,i}^2 \text{var}(R_i) + C_{i,i}^2 \text{var}(\overline{LR}_i)}$$
$$= \sqrt{C_{i,i}^2 (\text{var}(R_i) + \text{var}(\overline{LR}_i))}$$
$$= C_{i,i} \sqrt{\text{var}(LR_i) + \frac{\text{var}(LR_i)}{8}}$$
$$= C_{i,i} \sqrt{\text{var}(LR_i)(1 + \frac{1}{8})}$$
$$= C_{i,i} SD(LR_i) \sqrt{1 + \frac{1}{8}}$$

All symbols as previously defined.

Calculate the in-training estimates

Having calculated estimates for commencements, completions, cancellations/withdrawals and expired contracts, it is now possible to calculate estimated in-training counts. At this point, a slight change in the notation is needed in order to distinguish between estimates and counts for commencements, completions etc. This is achieved by the addition of a superscript to notation previously used. The equation is:

$$IT_j = IT_{j-1} + E_j^{Commence} + C_{0,j}^{\text{Re}commence} - E_j^{Complete} - E_j^{Cancel/withdraw} - E_j^{Expire}$$

where

IT_x is the estimated final in-training count at the end of the quarter occurring x quarters prior to the current quarter.

Note that the in-training estimate for a given quarter depends on that for the previous quarter. Thus these estimates must be produced in sequence starting at the earliest quarter being estimated and working toward the most recent quarter being estimated. This means that the subscript j follow this sequence,

j = 7, 6, … ,1.

Calculate the prediction errors of the in-training estimates

The prediction errors for the in-training estimates can be found from the prediction errors of the components that form the right-hand side of the equation above.

$$\Pr Err(IT_j) = \sqrt{\begin{aligned}&\Pr Err(IT_{j-1})^2 + \Pr Err(E_j^{Commence})^2 + \Pr Err(C_{0,j}^{\mathrm{Re}commence})^2 \\ &+ \Pr Err(E_j^{Complete})^2 + \Pr Err(E_j^{Cancel/withdraw})^2 + \Pr Err(E_j^{CExpire})^2\end{aligned}}$$

Note that the in-training count for eight quarters prior to the current quarter is considered to be final and therefore the prediction error for that quarter is zero. This allows the calculation to iterate from early quarters to recent quarters as was done previously.

Calculate the prediction intervals of the estimates

After all the above calculations have been made, 95% prediction intervals are calculated. These intervals indicate a range of values that are likely to include the (unknown) final count. For any of the estimates, a 95% prediction interval can be defined as follows:

$$Estimate \pm t_{7,0.975} \Pr Err(Estimate)$$

where

$t_{7,0.975}$ is the upper 2.5% point of the t distribution with seven degrees of freedom.

Potential weaknesses in the method

The method assumes that the way the data accumulated in the past will persist (at least approximately) into the future. In reality, we might expect administrative systems to change (hopefully improve) over time.

If the effect of the change is gradual enough, then it is possible for the estimation method to cope, since the time periods used for calculating average lag ratios (which advance one quarter at a time) will eventually include the quarters in which the change began. Depending on the nature of the change, the result would be to introduce some bias or some extra variability into the estimates. Slow change would mean that the estimates will still be useable.

If the effect of the change is too rapid (and is large), then the time periods used for calculating average lag ratios will be totally inappropriate. The result could be that the introduced bias and/or variability will make the estimates useless or (even worse) misleading. The great danger with rapid change is that it might take some time for the data to reveal that it taken place. Furthermore, the time periods used for calculating average lag ratios only advance one quarter at a time. It will take some time to catch up to where the change occurred.

The problems discussed above are common to any method of estimation that uses existing data to predict future data.

The estimation method also relies on the assumption that, for any given average lag ratio, the lag ratios that are used to calculate that average can be treated as a random sample. Analysis of the lag ratios reveals that sometimes this is not the case, but in most cases the departure from this assumption is not severe. When the effect is large enough to warrant attention, the affected estimates are reviewed. Documentation relating to the review for collection 60 (June quarter 2009), and subsequent collections, can be found in the *Adjustment notes for apprentice and trainee estimates* document, which is provided as a supporting document to the current collection.

The estimates are monitored and estimates made in the past are compared with the values that the corresponding reported counts eventually become. There are a small number of occasions where the method has underperformed but, in general, the estimates tend to be in the correct order of magnitude and the prediction intervals perform well.

Appendix A

Supporting data

The following pages contain tables similar to table 1 in the main text. They are included as supporting evidence for the assertions in the main text referring to the way data accumulates in the National Apprentice and Trainee Collection. These tables will allow interested readers to test these claims for themselves.

The range of tables is only a sample of the set of tables that could have been selected.

Note that all contract cancellations and withdrawals are treated as a combined category.

Victoria: Reported counts of completions

	Quarter completion occurred												
Collection reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
47 (Mar-06)	6772												
48 (Jun-06)	9370	7350											
49 (Sep-06)	10000	9843	6946										
50 (Dec-06)	10196	10268	8889	8049									
51 (Mar-07)	10296	10431	9348	10819	6976								
52 (Jun-07)	10389	10537	9669	11568	9499	6878							
53 (Sep-07)	10433	10623	9796	11854	9946	8866	7050						
54 (Dec-07)	10451	10662	9927	12005	10115	9236	8950	8110					
55 (Mar-08)	10466	10671	9943	12051	10229	9397	9530	10996	6787				
56 (Jun-08)	10475	10679	9958	12084	10297	9522	9735	11647	9073	6820			
57 (Sep-08)	10484	10707	9978	12118	10342	9560	9849	11953	9538	8993	7501		
58 (Dec-08)	10496	10720	9992	12138	10370	9575	9887	12060	9693	9352	8731	8213	
59 (Mar-09)	10495	10720	9991	12138	10370	9575	9886	12057	9689	9345	9118	11011	6621

Western Australia: Reported counts of completions

	Quarter completion occurred												
Collection reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
47 (Mar-06)	2469												
48 (Jun-06)	2892	2055											
49 (Sep-06)	2933	2448	2095										
50 (Dec-06)	2970	2542	2514	2022									
51 (Mar-07)	2982	2555	2592	2756	2121								
52 (Jun-07)	2983	2564	2625	2900	2683	1867							
53 (Sep-07)	2983	2569	2633	2934	2728	2175	1860						
54 (Dec-07)	2983	2570	2638	2947	2756	2272	2500	2431					
55 (Mar-08)	2983	2571	2641	2958	2768	2301	2601	3022	2589				
56 (Jun-08)	2985	2571	2641	2959	2772	2309	2631	3156	2948	2091			
57 (Sep-08)	2986	2571	2644	2960	2771	2311	2644	3193	3037	2488	2467		
58 (Dec-08)	2986	2572	2644	2961	2776	2344	2647	3206	3093	2594	2844	2797	
59 (Mar-09)	2986	2574	2644	2961	2777	2346	2651	3209	3101	2642	2940	3494	2604

Tasmania: Reported counts of commencements

	Quarter commencement occurred												
Collection reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
47 (Mar-06)	1630												
48 (Jun-06)	2285	959											
49 (Sep-06)	2370	1617	1153										
50 (Dec-06)	2392	1663	1680	1247									
51 (Mar-07)	2395	1674	1756	1674	1403								
52 (Jun-07)	2395	1678	1773	1731	2326	1255							
53 (Sep-07)	2395	1681	1779	1748	2411	1808	1332						
54 (Dec-07)	2395	1681	1779	1749	2417	1859	1822	1175					
55 (Mar-08)	2395	1681	1779	1749	2420	1865	1877	1723	1700				
56 (Jun-08)	2390	1681	1779	1742	2420	1865	1895	1822	2599	1467			
57 (Sep-08)	2395	1681	1779	1746	2422	1866	1902	1840	2694	2070	1217		
58 (Dec-08)	2395	1681	1779	1746	2422	1866	1901	1845	2708	2104	1811	1253	
59 (Mar-09)	2397	1682	1781	1751	2422	1867	1902	1849	2715	2106	1837	1617	1523

Australian Capital Territory: Reported counts of completions

	Quarter completion occurred												
Collection reported	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
47 (Mar-06)	1331												
48 (Jun-06)	1752	762											
49 (Sep-06)	1792	1174	765										
50 (Dec-06)	1807	1219	1137	658									
51 (Mar-07)	1812	1222	1165	987	770								
52 (Jun-07)	1813	1223	1175	1033	1298	466							
53 (Sep-07)	1789	1216	1177	1044	1371	687	724						
54 (Dec-07)	1805	1220	1178	1050	1400	734	1011	424					
55 (Mar-08)	1817	1228	1186	1108	1628	1003	1203	808	971				
56 (Jun-08)	1817	1228	1186	1107	1631	1006	1212	858	1679	772			
57 (Sep-08)	1817	1228	1186	1106	1632	1006	1213	866	1721	1023	675		
58 (Dec-08)	1817	1228	1186	1106	1632	1005	1212	867	1728	1049	963	512	
59 (Mar-09)	1817	1228	1186	1106	1632	1005	1212	867	1729	1050	999	787	1158

Queensland: Reported counts of cancellations and withdrawals

	Quarter cancellation or withdrawal occurred																
Collection reported	Mar-05	Jun-05	Sep-05	Dec-05	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
43 (Mar-05)	3220																
44 (Jun-05)	4730	3290															
45 (Sep-05)	5320	4967	3822														
46 (Dec-05)	5450	5264	4895	3742													
47 (Mar-06)	5529	5396	5203	5045	3926												
48 (Jun-06)	5566	5464	5348	5444	5092	3461											
49 (Sep-06)	5600	5517	5445	5644	5596	5012	3692										
50 (Dec-06)	5619	5559	5479	5753	5768	5452	5106	3744									
51 (Mar-07)	5649	5596	5526	5840	5906	5654	5648	5626	4297								
52 (Jun-07)	5671	5619	5555	5890	5962	5772	5862	6148	5918	3912							
53 (Sep-07)	5683	5627	5571	5923	6021	5847	6008	6436	6576	5892	4302						
54 (Dec-07)	5696	5640	5587	5946	6052	5895	6079	6561	6803	6473	6346	4512					
55 (Mar-08)	5712	5653	5599	5958	6072	5958	6122	6633	6916	6650	6898	6618	4214				
56 (Jun-08)	5714	5657	5605	5973	6085	5982	6151	6698	6980	6760	7149	7239	6395	4376			
57 (Sep-08)	5725	5666	5618	5983	6097	5992	6176	6731	7041	6846	7314	7539	7028	6624	4644		
58 (Dec-08)	5731	5666	5620	5991	6108	6002	6191	6753	7084	6900	7378	7655	7232	7105	6424	4093	
59 (Mar-09)	5731	5668	5622	5997	6113	6013	6206	6768	7103	6934	7428	7735	7372	7334	6938	6145	3953

Northern Territory: Reported counts of cancellations and withdrawals

	Quarter cancellation or withdrawal occurred																
Collection reported	Mar-05	Jun-05	Sep-05	Dec-05	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
43 (Mar-05)	208																
44 (Jun-05)	324	208															
45 (Sep-05)	356	329	231														
46 (Dec-05)	360	357	326	252													
47 (Mar-06)	361	362	350	376	221												
48 (Jun-06)	361	362	353	412	327	181											
49 (Sep-06)	361	362	353	418	355	321	168										
50 (Dec-06)	362	367	352	420	367	373	329	191									
51 (Mar-07)	362	369	352	421	372	375	367	313	241								
52 (Jun-07)	362	369	352	421	372	377	373	343	340	233							
53 (Sep-07)	362	369	352	421	372	377	373	347	356	325	226						
54 (Dec-07)	362	369	352	421	372	378	373	351	361	359	365	211					
55 (Mar-08)	362	369	352	421	374	380	373	353	361	361	381	322	228				
56 (Jun-08)	362	369	352	421	374	382	374	353	362	363	390	362	325	204			
57 (Sep-08)	362	369	352	421	374	382	374	353	363	366	394	377	370	346	158		
58 (Dec-08)	362	370	351	421	375	382	374	355	363	367	397	383	383	387	323	176	
59 (Mar-09)	363	369	351	421	373	384	373	357	365	367	397	383	391	401	360	322	166

South Australia: Calculated expired contracts

	Quarter cancellation or withdrawal occurred																
Collection reported	Mar-05	Jun-05	Sep-05	Dec-05	Mar-06	Jun-06	Sep-06	Dec-06	Mar-07	Jun-07	Sep-07	Dec-07	Mar-08	Jun-08	Sep-08	Dec-08	Mar-09
43 (Mar-05)	2416																
44 (Jun-05)	1513	1776															
45 (Sep-05)	1201	1166	1751														
46 (Dec-05)	955	921	1138	1590													
47 (Mar-06)	835	783	956	1156	1968												
48 (Jun-06)	730	663	765	927	1204	1727											
49 (Sep-06)	645	599	644	790	1003	1191	1745										
50 (Dec-06)	611	562	589	680	839	1009	1203	1566									
51 (Mar-07)	598	542	573	643	743	845	983	1132	2804								
52 (Jun-07)	593	539	550	619	693	746	892	915	1681	2177							
53 (Sep-07)	591	537	545	607	675	716	850	838	1488	1644	2133						
54 (Dec-07)	586	533	538	595	660	690	820	733	1166	1332	1423	1603					
55 (Mar-08)	582	529	535	586	651	674	802	682	1039	1172	1238	1025	1891				
56 (Jun-08)	568	511	524	570	608	646	757	635	907	998	964	748	1156	1437			
57 (Sep-08)	555	501	509	559	587	632	743	608	822	907	838	652	939	1020	1611		
58 (Dec-08)	553	500	505	551	582	625	736	590	782	857	789	612	820	824	1240	1635	
59 (Mar-09)	551	491	502	548	579	623	726	572	752	827	733	575	730	722	1033	1231	2287