TECHNICAL PAPER

NCVER

Methodological approaches for projecting completion rates for apprentices and trainees

Michelle Hall
Brian Harvey
National Centre for Vocational Education Research

INFORMING + INFLUENCING
THE AUSTRALIAN VET SECTOR

Publisher's note

The views and opinions expressed in this document are those of NCVER and do not necessarily reflect the views of the Australian Government, or state and territory governments. Any interpretation of data is the responsibility of the author/project team.

To find other material of interest, search VOCEDplus (the UNESCO/NCVER international database <http://www.voced.edu.au>) using the following keywords: apprentice; apprenticeship; apprenticeship contract; completion; data analysis; data collection; evaluation; outcomes of education and training; participation; providers of education and training; registered training organisation; qualifications; statistical method; statistics; students; trainee; vocational education and training.

This document should be attributed as Hall, M & Harvey, B 2021, *Methodological approaches for projecting completion rates for apprentices and trainees*, NCVER, Adelaide.

This work has been produced by NCVER on behalf of the Australian Government and state and territory governments, with funding provided through the Australian Government Department of Education, Skills and Employment.

COVER IMAGE: GETTY IMAGES

ISBN 978-1-925717-78-5
TD/TNC 144.17

Published by NCVER, ABN 87 007 967 311

Level 5, 60 Light Square, 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> <https://www.lsay.edu.au>
Follow us: <https://twitter.com/ncver> <https://www.linkedin.com/company/ncver>

About the research

Methodological approaches for projecting completion rates for apprentices and trainees

Michelle Hall and Brian Harvey, National Centre for Vocational Education Research

The purpose of this technical paper is to summarise exploratory work investigating the effectiveness of the life tables methodology, which is currently used to calculate completion rate projections for apprentices and trainees, against two alternative methodological approaches – Markov chains and machine learning.

The National Centre for Vocational Education Research (NCVER) publishes *projected* contract completion rates for recent commencing cohorts of apprentices and trainees. Reporting projected rates is necessary because *actual* completion rates cannot be known until enough time has passed for contracts to be completed and for the outcomes to be reported to NCVER.

This report includes:

- an overview of the three methodologies – life tables, Markov chains and machine learning – that were applied to calculate projected completion rates for apprentice and trainee contracts of training
- a comparison of the accuracy of the projections generated by the three methodologies
- an evaluation of the relative strengths and limitations of the three methodologies.

Key messages

- Between 2013 and 2015 – the three years of data for which projections were examined – the completion rate projections calculated using Markov chains and machine learning were generally more accurate than the rates achieved using life tables. When evaluated against *observed actual* completion rates:
  - For *non-trade* cohorts, the error rates were the lowest and most consistent for machine learning projections.
  - For *trade* cohorts, the error rates were the lowest and most consistent for Markov chains projections.
- Each method has strengths and limitations:
  - The *life tables* approach calculates completion rate projections for December quarter commencing cohorts, which are taken as proxy rates for annual commencing cohorts. This may account for the lower accuracy of life tables projections when evaluated against observed actual completion rates, which are calculated for annual cohorts.
  - The *Markov chains* approach produced the most accurate projections overall. However, this methodology generates projections with a 12-month delay, in contrast to the life tables and machine learning approaches.
  - The *machine learning* methodology produced highly accurate projections for non-trade cohorts, and the projections for trade cohorts were comparable with those generated by Markov chains. Unlike the other methods, strategies such as adding additional attributes or improving the data quality of existing attributes can be used to improve the performance of the model.

- All three methodological approaches draw on a five-year window of recent historical data to calculate projections. Due to the significant disruption to apprentices and trainees resulting from the COVID-19 pandemic, it is not clear whether the assumptions underlying the methodologies remain valid for years where the reference period includes disruption from COVID-19. For this reason, it is not considered appropriate to implement changes to the projection methodology used in NCVER reporting at this time.

Simon Walker
Managing Director, NCVER

Acknowledgements

The authors would like to thank Javen Qinfeng Shi from the Australian Institute for Machine Learning for feedback on the discussion of machine learning in an earlier version of this report.

Contents

Tables and figures

Tables

Figures

Introduction

Completion rates for annual cohorts of apprentices and trainees are published annually by the National Centre for Vocational Education Research (NCVER). Observed actual completion rates are calculated directly from the collected data for annual cohorts who commenced four years or more before the year in which the publication is released (five years or more for trade contracts). For example, the *Completion and attrition rates for apprentices and trainees 2020* publication (released in July 2021) reports the *observed actual* completion rates calculated for annual cohorts who commenced in 2014-2017 for non-trade contracts and 2014-2016 for trade contracts (NCVER 2021).

For more recent cohorts, NCVER calculates and reports *projected* completion rates by modelling recent historical data trends. Projected rates are necessary for recent years because it takes some time for training contracts to be completed (or cancelled/withdrawn) and for the outcomes to be reported to the National Apprentice and Trainee Collection. NCVER currently calculates completion rate projections for recent commencing cohorts using life tables methodology.

This technical paper outlines recent exploratory work that was conducted to evaluate the effectiveness of two alternative methodologies for calculating project completion rates for recent commencing cohorts: Markov chains (developed by NCVER) and machine learning (developed by the Australian Institute of Machine Learning).

The purpose of this technical paper is to:

- provide an introduction to completion rates for apprentices and trainees
- provide an overview of the three methodologies whose performance in projecting apprentice and trainee completion rates was evaluated
- describe exploratory work comparing the performance of the three methodologies in projecting contract completion rates for apprentices and trainees
- compare the relative strengths and limitations of the three methodologies evaluated.

Conceptually, completion rates can be calculated for contracts or individuals. The focus of this technical paper is on the former. The complexities associated with the calculation of completion rates for individuals, as well as the concept of attrition, are discussed briefly at the end of the report.

The National Apprentice and Trainee Collection

The National Apprentice and Trainee Collection contains data records that capture the history of apprentice and trainee contracts of training. Data are submitted to NCVER every quarter and contain the most recent information on the current and previous quarters (currently back to July 2011). These data include information on:

- contract events (for example, when they commence, complete or cancel/withdraw)
- the client (for example, demographics)
- the training (for example, the qualification being completed, level of education)
- the employer (for example, employer size, based on number of employees)
- the training provider (for example, type of provider).

The data allow the grouping of contracts into cohorts, which can be based on year of commencement (the most common) and quarter of commencement, as well as other aggregations. A particular aggregation of interest explored in this technical paper is that of trade occupations versus non-trade occupations.[1]

For a given cohort, the contracts can be further grouped according to their status:

- terminated contracts
  - completed contracts
  - cancelled/withdrawn contracts
- unreported contracts
  - live contracts: those that have yet to pass their expected completion date
  - expired contracts: those that have passed their expected completion date but have not reported a termination event.

Terms and definitions can be found on NCVER's *Apprentices and trainees collection* page[2].

Reporting and time delays

Information about contract events takes some time to appear in the National Apprentice and Trainee Collection. The two key reasons for this are:

- reporting delays: delays in reporting contract events to NCVER after they occur
- time delays: following commencement, the time taken to either complete or cancel/withdraw from a contract.

For example, suppose that a group of apprentices/trainees commence contracts of training with an expected duration of four years. All contract events, including commencements, take some time to be reported to NCVER and then to appear in the National Apprentice and Trainee Collection. This reporting delay can be up to one year for completions and longer for cancellations/withdrawals. In addition to reporting delays, completions and cancellations/withdrawals will occur some period of time after they commenced. For cancellations/withdrawals, the terminating event could occur at any time. For completions, it is possible that some apprentices/trainees could complete their contract prior to the expected duration of four years (for example, if some are granted credit); however, most of the completions will occur around four years later. A detailed discussion of reporting and time delays is described in the technical paper: *Estimation of apprentice and trainee statistics* (Harvey 2010).

Unreported rates

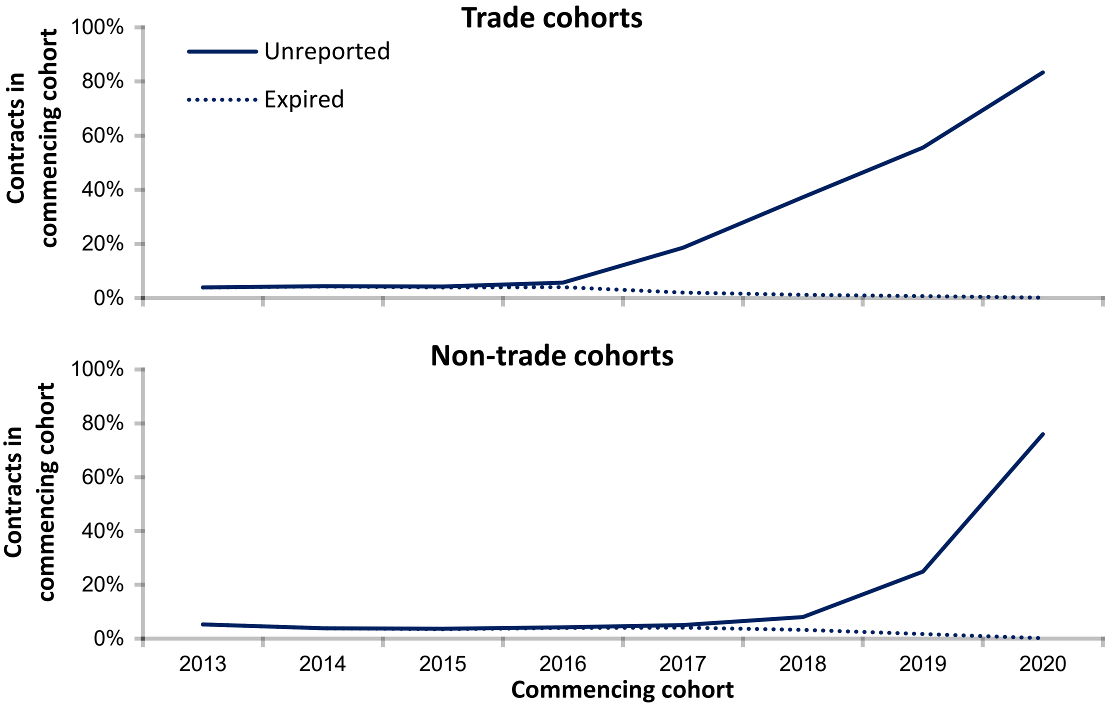
The reporting and time delays described above can be measured by looking at the unreported rate; that is, the proportion of contracts in a commencing cohort where a termination event has not yet been reported.

[1] Trade occupations are those classified as technicians and trade workers, whereas non-trade occupations are those not classified as technicians and trades workers. Occupations are classified using the Australian and New Zealand Standard Classification of Occupations (ANZSCO), first edition, revision 2 (ABS 2013).

[2] <https://www.ncver.edu.au/research-and-statistics/collections/apprentices-and-trainees-collection>.

Figure 1 shows the unreported rates for annual cohorts of contracts commencing between 2013 and 2020, separately for trade and non-trade cohorts, based on the latest available data. The unreported rate is comprised of expired contracts (depicted by the dotted line) and live contracts (not shown). As can be seen from figure 1, expired rates comprise only a small fraction of the unreported rates.

Figure 1 Unreported and expired rates for annual trade and non-trade cohorts, commencing 2013−20 (%)



Source: National Apprentice and Trainee Collection, no.107 (March 2021 estimates).

The general trend for the most recent commencing cohorts, both trade and non-trade, is that the unreported rate is high, while the expired rate is low. With many contracts still active, there has been little opportunity for terminating events to have been reported or for the expected completion date to have elapsed. For example, for the contracts that commenced in 2020 (trade and non-trade combined), only 20.3% had reported a terminating event by March 2021 and less than 0.2% of contracts had expired.

The unreported rate drops for cohorts who commenced in earlier years because more time has passed, allowing contracts to have reached a terminating event (completion or cancellation/withdrawal) and for those terminating events to have been reported. The expired rate increases as the unreported rate drops off as some contracts reach their expected completion date but fail to report a terminating event.

The rate of unreported contracts stabilises after four years for trade contracts and after three years for non-trade contracts. For example, based on the March 2021 data collection, which includes data to the end of December 2020, the unreported rate has stabilised for the cohort of trade contracts who commenced in 2016 and for the cohort of non-trade cohorts who commenced in 2017 (figure 1). The primary reason for this difference is that trade contracts tend to have longer durations than non-trade contracts, which explains why NCVER's completion rates publication has historically calculated completion rates one year more for non-trade contracts than for trade contracts.

Completion rates

A completion rate is essentially the proportion of all contracts in a cohort reported as terminated with completion.

Completion rates are calculated directly from the data for cohorts whose contracts have passed their expected date of completion and sufficient time has passed for the outcomes of the contracts to have been reported to NCVER. In our example, based on the data available in the March 2021 collection, *observed actual* rates are calculated directly from the data for cohorts commencing in 2016 and earlier.

However, completion rates calculated directly from the data are unreliable for recent commencing cohorts, where the unreported rate is still high. This is because it is still unknown how many live contracts will eventually report a completion once their expected term has come to an end. For these cohorts, a projection methodology can be applied to provide an indication of the likely completion rate. Projection methodologies are explored in the following section.

NCVER's annual completion and attrition rates publication is based on March quarter data of the following year. For example, the publication recording 2020 completion and attrition rates (released in July 2021) was produced using data from the March 2021 collection. Using the March quarter data allows extra time for information to be reported and mitigates, to some extent, the effect of reporting delays in the collection. The base population for the calculation of rates for any given year is the cohort of contracts who commenced training in that year. Rates can also be calculated for subcategories of the data (for example, trade occupations, non-trade occupations).

The following is a summary of the information given above:

- Reliable completion rates can only be calculated directly from the data when the unreported rate has fallen and reached a stable point.
- The unreported rate continues to fall until enough time has passed for contracts to have reached their expected completion date.
- The unreported rate is affected by delays in reporting termination events to the National Apprentice and Trainee Collection.
- It takes longer for unreported rates to fall for trade cohorts than for non-trade cohorts.

Projection methodologies

Having an indication of the likely completion rate for a given cohort well in advance of the expected completion date is an issue of considerable interest. To address this, completion rates can be approximated by modelling recent historical data to generate a projection. According to the Australian Bureau of Statistics (ABS) website:

> A projection indicates what the future changes in a population would be if the assumptions about the future trends actually occur. These assumptions are often based on patterns of change which have previously occurred. A projection is not making a prediction or forecast about what is going to happen, it is indicating what would happen if the assumptions which underpin the projection actually occur.[3]

This section provides a general introduction to three methodologies that can be used to generate projections – life tables, Markov chains and supervised machine learning.

Life tables methodology

NCVER currently calculates projected completion rates using a life tables methodology. This methodology has its foundations in demographic and actuarial studies and uses data on population, deaths and births for a recent reference period to calculate mortality, survivorship and life expectancy (ABS 2017—19). The life table can then be used to make inferences about hypothetical/future cohorts, on the assumption that age-specific mortality rates during the reference period will remain the same.

In the context of contract completion rates, the life tables approach has been adapted by:

- considering contract status (commencements, terminations) to replace births and deaths, thereby calculating a projected completion rate rather than a projected mortality rate.

For an in-depth explanation of how this methodology has been adapted for the National Apprentice and Trainee Collection, refer to Karmel and Mlotkowski (2010).

A key limitation of the life tables methodology is that it calculates completion rate projections for December quarter commencing cohorts and uses them as proxy rates for annual commencing cohorts.

Markov chains methodology

Absorbing Markov chains provide a means of modelling objects that progress through one or more transient states before reaching one or more final (absorbing) states. This methodological approach calculates the probabilities of objects transitioning between the different states and can be used to calculate the probability of an object eventually reaching one of the ‘absorbing’ states. Markov chains have applications in a wide variety of areas, including science, economics and mathematics. For a theoretical review see Isaacson and Madsen (1976).

NCVER currently uses Markov chains methodology to project completion rates for VET qualifications, whereby the lifetime of training can be expressed in terms of the probability of transitioning between one state (for example, active) to another state (for example, completed). Once the qualification has

[3] <http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+estimate+and+projection>.

been completed, the state will not change again, so completion is referred to as an ‘absorbing’ state. The probability that a qualification is eventually in the completed state is the completion rate. The projected completion rate for a given year is based on recent longitudinal data, which include the year for which the completion rate is being projected, the previous year and the following year. For more detailed information on the Markov chains methodology in this context, refer to Mark and Karmel (2010) and McDonald (2018).

The methodology used for VET qualification completion rates was altered slightly for application to the National Apprentice and Trainee Collection. This was mainly due to the duration of contracts of training tending to be longer than the duration of VET qualifications, particularly in trade occupations. For contract completion rates, the three-year window (used in the calculation of VET qualification completion rates) was modified to a five-year window to ensure sufficient data to estimate transitional probabilities. The window includes the year for which the completion rate is being projected, the preceding three years, and the following year. Note that the five-year window allows for contracts with long durations but also works for contracts with short durations.

A key limitation of the Markov chains methodology is the 12-month delay before projected rates can be calculated. This is because calculating the transitional probabilities that form the basis for the completion rate projection for a given year relies on data from a window that includes the following year. That is, rates for the most recent year cannot be estimated since there is no following year to provide data.

Similar to the life tables methodology, the Markov chains approach calculates projections based on transitional probabilities between different states, such as transitions from actively training to completion of a contract. For both methods, the only information used in the modelling is data on changes of contract status, which include the timing of the event, the previous contract status and the new contract status.

Machine learning methodology

Machine learning refers to a variety of automated methods whose aim is to uncover patterns in a set of data observations (Murphy 2012). Classification is one form of machine learning where the patterns are used to differentiate between two or more categories of interest. Classification methods are a form of *supervised* machine learning because the categories are made explicit in the training examples that are used to ‘train’ the classification model.

The first step in machine learning classification is to ‘train’ a model by showing it example records from each category of interest. The goal of training is to identify patterns among the features (also known as attributes) of the example records that distinguish between the categories. Following training, the model is evaluated on a set of new records, with the category labels removed. The model’s prediction of the category that each new record belongs to is then compared with the true category labels. If the model’s performance is satisfactory, it can be used to make predictions for new data records where the categories are not yet known.

In the context of apprentice and trainee completion rates, machine learning methodology has been developed by the Australian Institute for Machine Learning (AIML) to predict the final status (completed, cancelled/withdrawn) of apprentice and trainee contracts of training, based on the attributes of the contracts when they commenced. The example records used to train the model include information about the attributes of the client, the training being undertaken, and the employer, among others. After training and evaluating the model, it is used to predict whether new contracts will result in a completion. Cohort completion rates are calculated by averaging individual predictions within a cohort. Further detail of this approach is included in appendix A, including the full list of attributes included in the modelling.

With many attributes and the patterns driving training outcomes likely to be complex, deep learning approaches were considered appropriate for this task. Deep learning methods are powerful for learning rich representations (LeCun, Bengio & Hinton 2015). This methodology draws on more extensive information about each contract than do the life tables and Markov chains methodologies because it relies on patterns among contract attributes to differentiate between those that will later complete and those that will not complete.

Comparing the accuracy of projected completion rates

This section presents a comparison of the performance of the three projection methodologies described in the previous chapter.

To generate a projected completion rate for a given cohort of interest, each methodology follows a two-stage approach:

- building a model based on recent historical data
- applying the model to a new commencing cohort to make a projection of the completion rate for that cohort.

The data requirements for stage 1 (building the model) are different for each methodology. These data requirements are detailed in appendix B.

For this comparison, completion rates have been projected separately for the trade and non-trade cohorts who commenced in 2013, 2014 and 2015. The performance of each methodology has been evaluated by comparing the projected completion rate for each cohort to the observed actual completion rate for that cohort.[4] The years 2013 to 2015 were chosen to match the data that were readily available for the most direct comparison of the projections attained using the life tables, Markov chains and machine learning methodologies.

To summarise the outcomes of this evaluation exercise:

- Contract completion rates have been projected using three methodologies: life tables, Markov chains, and machine learning.
- Completion rates have been projected separately for trade and non-trade cohorts.
- Completion rates have been projected for cohorts commencing in 2013, 2014 and 2015.
- The accuracy of each projected rate has been evaluated against the observed actual completion rate for that cohort, as calculated directly from the data.

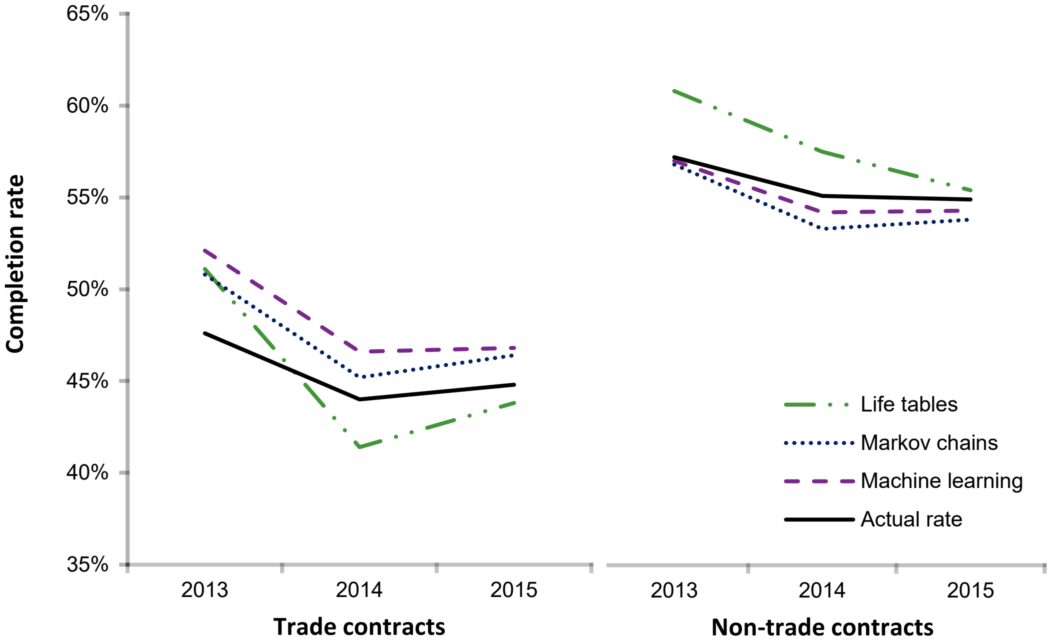
Projected completion rates: accuracy

Figure 2 shows a comparison of the projected rates for each of the methodologies for the trade and non-trade cohorts who commenced between 2013 and 2015. The Markov chains and machine learning methodologies calculate projected completion rates for annual commencing cohorts. Under the life tables methodology, projected completion rates are calculated for the December commencing cohorts, which are taken as proxy rates for the annual commencing cohort.

[4] The actual completion rates for these cohorts can be calculated directly from the data because enough time has passed for the unreported rates to have fallen and reached a stable level. However, it is important to note that the actual completion rate is only an approximation of the true completion rate for the cohort because some contracts never report a final status. The true completion rate may therefore be higher than the actual completion rate calculated from the data in instances where some of the unreported contracts are completions that have not been reported as such.

The observed actual completion rate for each annual cohort is depicted by the black solid line. The actual rates have been calculated directly from the data from a point in time five years after the commencement year for the cohort, as published by NCVER.[5]

Figure 2 Observed actual and projected contract completion rates by trade and non-trade occupations, for annual cohorts commencing 2013−15 (%)



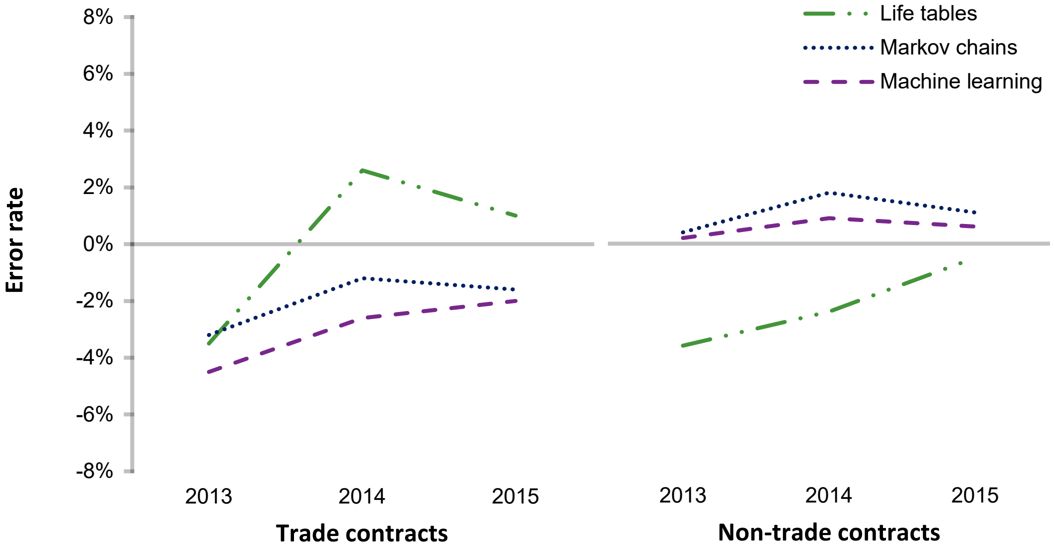
Source: Actual rates have been calculated from National Apprentice and Trainee Collection nos. 99, 103 and 107 (March estimates 2019−21). Data sources for the projected rates vary according to the methodology (see appendix B).

In addition to visualising the projected rates alongside the actual rates, it is useful to visualise the error associated with each projection (figure 3). For the three years examined, for trade and non-trade contracts, all three methodologies have absolute error rates of fewer than five percentage points.

In general, the projections made for non-trade contracts are more accurate than the projections made for trade contracts. For the years examined, the Markov chains methodology provides the lowest average absolute error rate for trade cohorts, while the machine learning approach gives the lowest average absolute error rates for non-trade cohorts.

[5] Actual completion rates for 2013 cohorts are drawn from the NCVER publication *Completion and attrition rates for apprentices and trainees 2018* based on March 2019 estimates (NCVER 2019); actual rates for 2014 cohorts are drawn from *Completion and attrition rates for apprentices and trainees 2019* based on March 2020 estimates (NCVER 2020b); actual rates for 2015 cohorts are drawn from *Completion and attrition rates for apprentices and trainees 2020* based on March 2021 estimates (NCVER 2021).

Figure 3 Error rates for projected contract completion rates by trade and non-trade occupations, for annual cohorts commencing 2013−15 (%)



Source: Error rates have been calculated based on observed actual completion rates, which have been calculated from National Apprentice and Trainee Collection nos. 99, 103 and 107 (March estimates 2019−21). Data sources for the projected rates vary according to the methodology (see appendix B).

In the case of the machine learning model, a single model was trained on data from the 2012 commencing cohort. This model generalised well to the 2013, 2014 and 2015 cohorts, with the error rate remaining under 5%. This is an important feature of the results, as the machine learning model needs to have the ability to generate reliable projections for cohorts commencing four years after the reference cohort in order to provide timely projections for new commencing cohorts.

For this analysis, data were not available for assessing model projections for a cohort four years after the reference cohort, but the accuracy of the projections for the cohort three years after the reference cohort is a preliminary indicator that this method is robust for future cohorts.

Evaluation

Each of the methodologies displays strengths and limitations. These are summarised in table 1 and discussed in detail throughout this section.

Table 1 Evaluation and comparison of projection methodologies

Criterion	Life tables	Markov chains	Machine learning
Accuracy	Projected completion rates are within **four** percentage points of observed actual rates for non-trade and trade occupations	Projected completion rates are within **two** percentage points of observed actual rates for non-trade occupations and within **four** percentage points for trade occupations	Projected completion rates are within **one** percentage point of observed actual rates for non-trade occupations and within **five** percentage points for trade occupations
Data requirements	Annual training activity from the current and previous four years	Annual training activity from the current and previous four years	All training activity associated with a cohort that commenced four years prior to the current year
Timeliness	Projections for December quarter commencing cohort made as soon as data are available for that cohort	Projections for annual commencing cohort made one year after data are available for that cohort	Projections for annual commencing cohort made as soon as data are available for that cohort
Assumptions	Projected completion rates for the December quarter commencing cohorts are based on recent five-year historical trends in annual data December quarter projections are taken as a proxy rate for the annual cohort	Projected completion rates for annual commencing cohorts are based on recent five-year historical trends in annual data	Projected completion rates for annual commencing cohorts are based on recent five-year historical trends in annual data
Prediction errors	None of the methods produce an estimate of prediction error or confidence intervals at the time the projections are made. The methods can only be evaluated for accuracy when enough time has passed for the observed actual completion rates to be known. Observed actual completion rates as calculated from the data may underestimate the true completion rates due to the persistence of unreported (including expired) contracts in the National Apprentice and Trainee Collection.		

Accuracy

At the trade/non-trade level of aggregation, annual completion rate projections for all three methods were within five percentage points of the actual rates for the cohorts assessed. Projections for the non-trades were generally closer to the observed actual rates than projections for the trades.

The Markov chains methodology had the lowest average error rate for trade contracts, whereas the machine learning approach had the lowest average error rate for non-trade contracts. The error rates for projections generated using Markov chains and machine learning were more consistent than the error rates for projections generated using life tables.

Data requirements

Each methodology has slightly different data requirements for modelling the recent historical data trends to make a projection (for detail, refer to appendix B). All three methodologies require a five-year window of recent historical training activity. The life tables and Markov chains approaches require records of all training activity that occurred within the five-year reference period, whereas the machine learning approach requires the commencing records and contract outcomes of the cohort that commenced in the first year of the reference period, and the commencing records of the cohort that commenced in the final year of the reference period (the cohort for which the projection is being calculated).

To illustrate, we can consider the projections that can be made based on the March 2021 collection (the latest available data):

- Using Life tables, the latest cohort for which projected completion rates can be calculated is the cohort that commenced in the December quarter of 2020:
  - Data are required on annual training activity from 2016 to 2020.
  - The probabilities of contracts changing states between each quarter from the March quarter 2016 to the December quarter 2020 are calculated.
  - These quarterly transitional probabilities are applied to the trade and non-trade cohorts that commenced in the December quarter 2020 to project rates for these cohorts.
- Using Markov chains, the latest cohort for which projected completion rates can be calculated is the annual cohort that commenced in 2019:
  - Data are required on annual training activity from 2016 to 2020.
  - The probabilities of contracts changing states between each year from 2016 to 2020 are modelled.
  - These annual transitional probabilities are applied to the trade and non-trade cohorts that commenced in 2019 to project rates for these cohorts.
- Using machine learning, the latest cohort for which projected completion rates can be calculated is the annual cohort that commenced in 2020:
  - Data are required on contract attributes and events (commencements, cancellations/withdrawals, and completions) for the cohort who commenced in 2016.
  - The patterns of training activity associated with completions and cancellations/withdrawals for the cohort commencing in 2016 are modelled.
  - These patterns of training activity are applied to the trade and non-trade contracts that commenced in 2020 to make predictions of the outcomes of each contract. The predicted outcomes are averaged across trade and non-trade contracts to derive projected completion rates for these cohorts.

Timeliness of projections

The Markov chains methodology requires a delay of one year to make projections by comparison with the life tables and machine learning approaches. Taking the 2020 commencing cohort as an example, a projection can be made for this cohort using:

- the life tables or machine learning methodologies, based on March 2021 estimates
- the Markov chains methodology, based on March 2022 estimates.

Representativeness of the reference data

Under the life tables methodology, projections are made for December quarter commencing cohorts, which might not adequately represent annual completion rates. This may lead to bias in the projections and misalignment with the actual annual completion rates for that cohort. This methodology assumes that patterns of transitions between contract states over the past five years will be representative of transitions between contract states for the projection cohort.

The Markov chains methodology calculates projected rates for annual commencing cohorts. This method also relies on the assumption that the patterns of training activity over a recent five-year period are representative of the patterns of training activity that will emerge for the projection cohort. However, the five-year reference period is offset by twelve months as compared with the life tables methodology, resulting in a 12-month delay for calculating projections.

The machine learning model calculates completion rate projections for annual commencing cohorts. This methodology has the assumption that the patterns in contract behaviour of a cohort that commenced four years earlier are representative of contract behaviour for the annual cohort for which the projection is being made. This methodology uses a single annual commencing cohort as the reference, rather than all training activity within the reference window (as is the case for life tables and Markov chains).

Adding new slices of data

In addition to comparing trade and non-trade occupations, projections can be calculated according to other characteristics of interest. This is true for the three methodologies examined. However, the reliability of those projections has not been assessed in this report.

In addition, the machine learning methodology works by making predictions for individual records. In terms of projecting completion rates, the individual predictions can then be aggregated to the cohort level. Furthermore, projecting rates for subsets of the cohort is a straightforward process. This contrasts with the life tables and Markov chains approaches, which calculate the projection at the aggregate level only and therefore need to be recalculated separately for any subset of interest.

Refining the algorithms

Improving the accuracy of projections

The life tables and Markov chains methodologies are fairly rigid algorithms, with limited opportunities for improvement. This contrasts with the machine learning methodology, which allows for strategies that can be adopted to improve model performance. For example, additional attributes that influence contract outcomes can be included in the model. The machine learning work conducted so far was limited to the attributes available in NCVER's apprentice and trainee collection. Additional data sources could be identified through future work.

Individual completion rates

Contract completion rates do not account for students who change employers and who thus may need to start a new contract of training. A student may be successful in completing their training, but during their training may have multiple contracts. Individual completion rates provide a different indicator of movement through the apprenticeship system from contract completion rates.

NCVER currently publishes completion rates for individuals based on the outcomes of contracts of training and an adjustment for a recommencement factor (Karmel 2011). These individual completion rates are only calculated for cohorts where sufficient time has passed for the outcomes of the contracts of training to be known. In addition, the adjustment is based on the number of recommencements, which may not be reported consistently across jurisdictions. An alternative approach, based on tracking client activity over time using the Unique Student Identifier (USI), would not have to deal with this issue. In theory, it should also capture training activity in instances where students change jurisdictions during their training. However, data-quality issues with the USI first need to be explored.

Extensions to the methodology

In the longer term, opportunities are available for using the machine learning modelling to not only make projections of completion rates for new cohorts, but also to understand the factors driving completion and non-completion. Identifying risk factors for non-completion is one direction that could be pursued in future work.

The supervised machine learning approach used in this work is powerful in detecting subtle and complex patterns in the underlying data. However, this sensitivity does come at a cost of interpretability of the model, where the factors affecting completion are not readily explicable. That said, additional methods have been developed to enable the drivers of outcomes to be recognised, which can also be explored in future work.

Limitations of the analysis

For this analysis, projected completion rates were compared for 2013, 2014 and 2015. Ideally, additional years would provide a more robust indication of the accuracy and reliability of the projections calculated by the various methodologies.

It is important to note that completion rates were analysed to the broad level of aggregation of trade and non-trade contracts, but the methodologies were not investigated to finer levels of aggregation (such as ANZSCO occupations).

The accuracy of the completion rate projections is evaluated against *observed actual* completion rates, as calculated from the data after sufficient time has passed to know the outcomes of contracts. However, some contracts never report a final status. The true completion rate may be higher than the observed actual rate, depending on the proportion of expired contracts (that is, those that have passed their expected completion date but have not reported a final status, typically 3–5%) that represent completions not reported to NCVER.

For the reasons given above, this paper is not intended to be a definitive evaluation of the three methodologies, but instead represents initial exploratory work intended to establish the relative strengths and weaknesses of three techniques for projecting completion rates for apprentices and trainees.

Other considerations

The COVID-19 pandemic

Each of the projection methodologies explored in this technical paper draws upon recent historical data trends to make inferences about current or future training patterns. In the case of the life tables and Markov chains methodologies, this relies on the assumption that the transitional probabilities of recent cohorts are likely to be similar to the transitional probabilities of the current cohort (for which a projection is being made). In the case of machine learning, the assumption is that the factors affecting completion and non-completion, and the extent to which they affect them, are similar for the current cohort and the cohort on which the model was 'trained'.

In the context of the COVID-19 pandemic, it is not clear whether these assumptions are reasonable, because training behaviour is likely to have been substantially disrupted. This is evident in the number of contract suspensions and the stark changes in commencement and cancellation/withdrawal patterns (see, for example, NCVER 2020a). New factors with unknown and possibly transient impacts on apprentice and trainee activity, which may not be captured in NCVER data, are likely to emerge (for example, eligibility for government support).

It is important to note that the machine learning approach described in this technical paper is experimental, and it is not yet possible to evaluate whether it will generalise to training activity in the future, once the disruptions of the pandemic no longer apply.

Expired contracts and attrition rates

This technical paper has focused on completion rates. A complementary statistic is the attrition rate. However, a fundamental question relates to the definition of attrition, particularly with regards to expired contracts.

Expired contracts are those for which the expected completion date has passed but no completion or cancellation/withdrawal has been reported; hence, the status of these contracts is not known. Past investigations have found that that some expired contracts have outcomes that could reasonably be treated as attrition, while other expired contracts are unreported completions. The current mix of completions and attrition among expired contracts is not known.

The life tables methodology does not include expired contracts as a form of attrition. In other words, projected completion and attrition rates do not sum to 100%, and rates of expired contracts are considered separately.

By contrast, the Markov chains methodology does treat expired contracts as attrition. This is due to a practical constraint; namely, if not removed from the 'unreported' state, the projections for completion rates are too high. This is because expired contracts that never report a status do not belong in a transient state, that is, the unreported state.

The machine learning model was developed using data records where the final contract status was known, either as a completion or a cancellation/withdrawal. Expired contracts were excluded from the dataset when training the model, and no investigation was conducted into whether expired contracts could be predicted according to their attributes. In practical terms, the machine learning methodology calculates cohort completion and cancellation/withdrawal rates by averaging across predictions for every contract in the commencing cohort. Although the machine learning model makes a prediction of either completion or cancellation/withdrawal for each contract, some contracts will never report either of these statuses, and will remain expired.

References

ABS (Australian Bureau of Statistics) 2006, *Australian and New Zealand Standard Industrial Classification (ANZSIC)*, revision 2, ABS cat.no.1292.0, ABS, Canberra.

——2017—19, *Life tables methodology*, ABS, Canberra, viewed June 2021, <https://www.abs.gov.au/methodologies/life-tables-methodology/2017-2019>.

——2013, ANZSCO — *Australian and New Zealand Standard Classification of Occupations,* version 1.2, cat.no.1220.0, ABS, Canberra, viewed 9 June 2021, <https://www.abs.gov.au/AUSSTATS/abs@.nsf/allprimarymainfeatures/4AF138F6DB4FFD4BCA2571E200096BAD?opendocument>.

——*Australian statistical geography standard (ASGS) remoteness structure*, viewed 25 July 2021, <https://www.abs.gov.au/websitedbs/d3310114.nsf/home/remoteness+structure>.

LeCun, Y, Bengio, Y & Hinton, G 2015, 'Deep learning', *Nature,* vol.521, no.7553, pp.436-444.

Harvey, B 2010, *Estimation of apprentice and trainee statistics*, NCVER, Adelaide, viewed 12 April 2021, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/estimation-of-apprentice-and-trainee-statistics>.

Isaacson, DL & Madsen, RW 1976, *Markov chains: Theory and applications*, John Wiley & Sons, New York.

Karmel, T 2011, *Individual-based completion rates for apprentices,* NCVER, Adelaide, viewed 12 April 2021, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/individual-based-completion-rates-for-apprentices>.

Karmel, T & Mlotkowski, P 2010, *Estimating apprentice and trainee completion and attrition rates using a 'life tables' approach,* NCVER, Adelaide, viewed 12 April 2021, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/estimating-apprentice-and-trainee-completion-and-attrition-rates-using-a-life-tables-approach>.

McDonald, B 2018, *Total VET program completion rates*, NCVER, Adelaide, viewed 19 April 2021, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/total-vet-program-completion-rates>.

Mark, K & Karmel, T 2010, *The likelihood of completing a VET qualification: A model-based approach,* NCVER, Adelaide, viewed 12 April 2021, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/the-likelihood-of-completing-a-vet-qualification-a-model-based-approach>.

Murphy, KP 2012, *Machine learning: A probabilistic perspective*, MIT Press, Cambridge, MA.

NCVER (National Centre for Vocational Education Research) 2019, *Completion and attrition rates for apprentices and trainees 2018*, NCVER, Adelaide, viewed 25 July 2021, <http://hdl.voced.edu.au/10707/513891>.

——2020a, *Apprentices and trainees 2020: June quarter: Australia*, NCVER, Adelaide, viewed 9 June 2021 <http://hdl.voced.edu.au/10707/562147>.

——2020b, *Completion and attrition rates for apprentices and trainees 2019*, NCVER, Adelaide, viewed 25 July 2021, <http://hdl.voced.edu.au/10707/545628>.

——2021, *Completion and attrition rates for apprentices and trainees 2020*, NCVER, Adelaide, viewed 25 July 2021, <https://www.ncver.edu.au/research-and-statistics/publications/all-publications/completion-and-attrition-rates-for-apprentices-and-trainees-2020>.

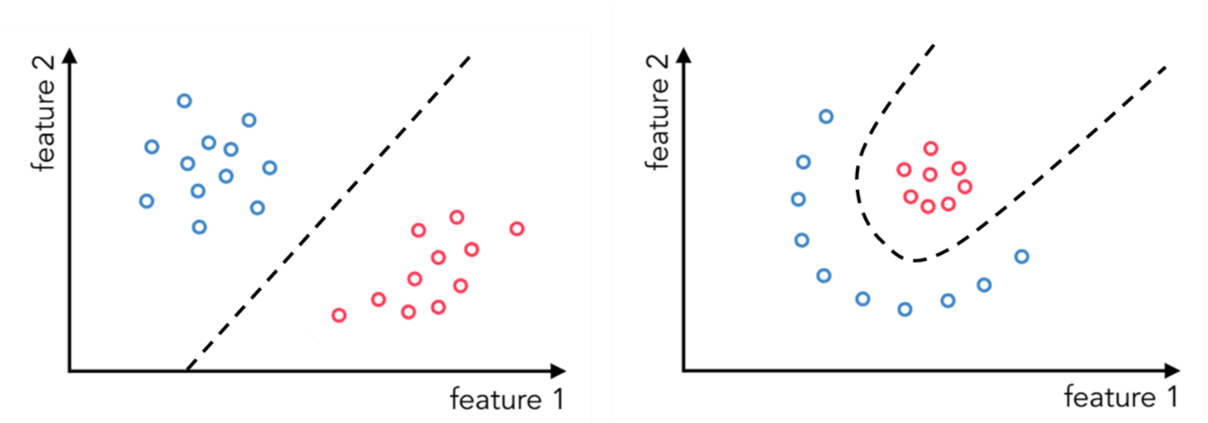
Watt, J, Borhani, R & Katsaggelos, AK 2020, *Machine learning refined: Foundations, algorithms, and applications*, Cambridge University Press, Cambridge, UK.

Appendix A – Methodological detail of machine learning modelling

In the context of projecting completion rates for apprentices and trainees, the goal of the machine learning model is to predict the final status (completed, cancelled/withdrawn) of apprentice and trainee contracts of training, based on the attributes of the contracts when they commenced. In other words, the model is used to categorise new contracts according to whether they are more likely to complete or more likely to cancel/withdraw. Many different machine learning approaches are available for solving this categorisation problem. One of the considerations in choosing which approach to use is the complexity of the data and the complexity of the patterns that might separate the categories of interest.

Figure A1 shows examples of two hypothetical data categories (blue and red data points), which are separated according to two features (feature 1 on the x axis and feature 2 on the y axis). In the example on the left, the separation between the two categories (dashed line) is straightforward, whereas in the example on the right the separation is more complex. The complexity of the separation between categories can increase very quickly when more features are included in the model.

Figure A1 Examples of simple and complex separation of data categories



Source: Adapted from Watt, Borhani & Katsaggelos (2020).

For projecting contract completion rates for apprentices and trainees, many attributes are available to be included in the model and the patterns driving training outcomes likely to be complex. For these reasons, *deep learning* approaches were considered appropriate for this task.

Table A1 shows the training contract attributes that were included in the machine learning models for trade and non-trade cohorts. Several attributes included in the non-trade model were not included in the trade model. These were identified through experiments that revealed that excluding them did not affect the model performance. These variables are identified in table A1.

Records with missing values for key variables such as contract status at completion and trade status (based on ANZSCO) were excluded from the analysis. This means that unreported contracts were not considered in the development of the machine learning model.

Table A1 Attributes of commencing apprentice and trainee contracts included in machine learning modelling

Variable/attribute	Variable included in trade model?
Client attributes	
Age at commencement	Yes
Gender	Yes
Country of birth	No
Main language spoken at home	No
Indigenous status	Yes
Disability status	Yes
State of residence	Yes
Postcode of residence	No
Postcode of residence region – whether client's residence in within state of the data submitter	Yes
Residence remoteness – Accessibility and Remoteness Index of Australia (ARIA+)[6]	Yes
Prior educational achievement flag	Yes
Prior educational achievement count	Yes
Highest school completed	Yes
Highest prior educational achievement	Yes
Whether currently at school	Yes
Whether an existing worker with the employer under the training contract	Yes
Contract attributes	
State administering the contract	Yes
Full-time status of the training contract	Yes
Contract status at commencement – commencement or recommencement	Yes
Calendar quarter of commencement	Yes
Training package identifier	Yes
Australian Qualification Framework level	Yes
Whether contract commenced as an approved school-based apprenticeship	Yes
ANZSCO – Most likely occupational outcome of training	No
Expected contract duration in days	Yes
On-the-job attributes	
Employer type	Yes
Industry of employer[7]	No
Workplace remoteness − Accessibility and Remoteness Index of Australia (ARIA+)	Yes
Workplace postcode	No
Off-the-job attributes	
Registered training organisation (RTO) identifier for off-the-job training delivery	No

[6] Based on the Australian statistical geography standard (ASGS) remoteness structure (ABS 2016).

[7] Based on the Australian and New Zealand Standard Industrial Classification (ANZSIC) (ABS 2006).

Appendix B – Data requirements for calculating projections

The exploratory analysis described in this technical paper compares projected completion rates for cohorts commencing in 2013, 2014 and 2015, separately for trade and non-trade contracts. The three methodologies being compared are life tables, Markov chains and machine learning. Each of these methodologies has different data requirements for calculating a projected completion rate for a given cohort. For the purposes of the exploratory analysis described in this paper, the data requirements for making projections are the same for trade and non-trade cohorts.

Data requirements for life tables

To calculate projected completion rates for the 2013 commencing cohorts (trade and non-trade), the life tables approach uses quarterly transitional probabilities, as calculated from the projection year (2013) and the preceding four years (2009-2012). However, projected completion rates are calculated for the December quarter commencing cohorts, which are taken as proxy rates for the annual commencing cohorts.

The data sources used to calculate the projected completion rates for annual cohorts commencing in 2013, 2014 and 2015 using the life tables methodology are provided in table B1.

Table B1 Data sources: life tables

Projection cohort	Source collection	Reference cohort/s
2013	Collection 79 (March 2014 estimates)	Contracts in-training at any point between 2009 and 2013
2014	Collection 83 (March 2015 estimates)	Contracts in-training at any point between 2010 and 2014
2015	Collection 87 (March 2016 estimates)	Contracts in-training at any point between 2011 and 2015

Data requirements for Markov chains

To calculate the projected completion rate for the 2013 commencing cohort, the Markov chains approach uses transitional probabilities, as calculated from the projection year (2013), the following year (2014), and the preceding three years (2010, 2011 and 2012). This methodology uses a five-year window of recent historical data to calculate transitional probabilities. There is a 12-month delay before a projection can be made for any given projection year by comparison with the life tables and machine learning approaches, because data from the following year are required for the calculations.

The data sources used to calculate the projected completion rates for annual cohorts commencing in 2013, 2014 and 2015 using the Markov chains methodology are provided in table B2.

Table B2 Data sources: Markov chains

Projection cohort	Source collection	Reference cohort/s
2013	Collection 84 (June 2015 estimates)	Contracts in-training at any point between 2010 and 2014
2014	Collection 87 (March 2016 estimates)	Contracts in-training at any point between 2011 and 2015
2015	Collection 91 (March 2017 estimates)	Contracts in-training at any point between 2012 and 2016

Note: Collection 84 (June 2015 estimates) was used as the data source for the projections for 2013 instead of Collection 83 (March 2015 estimates) due to data availability.

Data requirements for machine learning

To calculate the projected completion rate for the 2013 commencing cohort, the machine learning methodology uses data from an annual cohort where the commencing and terminating information is known. Ideally, the machine learning model would be based on a reference cohort that commenced four years earlier and would be retrained for each annual commencing cohort. For this exploration, however, the trade and non-trade models were developed based on training records for the 2012 commencing cohort only. These models were then applied to the commencing records for the trade and non-trade cohorts commencing in 2013, 2014 and 2015. The projection developed for the 2015 commencing cohort, based on the 2012 reference cohort, gives a preliminary indication of the robustness of historical data in predicting contract completions.

The data sources used to calculate projected completion rates for annual cohorts commencing in 2013, 2014 and 2015 using the machine learning methodology are provided in table B3.

Table B3 Data sources: machine learning

Projection cohort	Source collection	Reference cohort
2013	Collection 95 (March 2018 estimates)	Contracts that commenced in 2012
2014	Collection 95 (March 2018 estimates)	Contracts that commenced in 2012
2015	Collection 95 (March 2018 estimates)	Contracts that commenced in 2012

Note: The data source used for the machine learning modelling is a single, later collection than those used for the life tables and Markov chains approaches. However, only data on contracts that commenced in 2012 (including contract outcomes) were used from the March 2018 estimates to develop the machine learning model.

NCVER

National Centre for Vocational Education Research

Level 5, 60 Light Square, 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> <https://www.lsay.edu.au>
Follow us: <https://twitter.com/ncver> <https://www.linkedin.com/company/ncver>