

NATIONAL CENTRE FOR
VOCATIONAL EDUCATION RESEARCH
CONFERENCE PAPER

Maximising the use of data: vocational education and training

Tom Karmel

NATIONAL CENTRE FOR
VOCATIONAL EDUCATION RESEARCH

Presented at NatStats, Brisbane, 2013



Maximising the use of data: vocational education and training

Tom Karmel

*Paper presented to NatStats, Brisbane,
March, 2013*

Publisher's note:

On page 6, under NCVER data collections, last sentence in the first dotpoint should read:

Data are collected on over 2000 providers, covering 1.9 million students

© National Centre for Vocational Education Research, 2013



With the exception of cover design, artwork, photographs, all logos, and any other material where copyright is owned by a third party, all material presented in this document is provided under a Creative Commons Attribution 3.0 Australia
<www.creativecommons.org/licenses/by/3.0/au>.

NCVER is an independent body responsible for collecting, managing and analysing, evaluating and communicating research and statistics about vocational education and training (VET).

NCVER's inhouse research and evaluation program undertakes projects which are strategic to the VET sector. These projects are developed and conducted by NCVER's research staff and are funded by NCVER. This research aims to improve policy and practice in the VET sector.

This document should be attributed as Karmel, T 2013, *Maximising the use of data: vocational education and training*, NCVER, Adelaide.

TD/TNC 111.11

Amended April 2013

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

P +61 8 8230 8400 F +61 8 8212 3436 E ncver@ncver.edu.au W < www.ncver.edu.au>

About the research

Maximising the use of data: vocational education and training

Tom Karmel

This paper was presented to the NatStats conference in March 2013. I briefly describe the range of data collected by NCVET for the national vocational education and training (VET) system and how it is disseminated. I then talk about how analytical techniques are used to maximise what can be extracted from those collections; the current focus on expanding the range of data to contribute to the 'transparency agenda'; and current gaps in VET data collections. I end with some comments on the challenges NCVET faces in maximising the use of data, notably the use of 'privacy considerations' to constrain data use.

Tom Karmel
Managing Director, NCVET

Contents

Introduction	6
NCVER's collections	6
Using analytical tools to expand data use	7
Expanding the range of data on vocational education and training	9
Gaps in VET data	11
Concluding comments	12
References	14

Introduction

The National Centre for Vocational Education Research is the official statistical agency for vocational education and training statistics. It sees its vision as contributing ‘to improvement in Australia’s education and training system by ensuring that policy and practice can be based on sound evidence’ and its mission as being ‘Australia’s leading provider of high-quality, independent information on vocational education and training (VET) to governments, the education sector, industry and the community’.

NCVER has both statistical and research roles, but in this presentation I will concentrate on the former. That said, statistics and research make very productive partners, with the statistics feeding into research and the research feeding back into the way statistics are collected and disseminated. Indeed, in some situations it is difficult to draw a simple divide between them. For example, some of the analytical techniques used in the statistical collections were developed as part of research projects.

In this presentation I will talk briefly about NCVER’s statistical collections and the ways in which we disseminate data. I will then talk about how analytical techniques are used to maximise what can be squeezed out of those collections; the current focus on expanding the range of data in promoting the Council of Australian Governments’ transparency agenda; and current gaps in the VET data collections. I will end with some comments on the challenges we face in maximising the use of data.

NCVER's collections

NCVER's regular collections comprise:

- students and courses (sometimes referred to as the provider collection), an annual administrative collection covering all government providers (mostly Technical and Further Education Institutes) and the publicly funded activity of private providers. Data are collected on over 2000 providers, covering 1.9 million students
- the apprenticeship and traineeship collection, a quarterly administrative collection of the numbers of apprentices and trainees commencing, in-training, completing and withdrawing, covering some half a million apprentices and trainees at any point in time
- the annual student outcome survey, a sample survey of VET graduates and others who completed their training but not completed a qualification. This survey alternates between a large year (some 90 000 graduates and 210 000 module completers – those who have left VET without completing a qualification) and a small year (20 000 graduates and 60 000 module completers)¹
- a VET in schools collection, an annual collection of data from senior secondary board of studies
- a finance collection, a compilation of data from the Commonwealth and state training authorities
- a biennial employer survey, collecting the views held by employers on the VET system and their use of the system.

In addition, a number of ad-hoc surveys have been conducted in recent years: a tracker survey of apprentices and trainees who have either completed their training or withdrawn from it; a survey of

¹ Response rates are around 45% for graduates and 35% for module completers.

labour market outcomes of young persons two and a half years after completion of training; and a student intentions survey, to look at the question of whether students beginning a course intend to complete it.²

The statistical information is disseminated in a relatively standard way: hard copy publications, electronic spreadsheets, data cubes and de-identified unit record files. The information is free, unless a client wants a special tabulation in which case clerical time is charged for. Most of the information is available for free on the web, although users have to register for Vocstats (a sophisticated table builder) and have to apply for de-identified unit record files. The Vocstats registration is not a device to restrict access. Rather, it enables us to ensure that users know how to use the product. In terms of access to de-identified unit record files, users have to sign an undertaking that they are using the file for a specified research or analytic purpose, to avoid any possibility that personal information would be matched with other data sets, as low risk as that is. There are plans to build a portal to enable users to interact with data directly, but this is some way down the track.

Using analytical tools to expand data use

Most data sets held by statistical agencies are either administrative collections or sample surveys. Typically, administrative collections are treated as censuses, with tabulations drawing directly on the data while survey data are weighted to population totals using a range of standard survey analytic techniques.

One way of getting better value from a set of data is to apply techniques from the mathematical sciences in order to create or derive information that is not available in the data. One example is the use of imputation techniques to fill in missing data (the ABS has used such techniques for many years). Another example is the use of synthetic estimation to derive estimates of small area statistics based on a combination of fine grained census data with survey data at a coarser geographic level (Department of Education, Employment and Workplace Relations 2012).

NCVER has used a number of techniques to meet the demand for information when the underlying data does not immediately provide straightforward statistics. Four examples are given here, all of which demonstrate how statisticians can apparently create something out of very little.

Estimates of apprenticeship activity

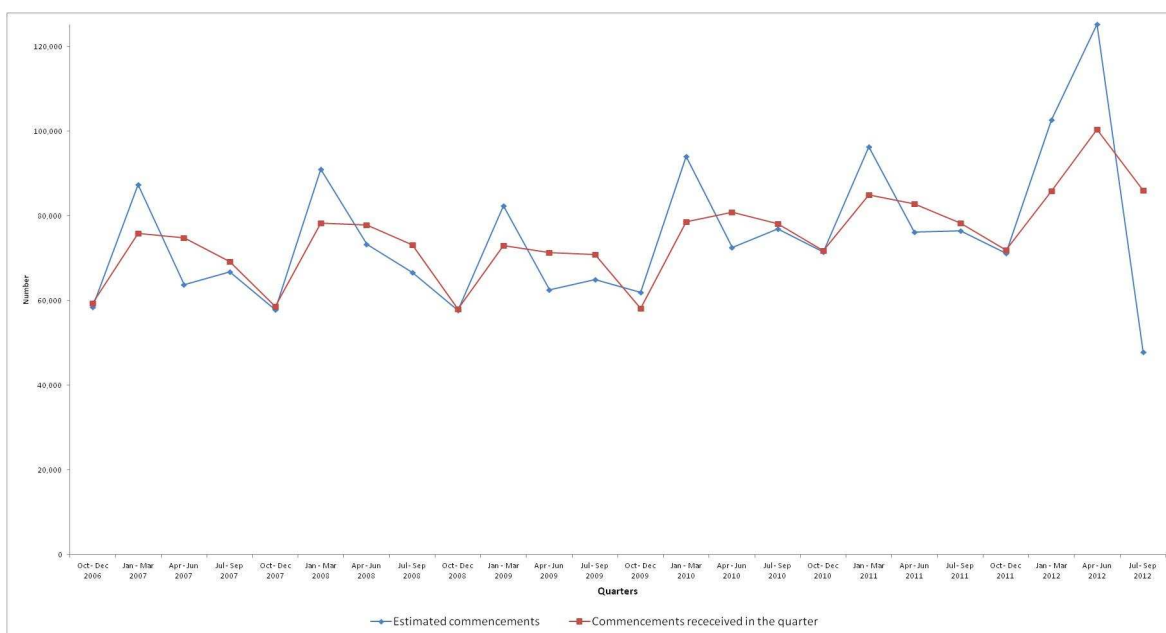
This is an interesting application of what appear to be standard sampling techniques. The problem is that there are serious lags in the process of collecting apprenticeship and traineeship data which occur because it takes considerable time for employers and state training authorities to lodge the data.

Without some sort of modelling approach we would have to publish 'received in the quarter'³ data. While one might think that there would be swings and roundabouts in such an approach, the effect of doing so is to create a phase shift in the data (figure 1).

² NCVER is also responsible for cleaning the data from the Longitudinal Survey of Australian Youth and loading it to the Australian Data Archive, but this collection is not part of the national VET statistics collections.

³ Because of lags data received in one quarter often refers to activity in an earlier quarter. Withdrawals/cancellations have the worst lags – up to eight quarters.

Figure 1 Estimated apprentice and trainee commencements for a quarter compared with commencements received in the quarter



This is very important in economic data for the simple reason that it is turning points that are of interest to policy makers. In figure 1 it can be seen that the ‘commencements received in the quarter’ distorts the seasonal pattern but also obscures structural changes. For example, the very rapid increase in commencements, followed by a dramatic decrease, in 2012 (due to a change in government incentives) would be distorted by the ‘received in a quarter’ approach. What makes this problem interesting is that we have to model administrative practice – this is not a simple sampling framework. The way we do this is to look at the ratio of reported activity to the final figure over a window that gives us eight observations. By assuming that these observations come from some underlying population (obviously a figment of our imagination) we can describe our estimation procedure in standard probabilistic terms, thus deriving ‘unbiased’ estimates and standard errors. The standard errors are particularly important because high standard errors alert us to instability in the underlying administrative practice. If the standard errors are high we inspect the eight observations of lags, looking for any structural shift in administrative practice. If it appears that there has been a shift we consider applying differential weights to the eight observations. (See Harvey 2010 for details on the methodology).

Estimating completion rates using Markov chains

For many years the Ministers with responsibilities for training included course completion rates in their list of key indicators for the efficiency of the sector. The lack of cohort data following students from commencement of a course to completion prevented the calculation of the indicator.

To fill this gap NCVER developed a modelling approach which exploits matched data over two years, and employs Markov chain methodology to come up with completion rates (Mark & Karmel 2010).

Applying mathematical logic to deriving completion rates for apprentices

One of the criticisms of the apprenticeship system is relatively low completion rates. However, the completion rates routinely estimated by NCVER are based on the contracts of training (a contract

between an individual, an employer, and a training provider) and some stakeholders argue that the completion rates are misleading because they do not allow for an individual changing an employer within a particular apprenticeship area. The direct calculation of completion rates for individuals (as distinct from contracts) is not possible because there are no data which allows identification of the first time an individual has commenced an apprenticeship. This lack of data has been overcome by collecting data on recommencements (that is the number of individuals moving from one employer to another) and applying mathematical logic to estimate the relationship between the individual completion rates and the contract completion rates (Karmel 2011).

Estimating the pay-off to an apprenticeship or traineeship

The apprenticeship and traineeship system is heavily subsidised by government. While it is difficult to argue with the proposition that apprenticeships and traineeships are a good method of training (combining formal training with direct experience), it is important to understand for which occupations the apprenticeship model is particularly beneficial. One of the issues in estimating the pay-off is coming up with a counterfactual – what the individual would be doing in the absence of the apprenticeship or traineeship. Based on a survey of the outcomes for both those who had and had not completed their apprenticeship, NCVET estimated the wages and employment outcomes for each individual, contingent on background characteristics such as education and field of the apprenticeship (Karmel & Mlotkowski 2011). By fitting models for both those who completed and did not complete their training, the pay-off to completion was calculated for each individual. This enabled the pay-off to be calculated for different types of apprenticeships and traineeships, and this was an important piece of evidence for the government in amending government subsidies.

Using life tables to create up to date apprenticeship completion rates

Apprenticeship completion rates are an important indicator of system performance, particularly relating to apprenticeships and traineeships. One of the problems with their calculation is that an apprenticeship can last up to four years, even when undertaken on a full-time basis. The standard way of calculating them is to follow a cohort of apprentices until they have all either completed or withdrawn from the apprenticeship. This entails waiting four years from commencement, with the implication the completion rate for that cohort may not be a good indicator of the completion rate for a newly commencing cohort. We have solved this conundrum by applying life tables' methodology, which involves calculating the completion rate over the last quarter for all cohorts currently in the system. By assuming that these transitional completion rates remain unchanged we can calculate the expected completion rate for the most newly commencing cohort (Karmel & Mlotkowski 2011).

Expanding the range of data on vocational education and training

One of the most heartening developments in recent years – from a statistician or analysts point of view at least – is the promotion of the 'transparency agenda' (Australian Government 2012). There are multiple motivations for this development. The first is the development of the training market with an emphasis on user choice. The second is accountability for very substantial government expenditure – governments want to know what they are getting for their 'investment in human capital', and the third is a desire to better understand the training market.

There is a range of developments now on the drawing board, the most important of which are listed below.

Collection of training activity from all registered training organisations

The VET students and courses collection derives from an administrative collection based on the public funding of students. While it captures all the activity of the government providers (mostly TAFEs) it only captures the government funded activity of private providers. Thus it misses out on the fee for service activity of private providers, and this activity has become increasingly more important. It has become clear that this partial collection was inadequate for regulators and policy makers. How can you regulate an industry or make sensible policy when a very substantial proportion of the activity is invisible?

A related issue is the need to collect performance or outcome data from all providers. Currently, the student outcome survey is restricted to the providers in the students and courses collection. It makes sense that any move to broaden that collection should be reflected in the coverage of the student outcome survey.

The publication of provider level data

Until recently, VET data was presented at an all of industry or state level. However, consumer choice and regulation requires information at the level at which training is delivered. Thus the VET sector is following in the footsteps of the long standing practice in higher education, and the more recent MySchools in the schooling sector, to publish data by provider. Some contextual data (numbers and types of students) have already been published by NCVET (2011) and the Department of Innovation in MySkills <www.myskills.gov.au>. The big challenge lying ahead will be the publication of performance information (and the development of performance indicators is something that is currently exercising NCVET).

The creation of a unique student identifier

The inability to track students over time has been a frustration for many years, particularly in the context of an emphasis of lifelong learning and on opening up 'pathways'. However, the lack of longitudinal data is an insufficient motivation for the creation of a unique student identifier (USI) – statistical matching can provide very serviceable data. There are two reasons to go down the USI path: the provision of an authenticated and comprehensive training transcript for individual learners, and the facilitation of individual based funding programs (that is entitlement models).⁴

The provision of more timely data

The world is becoming an impatient place. Currently, we publish, say, students and courses data for 2012 in July 2013. However, there is a growing expectation for real time data. This might be possible if all RTOs ran identical student management systems and all students enrolled online, but the reality is far from this. While 'real time' data is many years (decades or centuries?) away, there is a push for more frequent data submissions, and the development of quarterly reporting is under current consideration.

⁴ See <www.innovation.gov.au/Skills/National/UniqueStudentIdentifierForVET/USIConsultation/Documents/OverviewPaper.pdf> for details of this initiative.

Geocoding

Participation in VET by region and socioeconomic status has been difficult to measure. Clearly, regional participation depends on the geographic classifications embedded in our statistical collections. In the absence of individual socioeconomic status data our measurement of socioeconomic status has been based on the SEIFA index derived from the census. The application of this index depends on the geographic data held at the individual level. At the moment, postcode and locality data are collected and coded statistical local areas. In the future, it is planned to take advantage of the new ABS geographic classification (the Australian Statistical Geography Standard) and to geocode individual addresses to the statistical area 1 level, enabling a fine level application of SEIFA and the aggregation of data to a plethora of geographic regions.

Gaps in VET data

The most obvious gap in VET data is the lack of a comprehensive collection of training activity undertaken by all RTOs. As noted earlier, this situation has arisen because of the genesis of the current students and courses collection as an accountability tool for the disbursement of government funds. The implication of this partial collection, as large as it is, is that neither policy makers nor regulators really understand what is going on in the training market. There is now recognition of this situation and Ministers have agreed to a total collection and it is planned to begin a comprehensive collection in 2014. The foundation for the comprehensive collection is that provision of data compliant with the data AVETMIS (Australian Vocational Education and Training Management Information System) Standard will be mandated for registered training organisations delivering Australian Qualification Framework vocational education standards. While there has been some resistance from private training organisations (on the basis that private fee for service activity is only the business of the RTO), the policy view is that being able to issue accredited qualifications is a privilege and that the provision of data is a small price for this privilege.

A related gap is the lack of outcome data for graduates from private providers. The student outcome survey has been tied to the provider collection and so its coverage has reflected the coverage of the provider collection. A push to collect performance information on all providers means that this gap needs to be addressed. One promising approach here would be to control fieldwork costs by expanding the online version of the survey, and put the responsibility of getting graduates to fill in the survey instrument on to the RTO.

These are not the only gaps, though. Another weakness in the official VET statistics is the finance collection. The current collection is basically a compilation of government budget accounts (of funds going through state training authorities) and is hardly comprehensive. It is planned to review the collection, to see whether it is possible to broaden the collection to better capture the finances of VET. One conceptual issue to be thought through is the scope of training expenditure. For example, should incentives going to employers for taking on an apprentice or trainee be included? The difficulty of broadening the finance collection should not be underestimated; it is one thing mandating the provision of student and course data, it is another thing delving into the finances of providers.

A further gap is the lack of data on the VET workforce. This is important because the skills and qualifications of the workforce are key aspects of the quality of provision. The lack of data has basically arisen because in times past state governments have argued that the staff at government providers is an internal matter, since the government is the employer. The growth of the private training providers has made this a less convincing argument. NCVET has made a small step in filling

this gap by developing a data standard for the VET workforce. There has not, though, been any real consideration of a collection to date.

The final possible gap that is commanding attention is consumer information about RTOs. The current student outcome survey collects much relevant data: student satisfaction, employment outcomes, whether the training met the learner's objectives, whether the training is relevant to current employment and so on, but it is still worthwhile considering other types of information. For example, the higher education sector has a graduate destination survey, but also has a student engagement collection while students are still studying.

Concluding comments

On the whole, from an official statistician's point of view things are progressing positively in the VET sector. There is a clear recognition of the importance of data for the four key areas: data to inform the market, data to assist regulation, data to assist the delivery of government funding and data for policy development.

Progress, however, can be slow. I have been in my current role as managing director of NCVET for over ten years now, and many of the current data issues were issues ten years ago. I also discern a resistance to data collection. While cost is always a legitimate issue the mantras of 'commercial in confidence' and 'protection of privacy' are often employed to make the collection of data difficult. This also translates to the dissemination of data. While I absolutely agree that we should not give out unit record data with an individual's name and address, I think we worry too much about possible identification. Does a detailed statistical table with a cell size of one really tell anybody anything about an individual? If you knew all the characteristics of an individual then I guess you would deduce that there was no other person with those characteristics (if you assume that the data collection is perfect). But in this case perhaps we should not allow cells with no data as well.

The 'privacy' card has also been played to limit the use of data even at an aggregated level. It is now not possible to use data for a purpose which somehow strays outside the original purpose of collecting data. This is a great way of restricting information and ensuring that creative analysts cannot do things that program managers had not thought of. While privacy is a legitimate concern my view is that the pendulum has swung too far in that direction, at the cost of furthering knowledge and understanding of our society.

My point is that de-identified data sets are rather harmless artefacts; invaluable to the analysts but of little use to private detectives and debt collectors. Does it really matter how the data were collected? Of course we should be worried about abuse of information and identity theft and so on. But de-identified data should be viewed as pretty harmless and a resource to be used for analysis.

In terms of access to data, there is a worrying trend to limit use. Perhaps this is an outcome of the privacy lobby. For example, NCVET used to have access to a de-identified unit record file from higher education, but this has now ceased on the basis of legal advice that the data was being used for a purpose not intended by the original legislation. The legislation has been altered to allow certain parties access to the data, but governments have been given a privileged position relative to non-government researchers. In whose interest is this? Who knows what insights might have been gained by a researcher analysing the data?

My final comment relates to data mining. Data has become relatively cheap and computing power even cheaper. Good analysis though does not come at the push of a button or a swipe of a touch

screen. Rather it depends on a good understanding of the underlying issue and relevant theoretical frameworks, a good understanding of the data and a sound knowledge of statistical methodology. We need to be careful to guide users in their use of data. Computing power does not equal understanding (and if you look hard enough something will be statistically significant). But I would say that wouldn't I, given I trained as a mathematical statistician?

References

Australian Government 2012, Skills for all Australians: greater transparency of the VET sector, viewed 4 March 2013 <www.dpmc.gov.au/publications/skills_for_all_australians/chapter6_greater_transparency_of_the_vet_sector.html>.

Department of Education, Employment and Workplace Relations 2012, *Small area labour markets*, September quarter 2012, DEEWR, Canberra.

Harvey, B 2010, *Estimation of apprentice and training statistics*, NCVER, Adelaide.

Karmel, T & Mlotkowski, P 2010 *Estimating apprentice and trainee completion and attrition rates using a 'life tables' approach*, NCVER, Adelaide.

Karmel, T & Mlotkowski, P 2011, *The impact of wages and the likelihood of employment on the probability of completing an apprenticeship or traineeship*, NCVER, Adelaide.

Karmel, T 2011, *Individual-based completion rates for apprentices*, NCVER, Adelaide.

Mark, K & Karmel, T 2010, *The likelihood of completing a VET qualification: a model based approach*, NCVER, Adelaide.



National Centre for Vocational Education Research Ltd
Level 11, 33 King William Street, Adelaide, South Australia
PO Box 8288, Station Arcade, SA 5000 Australia
Telephone +61 8 8230 8400 Facsimile +61 8 8212 3436
Website www.ncver.edu.au Email ncver@ncver.edu.au