DISCUSSION PAPER

NCVER

Data linkage in VET research: opportunities, challenges and principles

Kristen Osborne, Craig Fowler and Michelle Circelli
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: data analysis; data collection; research methods; statistical data; vocational education and training.

This document should be attributed as Osborne, K, Fowler, C, Circelli, M 2018, *Data linkage in VET research: opportunities, challenges and principles,* 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 and Training.

COVER IMAGE: GETTY IMAGES/iStock

ISBN 978 1 925717 08 2
TD/TNC 130.09

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

Data linkage in VET research: opportunities, challenges and principles

Kristen Osborne, Craig Fowler and Michelle Circelli, National Centre for Vocational Education Research

This discussion paper explores the possibilities and risks that data linkage presents for the vocational education and training (VET) sector. Along with a broad overview of the nature of data linkage, it highlights possible applications for data linkage in the VET sector and examines the key challenges associated with its use.

A number of case studies are reviewed to illustrate the advantages data linkage can offer, as well as the challenges that may arise. In order to better understand the options for data linkage from an education and employment perspective, a ‘map’ of relevant Australian datasets is presented, along with a list of data sources that may be of use to VET research. As well as selected Australian datasets, the paper reviews some international datasets of potential interest for VET research in Australia.

Using the insights gained from past data-linkage projects and taking into account the privacy and ethics concerns, the paper presents a set of six principles for data linkage. These principles provide researchers with basic steps for guidance when embarking upon a data-linkage project. Finally, future directions for data linkage in VET research are explored.

Key messages

- Data linkage is a powerful tool for research and policy in the VET sector, with a number of past research projects demonstrating its feasibility and utility.
- A number of challenges are associated with data linkage, chiefly privacy and ethics concerns, but additionally issues such as:
  - the various data custodians to be dealt with
  - the legislative environment in the area of data linkage
  - the costs associated with any data-linkage project
  - the availability of technical skills and infrastructure resources.
- Data linkage can be used to capture a more complete picture of an individual’s lifelong journey through education and employment, including their pathway through VET. The success of this depends on the cooperation of data custodians to foster linkage.
- Examples of future projects in the VET sector that may benefit from the use of data linkage include investigation of long-term student outcomes, analyses of the return on investment for particular programs, or examination of childhood and youth factors as they relate to VET study.

Dr Craig Fowler
Managing Director, NCVER

Contents

Tables and figures

Tables

Figures

Introduction

This paper presents an overview of data linkage, profiling both the opportunities and challenges.

Data linkage is a powerful and useful tool, and its use is expanding in the area of research and policy. Accordingly, the meaningful and ethical use of data in vocational education and training (VET) research calls for an assessment of both the opportunities and challenges of data linkage. In an overview of data linkage, this paper looks at these issues. It also includes a 'map' of selected resources for VET-related data linkage and a set of research principles to guide data-linkage activities, together with a selection of case studies and examples of relevant datasets and related initiatives (given in the appendix). The paper aims to be both a primer on data linkage and a source of future directions in the VET sector for this invaluable resource.

What is data linkage?

There are two broad types of data linkage: deterministic and probabilistic.

Deterministic

- Deterministic data linkage involves connecting data relating to an individual entity (for example, person, organisation or region) from different sources, such as administrative datasets or surveys, to form a new dataset using a given set of one or more identifiers that create a linkage key (see below definition for identifiers; National Statistical Service nda; Dusetzina et al. 2014)
- Connecting or linking identifier information that can be used to form a new dataset includes combinations of demographic details (for example, name, date of birth, gender) or other sources of connecting information such as unique identifiers (for example, medical record or student number). These variables are normally used to construct a unique linking identifier (linkage key) for each entity within the dataset.

Probabilistic

- Where a single and unique linking identifier cannot be constructed for each entity within the dataset, probabilistic data linkage uses characteristics of entities in separate datasets to statistically match the information for individuals (Sayers et al. 2016, pp.954—64).
- Essentially the datasets are linked using an algorithm that attempts to determine the probability that any two records in the datasets are a match and ranks the matches by probability. Cut-off thresholds must be developed, with matches indicated if the probability is above the cut-off. Although useful when unique identifiers are not possible, this method does rely on probability to assign matches rather than creating unique matches based on an individual's identifying information.

There are also options to connect datasets by time period, such as information on VET training completions and job vacancy data over a period of years. This type of matching is much more general, but can be used to assess broad trends and similarities across two areas of information.

Links can be made across different types or formats of data, and between multiple datasets. A link between National VET Provider Collection data, in order to retrieve an individual's training activity, and the Australian Bureau of Statistics (ABS) Census of

Population and Housing data, for information on income and housing status, is an example of a link between an administrative dataset and a survey dataset. Both of these could also be linked to a longitudinal survey, such as the Longitudinal Surveys of Australian Youth (LSAY), which could offer information on transitions from school and aspirations, or test results such as the National Assessment Program — Literacy and Numeracy (NAPLAN), which provides a marker of academic performance in school. Theoretically, as long as there is enough identifier information in common between two or more datasets (or they share a linking key), they can be linked. However, the quality of the original data and the approach taken to linking them will have a substantial impact on the quality of the results and the validity of the linked data for analyses.

Data linkage offers the opportunity for enhanced analyses and has the potential to provide valuable multifaceted insights for policy and future research.

Data linkage often involves the use of longitudinal datasets. A longitudinal dataset is one that involves ongoing engagement with individuals over a period of time, enabling the tracking of individual experience and outcomes. One example is LSAY, mentioned above, a study that interviews young Australians from age 15 to age 25 as they move through secondary school, employment and further education. Such longitudinal studies are a vital part of understanding population experiences. The recent Commonwealth review, found these longitudinal data assets to have delivered ‘significant public value’ (Department of Social Services 2016). The integration of longitudinal and non-longitudinal sources can result in even greater overall value from all datasets involved.

Why link data?

Data linkage offers the opportunity for enhanced analyses and has the potential to provide valuable multifaceted insights for policy and further research. The health sector in Australia has been one of the leaders in data linkage, with the Population Health Research Network[1] established to facilitate the linking of health data with education, social, child welfare and criminal justice data, with the aim of informing national policy.

Other data-linkage activities using data held by different government agencies have enabled better insights into the Australian economy. For instance, the Business Longitudinal Analysis Data Environment (BLADE, previously known as EABLD) links governmental administrative data from organisations such as the Australian Tax Office (ATO) with data from ABS surveys of Australian entities. This has proved useful in examining the role of entrepreneurial start-up companies in job creation (Department of Industry 2015).

The Department of Human Services also currently uses data matching as part of its fraud and non-compliance strategy, matching the records of Centrelink clients with information from the ATO in a program designed to detect individuals who are receiving incorrect payments (Department of Human Services 2017).

As well as enabling research to be undertaken, data linkage can reduce the need for developing new national surveys, or extend the use of existing national surveys, creating efficiencies in the utilisation of resources while reducing respondent burden. While survey data are limited by the quality of participants’ recall, there is no such limitation with historical data in administrative datasets. This is the case as long as a process of regular system downloads is preserving the real-time administrative record system, and it must also be acknowledged that administrative data entries are not error free. Linking

[1] <http://www.phrn.org.au/>.

data also provides opportunities to use surveys to obtain information that cannot be obtained from administrative data alone, answering the ‘how’ questions rather than the ‘what’.

In the VET sector specifically, linking data from the National VET Provider Collection (which includes information on VET students, program enrolments, subject enrolments, program completions and source of funding) with surveys such as the ABS Census, the National Health Survey or the ABS General Social Survey, or with proprietary data such as that collected by employment organisations such as SEEK, can provide more in-depth knowledge of pathways through, outcomes from, and the impact of, VET on the different aspects of an individual’s life.

Challenges

While linking data has clear benefits, it must be a highly considered process. Different datasets require different linkage approaches, and the style of linkage must be appropriate to the variables and values present in all of the proposed sets. Additional complexities include:

- *Dealing with various data custodians*: each of these may have different concerns, legal requirements and approval processes. These processes can often be lengthy and may require ongoing discussion about the nature of the project and data to be used. The cooperation between agencies involved is a key factor in the success of any linkage project, and building goodwill with these agencies is vital. Some data custodians, such as the Commonwealth, require a third party Accredited Integrating Authority to be involved during high-risk projects[2] (National Statistical Service ndb). An Accredited Integrating Authority is an organisation that has been accredited by the Cross Portfolio Data Integration Oversight Board of the National Statistical Service to undertake high-risk data-integration projects. Only three such organisations have so far been accredited: the ABS; the Australian Institute of Health and Welfare; and the Australian Institute of Family Studies.
- *Ensuring compliance with all relevant state/territory and local government legislation and policies*: some jurisdictions have taken steps to amend legislation to enable their data to be used more effectively for research and data linkage. In 2015, New South Wales amended its Privacy and Personal Information Protection Act to ‘allow public sector agencies to disclose personal information to interstate persons or bodies or Commonwealth agencies for certain purposes, and to collect, use and disclose personal information for certain research purposes’ (New South Wales Parliament 2015). However, the capacity of legislation to keep pace with technological developments and possibilities in this area is limited, which leaves some aspects of data linkage in a legal ‘grey area’, to be navigated as best as possible.
- *Resourcing costs associated with data linkage*: these vary depending on the data and methods involved, particularly if specific technical skills or software are required. Data can be stored and recorded in a variety of ways, meaning it may be difficult to

[2] High-risk projects are defined by the National Statistical Service as those which have a post-mitigation ‘“high” risk of harm to data providers (including persons, families, households, or organisations that have contributed data) or a loss of public trust in the Australian Government or its institutions’ (National Statistical Service 2013).

All data-linkage projects must consider often complex legal and/or ethical concerns to ensure privacy compliance.

access the data in the required format. Researchers must also navigate the difference in data quality between the datasets and possibly deal with corrections or updates to linked datasets when they occur. Various infrastructure and technical skill requirements must be met in order to resource the linking of data. It may take time to implement these conditions appropriately if an organisation pursuing linkage does not have them in place; furthermore, these resource requirements will vary between projects and may change over time.

- *Security of and access to linked data*: how to secure the data, and whether they can be made available for others to access, needs to be determined. While it may be more desirable to retain a linkage dataset for future use and dissemination, there are privacy concerns associated with making the linked dataset more widely available. Some datasets may be confidentialised and kept for future use, while others need to be destroyed after the research has been completed. Although destroying the dataset after use avoids the need to maintain its security, it also leads to the loss for future analyses of valuable linked datasets, into which resources have been invested.

These challenges mean that not all proposals for data linkage are successful. In 2013, the Standing Council on School Education and Early Childhood (now the Council of Australian Governments [COAG] Education Council) and the Standing Council on Tertiary Education, Skills and Employment (now the COAG Industry and Skills Council) endorsed the Transforming Education and Training Information in Australia (TETIA) proposal (ABS 2013). TETIA was aimed at creating an integrated, coordinated and high-quality education-focused database, whose objective was to improve education and training in Australia at all levels (ABS 2013). The project, despite the efforts of those involved, has never progressed to the point of usability. Other planned data linkage proposals may now supersede this.

Privacy, ethics and data linkage

A particularly important challenge for data linkage is the protection of the privacy of the individuals within the linked datasets, and all data-linkage projects must consider often complex legal and/or ethical concerns to ensure privacy compliance. For linkage activities that use Commonwealth data for statistical and research purposes, seven principles for data integration apply[3] (National Statistical Service 2010). The seven principles include undertaking a risk assessment and managing or mitigating the risks where possible. The Office of the Australian Information Commissioner also provides a set of guidelines for data linkage for matching across Australian government agencies[4] (Office of the Australian Information Commissioner 2014), which are more technical in nature than the Commonwealth principles. There remains a lack of national consensus on the exact ethical principles or guidelines that should guide data linkage.

Options are available for mitigating the privacy risks; for example, by obtaining consent from individuals in the datasets; creating a one-off rather than an enduring linkage; and establishing limited access to the linked dataset. The National Centre for Vocational Education Research (NCVER) always undertakes a privacy impact assessment prior to any

[3] See <http://www.nss.gov.au/nss/home.NSF/pages/High+Level+Principles+for+Data+Integration+-+Content?OpenDocument> for full list of principles.

[4] See <https://www.oaic.gov.au/agencies-and-organisations/advisory-guidelines/data-matching-guidelines-2014> for full list of guidelines.

data-linkage research, a practice which is vital to safeguarding the privacy of those involved.

Data custodians may also require a data-linkage project to gain initial approval from a Human Research Ethics Committee, established within the guidelines of the National Statement on Ethical Conduct in Human Research. This process requires researchers to explain the project and its potential risks to, and impacts on, the individuals involved. It may be necessary for researchers to adapt risk management techniques in order to reduce the possibility of adverse effects for individuals.

Consideration of community attitudes towards data linkage and privacy is an important part of assessing the future of data linkage. The recent Australian Community Attitudes to Privacy Survey, run by the Office of the Australian Information Commissioner, found that 46% of Australians are comfortable with having their information used by government for research or policy-making, with 40% not comfortable and the remainder unsure (Office of the Australian Information Commissioner 2017). The survey results indicate that this area is a sensitive one, and care should be taken not only to ensure the privacy of individuals, but also to remain transparent in how this is being achieved.

There have been government moves to offer the general public greater protection in the area of data and personal information. A recent amendment to the Privacy Act now requires a range of organisations, agencies and other entities to notify the Australian Privacy Commissioner in the event of some types of data breach, such as negligent disclosure, accidental loss of data or a cyber security-related incident (Australian Parliament 2016). Some states, such as Victoria, have introduced data-protection acts aimed at safeguarding personal information and ensuring data security (Victorian Parliament 2014). There continues to be an ongoing review of privacy legislation Australia-wide in this area, and researchers must pay careful attention to ensure compliance with the current law both at Australian Federal and state/territory level.

Another ethical issue is that of informed consent. Consent from participants is an important part of any research. With data linkage, special care must be taken to inform individuals about potential linkages that may occur with their data, and what these linkages may be used for. It is unacceptable ethically (although there is no legal requirement to gain informed consent) to simply take previously collected data and to embark upon a program of linkage without regard for the individuals from whom the data were collected. However, there are some situations in which consent cannot feasibly be sought, usually with administrative or historical datasets such as ABS Census data, in which case there still may be ethical options for linking available.

Through the linkage of various data assets, the burden on students to provide the same information to different surveys or research projects can be reduced.

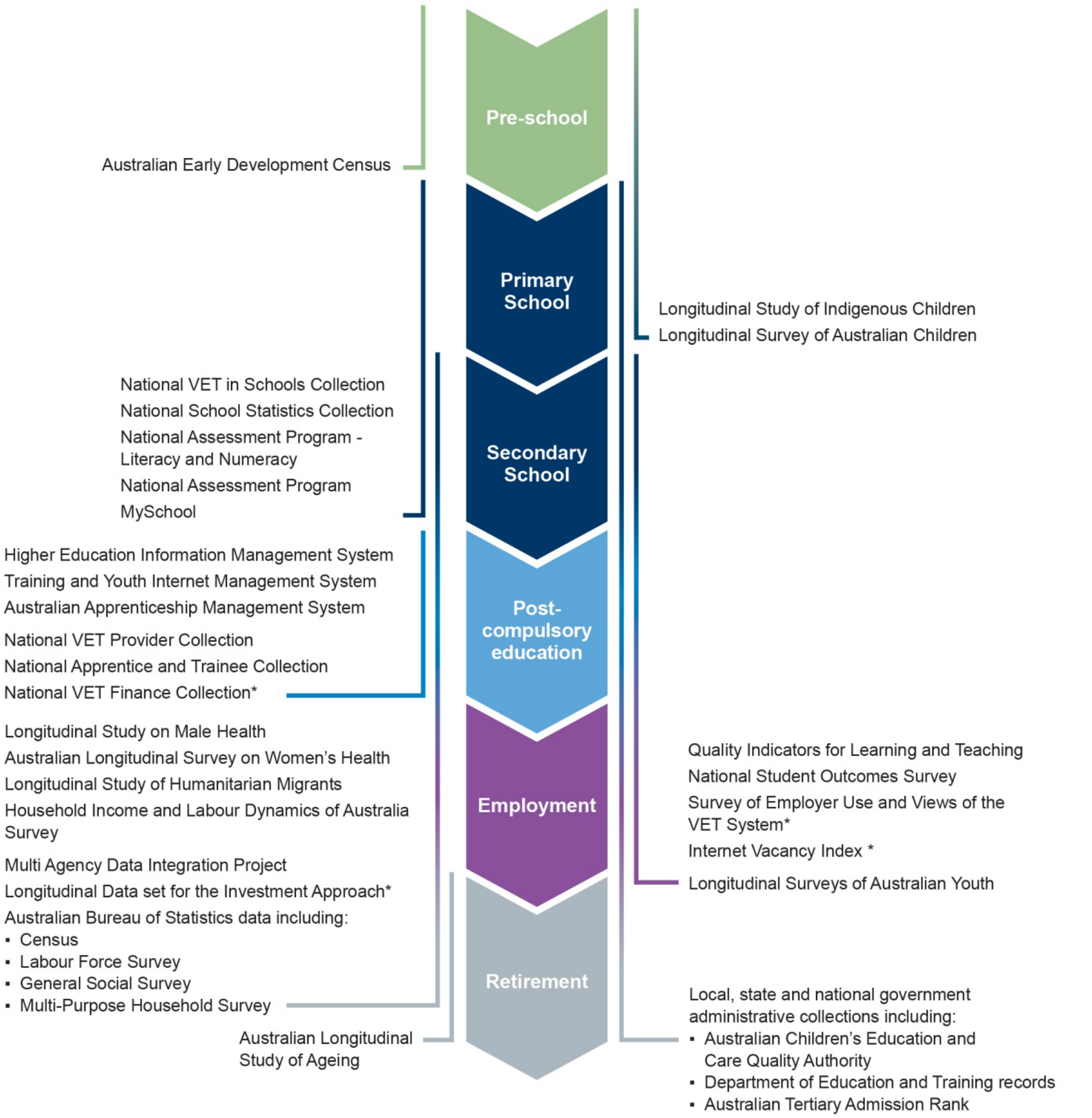
Linking data: an educational and employment view

Data linkage has specific benefits to the VET sector, and the educational community in general. The linkage of various data assets means that the burden on students to provide the same information to different surveys or research projects can be reduced. The range of VET-related data sources available are diverse and varied. The linkage of these existing sources then enables a better understanding of an individual's lifelong journey through education and employment, and their involvement with the VET sector along the way.

Figure 1 shows various national datasets and surveys that may be involved in the education and employment story of an individual, from birth to death. It must be noted that the list of datasets and surveys included in figure 1 is not intended to be exhaustive, a brief description of these datasets and surveys is given in the appendix.

As well as this broad overview, it is useful to illustrate with some past examples of successful data linkage, as employed in VET and related fields. For this purpose, three case studies are presented that describe the linkage of various survey, test and administrative datasets, together with the challenges involved and the potential applications to future settings.

Figure 1 Examples of data collected throughout an individual’s lifespan relevant to education and employment


Pre-school
Primary School
Secondary School
Post-compulsory education
Employment
Retirement
Australian Early Development Census
National VET in Schools Collection
National School Statistics Collection
National Assessment Program - Literacy and Numeracy
National Assessment Program
MySchool
Higher Education Information Management System
Training and Youth Internet Management System
Australian Apprenticeship Management System
National VET Provider Collection
National Apprentice and Trainee Collection
National VET Finance Collection*
Longitudinal Study on Male Health
Australian Longitudinal Survey on Women’s Health
Longitudinal Study of Humanitarian Migrants
Household Income and Labour Dynamics of Australia Survey
Multi Agency Data Integration Project
Longitudinal Data set for the Investment Approach*
Australian Bureau of Statistics data including:
▪ Census
▪ Labour Force Survey
▪ General Social Survey
▪ Multi-Purpose Household Survey
Australian Longitudinal Study of Ageing
Longitudinal Study of Indigenous Children
Longitudinal Survey of Australian Children
Quality Indicators for Learning and Teaching
National Student Outcomes Survey
Survey of Employer Use and Views of the VET System*
Internet Vacancy Index *
Longitudinal Surveys of Australian Youth
Local, state and national government administrative collections including:
▪ Australian Children’s Education and Care Quality Authority
▪ Department of Education and Training records
▪ Australian Tertiary Admission Rank

* Although these collections do not include individual level information, they are included for their relevance to an individual’s education journey.

International datasets

In addition to Australian datasets, numerous international datasets are available that could be used for data linkage. These international datasets offer valuable opportunities to use high-quality internationally validated data on Australians through data linkage, applying the six principles for data linkage research we subsequently set out in figure 3.

Some data integration with international datasets already exists. For example, the Longitudinal Surveys of Australian Youth (LSAY) recruits participants who have taken part in the Programme for International Student Assessment (PISA) and contains PISA achievement scores as a result of this partnership. This is an integration of the starting sample for LSAY with PISA, not an example of standard data linkage.

Other relevant datasets include the Survey of Adult Skills, run by the Programme for International Assessment of Adult Competencies (PIAAC), the Teaching and Learning International Survey (TALIS), the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS) and the International Computer and Information Literacy Study (ICILS). Australians are included among those surveyed for all of these datasets, making them particularly relevant to Australian data-linkage research.

PISA, PIRLS, TIMSS and ICILS have similarities to the Australian NAPLAN, in that they assess school-age students in various academic areas, such as reading ability or science and mathematics achievement. They can therefore assist researchers and policymakers to build an understanding of a student's abilities and performance throughout the schooling process. Options for linking these datasets might be to secondary school results, to information on students in the National VET Provider Collection or to Higher Education Information Management System data. The Teaching and Learning International Survey targets teachers and school principals, offering insight into the learning environment, which might, for example, be linked to existing school administrative data, as was done in England with the 2013 TALIS data (United Kingdom Department for Education 2014). The International Computer and Information Literacy Study addresses the increasing importance of computer literacy skills, and through data linkage may offer a way to track how those with high levels of these skills move through the education and employment space, compared with those with lower development of these skills. The Survey of Adult Skills (PIAAC) collects data on the academic skills of literacy and numeracy (and other skills) among those over 16 years of age. Unlike those still in school, there is no routine Australian-based testing of these skills in adults, making PIAAC a useful resource. Linkage with this dataset might include ABS Census data or other governmental administrative collections to gain a better understanding of the employment and health outcomes for these individuals.

As well as linking existing datasets to international datasets, further opportunities exist similar to that of the LSAY—PISA sample relationship. If longitudinal studies recruit their sample from those taking part in an existing or concurrent study, agreements to undertake data linkage can be arranged at the outset to enable a richer dataset. This also aims to reduce the initial costs of developing a longitudinal study.

Case studies

In this section, a selection of case studies is presented. These three case studies demonstrate some of the different types of approaches, challenges and outcomes that data-linkage projects can encompass. They were selected to provide an illustrative example of their linkage approach and their relevance to the VET sector. Each study has different insights to offer into consent as it is applied in research. Of note is the value each report provided in itself, by substantiating and recording for the public record their method and results.

The first case study demonstrates issues relating to consent and data matching, with consent personally sought from all of the individuals involved. The second case study is a recent example of the use of existing large-scale de-identified datasets to gain longitudinal insights into a population. The third case study shows an example of research potential where a permanently linked dataset is available.

Each of these case studies offers a different insight into data linkage. Table 1 shows the key similarities and differences in the cases examined here.

Table 1 Linkage case studies

	Case study	Value of linkage	Type of datasets linked	Data sample	Match rate
1.	*Linking NAPLAN scores to the Longitudinal Surveys of Australian Youth* (Lumsden et al. 2015)	Enables an assessment of whether there is agreement between the NAPLAN and PISA distributions; Demonstrates the values of linking early education achievement data and data on young people's transitions and education and employment outcomes	Test (NAPLAN) and survey/test (LSAY/PISA)	Sub-sample of LSAY cohort	98.1%
2.	*VET in Schools students: characteristics and post-school employment and training experiences* (Misko, Korbel & Blomberg 2017)	Enabled longitudinal follow-up on employment outcomes for VET in Schools students	Administrative (National VET in Schools Collection) and survey (ABS Census of Population and Housing)	Students present in the 2006 National VET in Schools Collection and the 2011 Census of Population and Housing	50.5%
3.	'Do thin, overweight and obese children have poorer development than their healthy-weight peers at the start of school? Findings from a South Australian data linkage study' (Pearce et al. 2016).	Compared a longitudinal cohort based around the Australian Early Development Census (AEDC) to reveal relationships between weight, development and the domains measured on the AEDC.	SA–NT DataLink used, specifically survey (AEDC) and administrative (pre-school health checks, perinatal hospital records and the student school enrolment Census).	Children with records linked through the SA–NT DataLink.	99.5% in this case, each linkage using SA–NT DataLink. DataLink is comprised of different datasets and match rates or false links vary.

Case study 1

Linking NAPLAN scores to the Longitudinal Surveys of Australian Youth (Lumsden et al. 2015)

This project explored linking NAPLAN scores to Longitudinal Surveys of Australian Youth (LSAY) data, and used this link to compare the NAPLAN and Programme for International Student Assessment scores of students. The majority (98%) of consenting LSAY participants were linked with their NAPLAN results, and an analysis showed that NAPLAN and PISA results do have reasonable agreement (refer to the appendix for explanation of the various datasets involved in this project).

A risk assessment and mitigation strategy was used, focusing on the likelihood of a breach of confidentiality and privacy, and any consequences of such a breach. This process involved consultation with the National Statistical Service, the Australian Institute of Health and Welfare, LSAY and NAPLAN data custodians (the Australian Government Department of Education and Training, and various state/territory test administration authorities), and the survey fieldwork contractor. The level of risk present and the proposed mitigation strategies determined whether an outside Accredited Integrating Authority needed to be involved in the linkage process. The project was given a low risk rating and accordingly an Accredited Integrating Authority was not required.

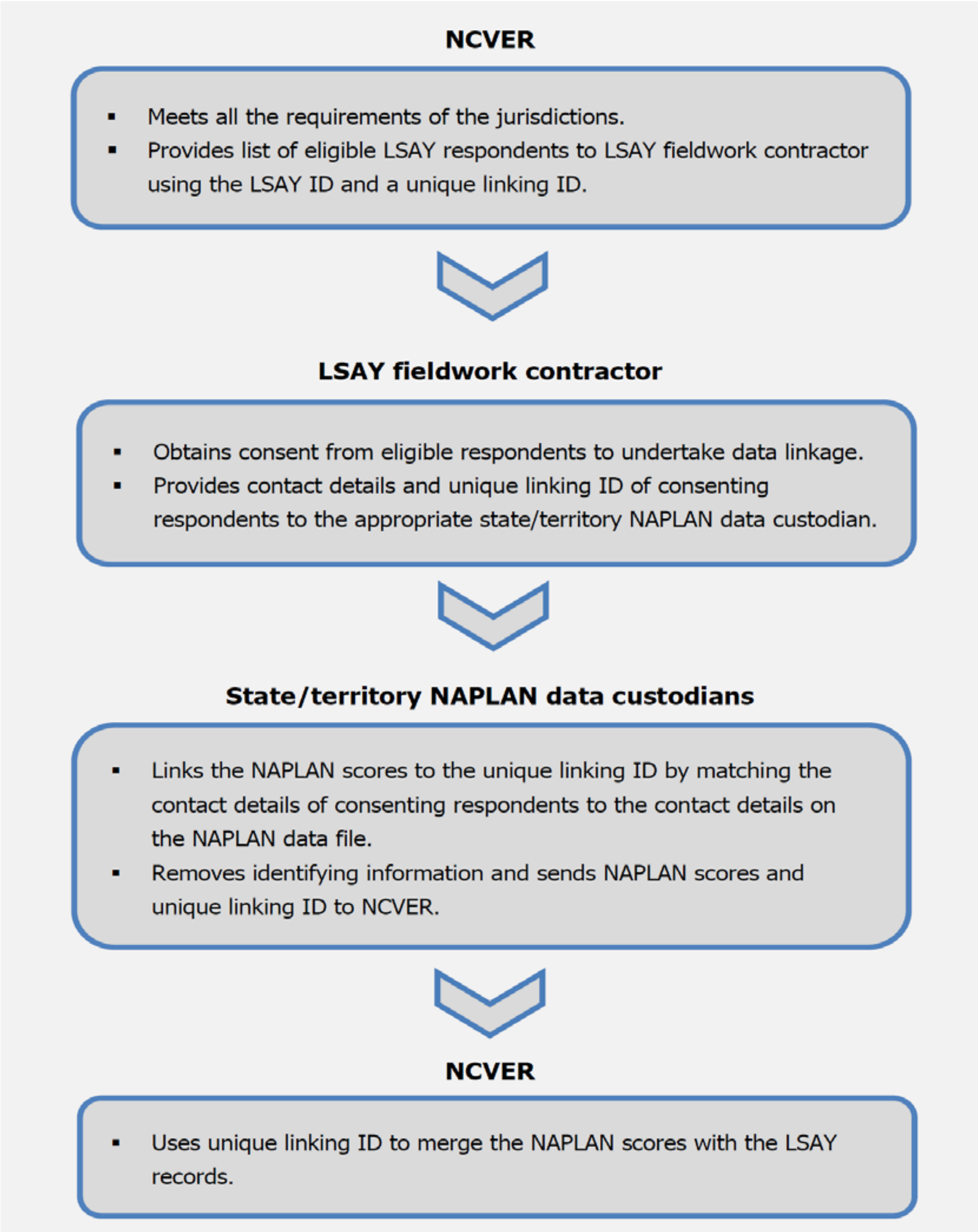
Participants were eligible for linkage if they were in Year 10 in 2009 (having completed NAPLAN in 2008 when these students were in Year 9) and remained in the LSAY sample until 2014. A set of 3800 individuals were eligible for consent, with 1644 participants asked for their consent based on a number of interview, geographical and timing factors. A variety of methods were used to gain consent, with participants asked permission to link their NAPLAN results with their LSAY (and therefore PISA) data. Written consent was sought prior to the LSAY interview, while oral and online consent was collected as part of the computer-assisted telephone or computer-assisted online LSAY fieldwork interview. Of those asked, 732 (44.5%) provided consent. While there may have been individual hesitancy about data linkage, other factors, such as method of collection and the relatively short time window for returning consent forms, likely impacted on this response rate. Written methods for obtaining consent for data linking resulted in lower rates of consent than the use of telephone or online process. The respondents' reasons for declining consent were not available (we would recommend collection of these reasons for similar future work) and therefore any specific concerns could not be investigated.

The two datasets were deterministically linked, as illustrated in figure 2. Individuals were matched on first and last name, gender, month of birth, year of birth, school name and school postcode. Individuals who were missing the linking variables were excluded. A 98.1% matching rate was achieved across the two datasets, resulting in the linkage of data from 673 individuals. The matched dataset was used to analyse the level of agreement between PISA and NAPLAN scores for individuals, using both hierarchical regression and correlation. The results showed a strong correlation between scores on the two tests, an important outcome for future research.

The authors reflected on the issues that arose, including the difficulty of coordinating the many different jurisdictions involved in the project. The process of obtaining consent was described as lengthy, and the greater usefulness of seeking consent for linkage at the beginning of any longitudinal collection was underlined.

An extension of this linkage project could enable research connecting education outcomes with individuals' transitions to further study and employment. Joining multiple years of NAPLAN results to LSAY data creates a longitudinal dataset of academic performance, which can be considered alongside the information given by LSAY participants. Further directions identified by the authors included linkages of the LSAY data with the National Schools Statistics Collection, the ABS Census and Medicare administrative data.

Figure 2 Process of linking LSAY and NAPLAN datasets



NCVER

- Meets all the requirements of the jurisdictions.
- Provides list of eligible LSAY respondents to LSAY fieldwork contractor using the LSAY ID and a unique linking ID.

LSAY fieldwork contractor

- Obtains consent from eligible respondents to undertake data linkage.
- Provides contact details and unique linking ID of consenting respondents to the appropriate state/territory NAPLAN data custodian.

State/territory NAPLAN data custodians

- Links the NAPLAN scores to the unique linking ID by matching the contact details of consenting respondents to the contact details on the NAPLAN data file.
- Removes identifying information and sends NAPLAN scores and unique linking ID to NCVER.

NCVER

- Uses unique linking ID to merge the NAPLAN scores with the LSAY records.

Source: Lumsden et al. (2015).

Case study 2

VET in Schools students: characteristics and post-school employment and training experiences (Misko, Korbel & Blomberg 2017)

This research examined the employment destinations of VET in Schools (VETiS) students by linking data from the 2006 National VET in Schools Collection (held by NCVER) with data from the 2011 Census of Population and Housing (held by the ABS). VETiS students are individuals who undertake VET study as part of their secondary school certificate. Employment and further study outcomes are important to school leavers, and data linkage can offer insights into these outcomes (refer to the appendix for explanation of the various datasets involved in this project).

Four geographic variables (postcode, locality code and two statistical area codes) were used to link the two datasets deterministically, along with date of birth, age, gender and country of birth. This produced a match in the 2011 ABS Census dataset for 50.5% of the individuals present in the 2006 VETiS collection. Although this rate is lower than similar data linkages, it provided over 84 000 linked records for examination. The dataset was weighted for postcode, sex, age, Indigenous status and country of birth to address under-representation in the linked dataset, although some over-representations were present in the final dataset.

Because there was no non-VETiS control group in the linked dataset, the research could not make claims about the effects of the VETiS program on the individuals studied in terms of direct employment outcomes. However, it did enable broad examination of the outcomes that VETiS students achieved after their study. The research found that students of trade VETiS subjects (for example, construction, carpentry and furnishing) were those most likely to enter an occupation related to their VETiS study later in life and that a sizable proportion of VETiS students did indeed choose occupations that aligned with their VETiS study program. This shows data linkage, when well executed, can provide some of the benefits of a longitudinal survey without issues such as non-response and long-term funding concerns, although there are representation and bias concerns when a high linkage rate cannot be achieved. Even with approximately 50% of individuals in the dataset unable to be linked, 84 412 individuals were successfully linked and available for analyses, a number far greater than that achieved by any national longitudinal study.

The authors underscored the importance of future research utilising a research design with a suitable control group to enable examination of the effect of VETiS on student outcomes. This might be possible by linking administrative student data or the results of academic tests such as NAPLAN with ABS Census data, alongside the VETiS data, to form a three-way linkage. The linkage of multiple datasets can be a way to form a more complete picture of student education and employment experiences.

The linked dataset from this project continues to be made available by the ABS to researchers on application, with strict privacy controls. This enables future insights to be gained by a range of research organisations with different perspectives and goals.

Case study 3

‘Do thin, overweight and obese children have poorer development than their healthy-weight peers at the start of school? Findings from a South Australian data linkage study’ (Pearce et al. 2016)

This research used the SA—NT DataLink[5], a linked population dataset, which can be used for many different types of health research. For this project, data from the Australian Early Development Census (AEDC), pre-school health checks, perinatal hospital records and the student school enrolment Census were all used as part of a linked dataset, available via SA—NT DataLink (refer to appendix entry ‘SA—NT datalink’ for explanations of the various SA—NT datalink datasets involved in this project South Australian Perinatal Statistics Collection and South Australian School Enrolment Census).

As SA–NT DataLink is the linkage organisation, individual researchers do not have to coordinate data custodians and perform the linkage.

The SA—NT DataLink is matched deterministically, with data custodians for each of the included datasets providing identifying information such as name, age, gender and address. A linkage algorithm then matches the records across different datasets. This process produces a small number (approximately 0.1%) of false matches and some duplicate cases, which must be removed before analyses. As SA—NT DataLink is the linkage organisation, individual researchers do not have to coordinate data custodians and perform the linkage. Many of the projects using SA—NT DataLink draw from more than two datasets, either for direct research purposes or to adjust for confounding factors.

Using the linked data, researchers compared the developmental status of 7553 children who were underweight, healthy, overweight and obese, determined via body mass index (BMI). The authors commented that, to their knowledge, this was the first study to compare BMI to a global measure of childhood development. The use of data from the AEDC enabled researchers to compare scores across physical, social, emotional and cognitive development areas to develop a complete picture of the individuals within the dataset, while the data from the school census and perinatal birth records enabled adjustment for socioeconomic and perinatal health factors.

The results showed that neither children who were underweight or overweight had increased developmental vulnerabilities. However, obese children were vulnerable on a number of domains used in the AEDC; these were the physical health and wellbeing domain, and the social competence domain. This provided an important direction for policymakers in the area of childhood development and education, highlighting the needs of obese children. The authors additionally warn that children in the overweight category are considered at risk of becoming obese, and that the results indicate the importance of supporting healthy bodyweight in children to protect their development.

The SA—NT DataLink is available for researchers on application and has been used for over 50 published projects, with many currently in progress. These range in scope from childhood health to cancer treatment and survivors, to kidney dialysis patients. The SA—NT DataLink demonstrates the value of an organisation facilitating good-quality data linkage and allowing researchers to access a range of relevant datasets.

[5] See <https://www.santdatalink.org.au/>.

Principles for high-quality data-linkage research

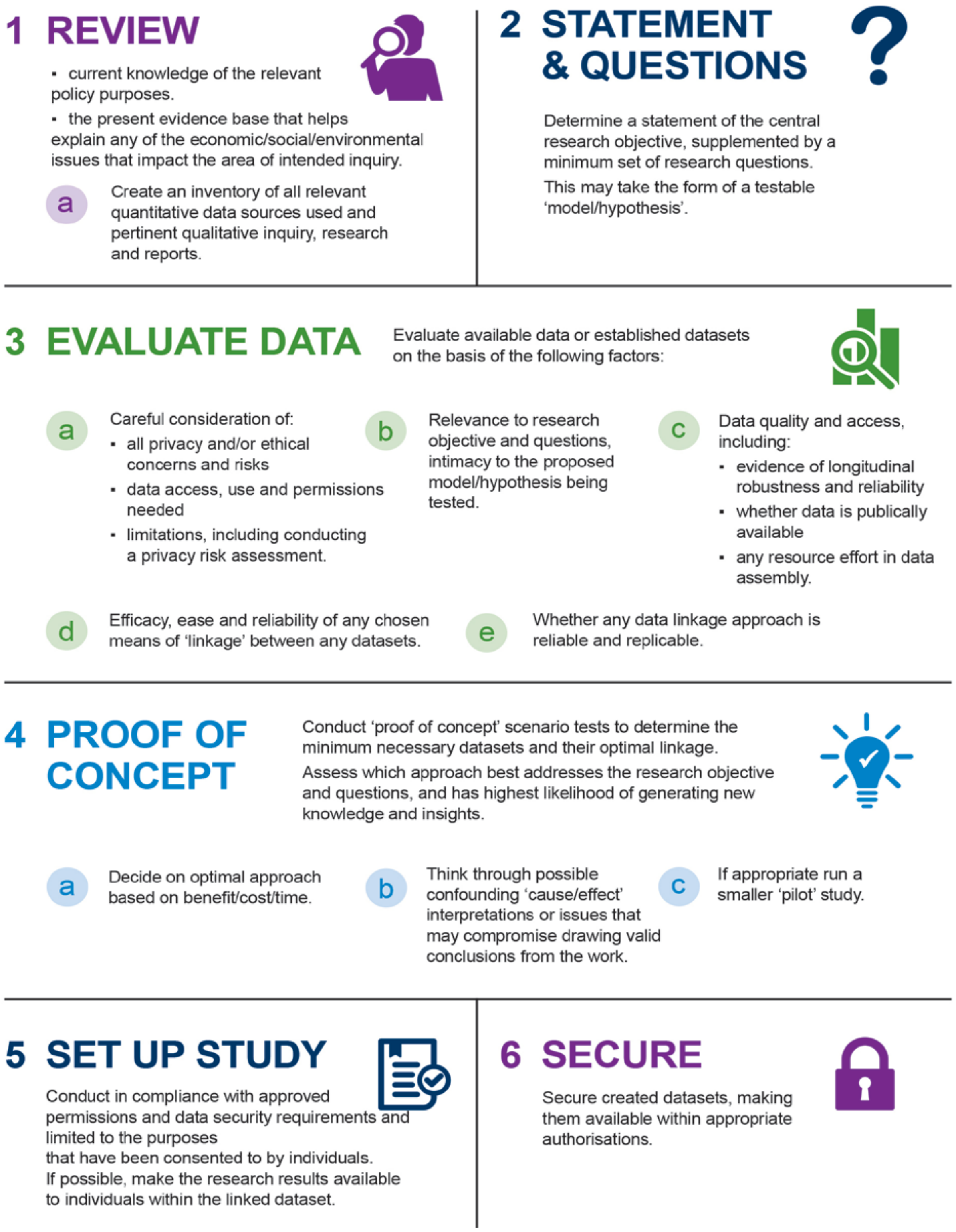
Given the complexity of undertaking research using data linkage, it is advisable to follow a 'best practice' process of resourcing, developing, undertaking and reporting such projects. Guidelines do exist, such as the 'five safes', a model developed initially in the United Kingdom by the Office for National Statistics and which is now widely used by data-integration organisations, including the ABS (Ritchie 2008). Note the five safes are applied to all ABS data, as well as being used for data linkage. This model prescribes that the project, the researchers, the setting, the data and the output should all be 'safe'; that is, considered for appropriateness, security and privacy. For example, for the basic confidentialised unit record files (CURFs) made available by the ABS, the ABS (2017) ensures:

- safe projects, by having users sign a declaration about how the data will be used
- safe researchers, by requiring users to register and sign this declaration
- safe settings, by requiring users to store the data securely
- safe data, through de-identification
- safe outputs, by giving users guidelines or rules about what may be published or shared.

The level of control can be adjusted according to the sensitivity of the data being provided to users for each 'safe' principle.

Figure 3 shows six principles for data-linkage research, developed by NCVER to guide data linkage projects. The focus here is on the steps researchers can take to produce the highest quality research — with the best chance of leading to practical policy outcomes — using data linkage, while mitigating risk. It is impossible to predict and cover every risk, although the accumulation of experience through repeated linking can identify a body of chief risks on which to focus. As such, these principles are meant as a guide to be thoughtfully applied by researchers.

Figure 3 NCVER's six principles for data linkage research



1 REVIEW

- current knowledge of the relevant policy purposes.
- the present evidence base that helps explain any of the economic/social/environmental issues that impact the area of intended inquiry.

a Create an inventory of all relevant quantitative data sources used and pertinent qualitative inquiry, research and reports.

2 STATEMENT & QUESTIONS

Determine a statement of the central research objective, supplemented by a minimum set of research questions.

This may take the form of a testable 'model/hypothesis'.

3 EVALUATE DATA

Evaluate available data or established datasets on the basis of the following factors:

a Careful consideration of:
- all privacy and/or ethical concerns and risks
- data access, use and permissions needed
- limitations, including conducting a privacy risk assessment.

b Relevance to research objective and questions, intimacy to the proposed model/hypothesis being tested.

c Data quality and access, including:
- evidence of longitudinal robustness and reliability
- whether data is publically available
- any resource effort in data assembly.

d Efficacy, ease and reliability of any chosen means of 'linkage' between any datasets.

e Whether any data linkage approach is reliable and replicable.

4 PROOF OF CONCEPT

Conduct 'proof of concept' scenario tests to determine the minimum necessary datasets and their optimal linkage.

Assess which approach best addresses the research objective and questions, and has highest likelihood of generating new knowledge and insights.

a Decide on optimal approach based on benefit/cost/time.

b Think through possible confounding 'cause/effect' interpretations or issues that may compromise drawing valid conclusions from the work.

c If appropriate run a smaller 'pilot' study.

5 SET UP STUDY

Conduct in compliance with approved permissions and data security requirements and limited to the purposes that have been consented to by individuals.

If possible, make the research results available to individuals within the linked dataset.

6 SECURE

Secure created datasets, making them available within appropriate authorisations.

Data linkage and VET in the future

Data linkage would benefit from a national register of administrative data and statistical assets to better understand the possibilities for future linkage projects.

The possibilities for the scope and use of data linkage in the future are growing, for the VET sector specifically and research generally. Data linkage has the potential to not only streamline the collection of information by reducing the need to 're-collect' certain information, but it also brings further opportunities to extract meaning from existing data. Data linkage is a powerful way of creating information on individuals' 'pathways' through life, education and employment. It can also create ways to measure the benefits or return on investment from programs without having to resort to large-scale longitudinal follow-up projects. A number of data archives or repositories aimed at making data more available to researchers and assisting with data-based research now exist, such as the Australian Data Archive, the National Australian Data Service and Data.gov.au. Although these are useful, they are limited in the data they contain or reference, and data linkage would benefit from a national register of administrative data and statistical assets, in that it would enable a more coherent assessment of the possibilities for future linkage projects. An example of one such national register that is currently under development arose from reporting of developmental consultations during 2015 for an Essential Statistical Assets list, under an Essential Statistical Infrastructure consultation exercise. Note that this is not a current or complete register but reflects the engaged stakeholders from a past exercise. A formal national register would need to reflect a baseline initiative with a full coverage compulsory stakeholder audit, and be maintained in real time. Furthermore, beyond a register of this type, a register of data linkage projects held (or a category within national register) would be useful. The National Statistical Service initiated such a mechanism via its voluntary 'Public register of data integration projects'[6]. Again, this is useful but to be a comprehensive resource, this would need a fuller baseline initiative and mandatory real time listing.

Data linkage is the current focus of numerous government initiatives and legislation, and linked datasets are in use for an ever-widening range of purposes. It is widely recognised that the use of linked datasets provides many opportunities for improving policy and outcomes, a recognition that has prompted initiatives such as the Data Integration Partnership for Australia (DIPA). DIPA is an upcoming Australia-wide organisation designed to govern data integration across the Australian Public Service. Established by the 2017—18 Commonwealth Budget, it will be administered by the ABS and will promote data linkage, with the aim of increasing the use and value of governmental data assets. It will create and administer de-identified, confidentialised datasets from the existing data assets held by government agencies and will include information on all areas of Australian life, including health, education, finance and the environment. It will also aim to improve data-linkage methodologies and infrastructure, while supporting existing linked datasets such as the Business Longitudinal Analysis Data Environment. This proposed organisation highlights the central importance of data linkage in government-supported research in the future, and if successful it could enable access to an extensive range of linked datasets for research in a variety of areas.

[6] See <http://www.nss.gov.au/nss/home.NSF/pages/Data%20Integration%20Find%20A%20Project>

One program already in use is the Multi-Agency Data Integration Project (MADIP)[7]. This is a partnership that combines a range of existing governmental data such as Medicare claims, government payments and the 2011 ABS Census and is the product of work by six different participating governmental agencies. MADIP is one example of how linked information has the potential to provide insight and understanding of policies and programs, and enable better targeting of these services to those who need them.

A recent report from the Productivity Commission on the status of the national education evidence base specifically mentioned the importance of data linkage to future education research. The report found that data linkage could be improved in this area through the implementation of a ‘national education master linkage key’, also known as a ‘unique student identifier’ (USI; Productivity Commission 2016, p.200). USIs are already in use for student records in the VET system with data availability for VET data records on training activity from 2015. The original aim, as proposed by COAG (2009), was the possible integration of a national student identifier, such as the USI, into all forms of education — from early schooling through to higher education and VET. This approach when implemented would create many valuable opportunities for data linkage between educational records and databases across jurisdictions. However, restrictions apply to these types of useful linkage keys; for example, under the relevant Student Identifiers Act, the Registrar of the USI office *may* authorise disclosure for research related to education or training that meets the requirements of the relevant Ministerial Council, currently the COAG Industry Skills Council (Australia Parliament 2014). The process of this authorisation has not yet been formalised.

[7] <http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Statistical+Data+Integration+-+MADIP>

Conclusion

Linkage projects could collect data potentially across the whole of a person's lifespan, and inform decisions on policy at every major point in that time.

This review and the case study evidence presented confirm that data linkage has the potential to be a powerful tool for research and policy in the VET sector. However, like any tool, data linkage must be used responsibly and for its intended purpose. With a carefully considered approach, data linkage can produce quality research with the capacity to improve and inform public policy.

Future projects in the VET sector might make use of data linkage to analyse the benefits of various programs and make determinations on return on investment, using sources such as the National VET Provider Collection alongside governmental administrative collections such as that of the Australian Tax Office, or data from the ABS Census. This would enable the tracking of training outcomes over significant timeframes, negating the need for a new longitudinal study, with its associated costs. Longitudinal studies would then be freer to focus on areas that administrative data cannot address, such as insights into the behaviours, thoughts and feelings of individuals. The six principles for data linkage (figure 3) can be applied to assess such projects and to guide the successful conduct of such research.

A historical perspective is also possible with data linkage; for example, projects linking a student's VET data to information from earlier in their life, such as school or NAPLAN results. This approach offers the opportunity to investigate factors that may predict a student's experience in the VET sector and therefore lead to policy aimed at supporting those at risk and social equity groups.

More cooperation and stronger relationships between key data custodians in the VET and related education sectors are central to building data-linkage capacity. Such approaches may include the use of existing linkage keys or adding the development of linking keys in each dataset to facilitate linkage. Data linkages between datasets may be possible to automate using machine learning, enabling a dataset that is continuously updated and available for use. The issues associated with the use of name and geographical location to match individuals might better be addressed by a number of interested data custodians.

As well as specific projects, it is important for the VET sector to invest in the capacity and skills needed to perform data linkage. Each instance of data linking requires clear planning, appropriate authorities and funding resources to ensure that maximum benefits are realised. Every successfully performed data-linkage project deepens the understanding of how to obtain the best and most practically useful information from these exercises.

The broader lifespan perspective offered by data linkage offers significant value to the VET sector: linkage projects could use data from across potentially the whole of a person's lifespan and inform decisions on policy at every major point in that time. The ability to look backwards and forwards from the point at which someone is involved in VET is a valuable opportunity for all researchers and policymakers.

References

ABS (Australian Bureau of Statistics) 2013, *Education and training newsletter, October 2013,* cat.no.4211.0, ABS, Canberra, viewed September 2017, <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4211.0main+features60October+2013>.

——2016, *Essential statistical assets for Australia, 2015,* cat.no.1395.0, ABS, Canberra, viewed January 2018, <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1395.0>.

——2017, *ABS confidentiality series, August 2017,* cat.no.1160.0, ABS, Canberra, viewed September 2017, <http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1160.0Main%20Features4Aug%202017?opendocument&tabname=Summary&prodno=1160.0&issue=Aug%202017&num=&view=>.

Australian Parliament 2014, *Student Identifiers Act 2014*, Parliament of the Commonwealth of Australia, Canberra, viewed October 2017, <https://www.legislation.gov.au/Details/C2014A00036>.

——2016, *Privacy Amendments (Notifiable Data Breaches) Bill 2016 explanatory memorandum*, Parliament of the Commonwealth of Australia, Canberra, viewed September 2017, <http://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r5747_ems_ed12b5bb-d3b3-4a6a-9536-53bb459a00df/upload_pdf/6000003.pdf;fileType=application%2Fpdf>.

Department of Human Services 2017, *Centrelink program data matching activities*, Department of Human Services, Canberra, viewed August 2017, <https://www.humanservices.gov.au/corporate/publications-and-resources/centrelink-program-data-matching-activities>.

Department of Industry 2015, *Australian innovation system report*, Department of Industry, Canberra, viewed September 2017, <https://industry.gov.au/Office-of-the-Chief-Economist/Publications/Documents/Australian-Innovation-System/Australian-Innovation-System-Report-2015.pdf>.

Department of Social Services 2016, *Review of Australia's longitudinal data system*, Department of Social Services, Canberra, viewed September 2017, <https://s3-ap-southeast-2.amazonaws.com/ehq-production-australia/5af646ff85f4cb0d1f20ec02db2a35e88633507a/documents/attachments/000/059/186/original/Review__Final_Report_-_NCLD.pdf?1499924852>.

Dusetzina S, Tyree S, Meyer A, Green, L, & Carpenter, W 2014, *Linking data for health services research: a framework and instructional guide*, Agency for Healthcare Research and Quality (US), Rockville, Maryland, viewed Janurary 2018, <https://www.ncbi.nlm.nih.gov/books/NBK253312/>

Lumsden, M, Semo, R, Blomberg, D & Lim, P 2015, *Linking NAPLAN scores to the Longitudinal Surveys of Australian Youth,* NCVER, Adelaide, viewed August 2017, <http://www.lsay.edu.au/publications/2829.html>.

Misko, J, Korbel, P & Blomberg, D 2017, *VET in Schools students: characteristics and post-school employment and training experiences,* NCVER, Adelaide, viewed November 2017, <https://www.ncver.edu.au/publications/publications/all-publications/vet-in-schools-students-characteristics-and-post-school-employment-and-training-experiences>.

National Statistical Service 2010, *High level principles for data integration involving Commonwealth data for statistical and research purposes*, National Statistical Service, Canberra, viewed August 2017, <http://www.nss.gov.au/nss/home.NSF/pages/High+Level+Principles+for+Data+Integration+-+Content?OpenDocument>.

——2013, *Data integration involving Commonwealth data for statistical and research purposes: risk assessment guidelines*, National Statistical Service, Canberra, viewed October 2017, <http://www.nss.gov.au/nss/home.NSF/pages/Data+integration+projects+%E2%80%93+how+to+determine+the+risk+level?opendocument>.

——nda, *Data linking: what is data linking? Information sheet one,* National Statistical Services, Canberra, viewed September 2017, <http://www.nss.gov.au/nss/home.nsf/533222ebfd5ac03aca25711000044c9e/91242a5a14b12e26ca257ba8007b0819/$FILE/data%20linking%20w.pdf>.

——ndb, *Statistical data integration involving Commonwealth data*, National Statistical Service, Canberra, viewed August 2017, <http://www.nss.gov.au/nss/home.NSF/pages/Data+Integration:+FAQ's?OpenDocument#Anchor23>.

New South Wales Parliament 2015, *Explanatory note, Privacy and Personal Information Protection Amendment (Exemptions Consolidation) Bill 2015*, Parliament of New South Wales, Sydney, viewed September 2017, <https://www.parliament.nsw.gov.au/bills/DBAssets/bills/ExplanatoryNotes/3216/XN%20Privacy.pdf>.

Office of the Australian Information Commissioner 2014, *Guidelines on data matching in Australian Government administration*, Office of the Australian Information Commissioner, Canberra, viewed September 2017, <https://www.oaic.gov.au/agencies-and-organisations/advisory-guidelines/data-matching-guidelines-2014>.

——2017, *Australian Community Attitudes to Privacy Survey*, Office of the Australian Information Commissioner, Canberra, viewed August 2017, <https://www.oaic.gov.au/engage-with-us/community-attitudes/australian-community-attitudes-to-privacy-survey-2017>.

Pearce, A, Scalzi, D, Lynch, J & Smithers, L 2016, 'Do thin, overweight and obese children have poorer development than their healthy-weight peers at the start of school? Findings from a South Australian data linkage study', *Early Childhood Research Quarterly*, vol.35, pp.85—94.

Productivity Commission 2016, *National education evidence base: Productivity Commission inquiry report,* Productivity Commission, Canberra, viewed August 2017, <http://www.pc.gov.au/inquiries/completed/education-evidence/report>.

Ritchie, F 2008, 'Secure access to confidential microdata: four years of the Virtual Microdata Laboratory', *Economic & Labour Market Review*, vol.2, no.5, pp.29—34.

Sayers, A, Ben-Shlomo, Y, Blom, AW & Steele, F 2016, 'Probabilistic record linkage', *International Journal of Epidemiology*, vol.45, no.3, pp.954—64.

United Kingdom Department for Education 2014, *Teachers in England's secondary schools: evidence from TALIS 2013*, Department for Education, London, England, viewed August 2017, <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/322910/RR302_-_TALIS_report_NC.pdf>.

Victorian Parliament 2014, *Privacy and Data Protection Act 2014*, Parliament of Victoria, Melbourne, viewed October 2017, <http://www.austlii.edu.au/cgi-bin/viewdb/au/legis/vic/num_act/padpa201460o2014317/>.

Appendix – examples of relevant datasets and related initiatives

Australian Apprenticeships Management System (AAMS)	Australian Apprenticeships Management System is the upcoming national system for the administration of apprenticeships, including payment of government incentives to employers and apprentices, and storage of contact information and contracts of training.
ABS Census of Population and Housing (ABS Census)	<http://www.abs.gov.au/websitedbs/Censushome.nsf/home/2016> The ABS Census is a survey conducted at five-yearly intervals which aims to capture information on all people in Australia, including demographic, income, education and other factors.
Australian Bureau of Statistics General Social Survey (ABS GSS)	<http://www.abs.gov.au/ausstats/abs@.nsf/mf/4159.0> The ABS General Social Survey is a survey of individual and community wellbeing, with a focus on social capital. It collects data on a range of social and demographic factors such as housing, health, community involvement, financial stress and sexual orientation. The survey is currently collected at four-year intervals via household survey. The most recent survey was conducted in 2014 and yielded responses from 12 932 dwellings.
Australian Bureau of Statistics Labour Force Survey (ABS LFS)	<http://www.abs.gov.au/Employment-and-Unemployment> The ABS Labour Force Survey is an official survey of employment and unemployment, with data collected and released by the ABS on a monthly basis. The survey is conducted through face-to-face and phone interviews of households, in which approximately 70 questions are asked.
Australian Bureau of Statistics Multi-Purpose Household Survey (ABS MPHS)	<http://www.abs.gov.au/ausstats/abs@.nsf/0/6B49F5106520A6A7CA2572C100244ACB?opendocument> The ABS Multi-Purpose Household Survey is a national Australian survey of households on issues related to barriers and incentives to labour force participation, retirement and retirement intentions, and work-related injuries. These topics are rotated through every two to four years, ensuring ongoing data collection of each issue. The MPHS is conducted in conjunction with the monthly Labour Force Survey.
Australian Children's Education and Care Quality Authority (ACECQA)	<http://acecqa.gov.au/> ACECQA is responsible for overseeing the National Quality Framework, where it applies to children under the age of 13. This involves assembling and storing data on assessment and ratings of educational and care services for children aged under 13 years. ACECQA maintains three registers: Education and Care Services, listing the services granted approval to operate under the National Qualification Framework; Approved Providers, listing individuals or entities authorised to operate and approved education and care service; and Certified Supervisors, listing persons holding a

	supervisor certificate.
Australian Census Longitudinal Dataset (ACLD)	<http://www.abs.gov.au/websitedbs/Censushome.nsf/home/acld> The ACLD bring together data from the ABS Census at different time points to build a longitudinal dataset. The initial release used a 5% sample from the 2006 ABS Census with matching records from the 2011 ABS Census, with plans to add data from subsequence ABS Censuses and other administrative datasets.
Australian Longitudinal Study on Male Health — 'Ten to Men' (ALSMH)	<http://www.tentomen.org.au/> 'Ten to Men', an Australia-wide study of males aged 10 and up, follows participants over time, with the aim of improving health for Australian men and boys. Just under 16 000 participants, aged 10–55 years, were recruited initially. Participants complete paper questionnaires every two to three years on a range of health and wellbeing topics.
Australian Curriculum, Assessment and Reporting Authority (ACARA)	<https://www.acara.edu.au> ACARA is the Australian body for the development of national schooling curriculum, administration of national assessments and associated reporting.
Australian Early Development Census (AEDC)	<https://www.aedc.gov.au/> The AEDC is a nationwide data collection of early childhood development at the time children commence their first year of full-time school. The AEDC provides evidence to support health, education and community policy and planning. The AEDC is held every three years, with the 2015 AEDC data collection being the third collection and a collection planned in 2018. The Census involves teachers of children in their first year of full-time school completing a research tool, the Australian version of the Early Development Instrument. The AEDC covers five key domains: physical health and wellbeing; social competence; emotional maturity; language and cognitive skills; and communication skills and general knowledge.
Australian Government Department of Education and Training (DET)	<https://www.education.gov.au/> The Australian Government Department of Education and Training is responsible for national policies and programs that help Australians access quality and affordable early child care and childhood education, school education, higher education, vocational education and training, international education and research. The administrative records held by the Department represent a useful dataset for possible linkage.
Australian Longitudinal Study of Ageing (ALSA)	<http://www.flinders.edu.au/sabs/fcas/alsa/> The general purpose of the ALSA study is to gain further understanding of how social, biomedical and environmental factors are associated with age-related changes in the health and wellbeing of people aged 70 years and over. Emphasis is given in the overall study to defining and exploring the concept of healthy active ageing, particularly in a South Australian context. The ALSA commenced in 1992 with 2087 participants aged 65 years or more.

Australian Longitudinal Study on Women's Health (ALSWH)	<https://www.alswh.org.au/> ALSWH aims to provide data and assist policy-making on issues related to women's health. It assesses women's physical and mental health, as well as psycho-social aspects of health (such as socio-demographic and lifestyle factors) and their use of health services. It consisted of over 58 000 women in three cohorts who were aged 18–23, 45–50 and 70–75 years when surveys began in 1996. In 2012–13 more than 17 000 young women aged 18–23 years were recruited to form a new cohort.
Business Longitudinal Analysis Data Environment (BLADE)	<https://industry.gov.au/Office-of-the-Chief-Economist/Data/Pages/Business-Longitudinal-Analysis-Data-Environment.aspx> BLADE is not a dataset; rather, it is a methodology for linking businesses using the Australian Business Number (ABN). In this way it allows the integration of administrative and survey data in order to assess business performance and dynamics, business demography and characteristics.
Longitudinal Study of Humanitarian Migrants — 'Building a New Life in Australia' (BNLA)	<http://www3.aifs.gov.au/bnla/> BNLA is a long-term study of humanitarian migrants in Australia. Approximately 2400 individuals and families are part of the study, from which data are collected via home visits and telephone interviews. Participants are asked questions on a range of topics, including pre-migration experiences, employment and income and life satisfaction.
Higher Education Information Management System (HEIMS)	<http://heimshelp.education.gov.au> HEIMS is a system that collects and makes relevant information on students in higher education in Australia available to higher education providers, such as usage of Commonwealth assistance.
Household, Income and Labour Dynamics in Australia Survey (HILDA)	<http://melbourneinstitute.unimelb.edu.au/hilda> HILDA is a household-based panel survey that collects information about economic and personal wellbeing, labour market dynamics and family life. It aims to provide policy-makers with unique insights about Australia, enabling them to make informed decisions across a range of policy areas, including health, education and social services. HILDA currently follows over 17 000 Australians, and participants are followed over the course of their lifetime.
International Computer and Information Literacy Study (ICILS)	<https://icils.acer.org/> ICILS is an international assessment of computer and information literacy, conducted in Year 8 or its national equivalent. It includes student, teacher and school-level questionnaires to develop information on the context in which this literacy is developed. In 2013 close to 60 000 students were involved in ICILS, across over 3300 schools and 21 different educational systems.

Internet Vacancy Index	<https://data.gov.au/dataset/internet-vacancy-index> This index is based on a count of online job advertisements newly lodged on three main job boards (SEEK, CareerOne and Australian JobSearch) during the month. The output includes a time series of vacancies at various levels, including at national, state and regional levels and at an occupational level.
Longitudinal Dataset for the Investment Approach (JASON)	<http://www.aihw.gov.au/data/priority-investment-approach-dataset/> JASON is a longitudinal quarterly dataset managed by the Australian Department of Social Service and contains data on individuals receiving government payments such as pensions, ABSTUDY, family tax benefit, youth allowance and concession cards.
Longitudinal Survey of Australian Children — 'Growing up in Australia' (LSAC)	<http://growingupinaustralia.gov.au/> LSAC follows the development of 10 000 children and families from all parts of Australia. The study commenced in 2004 with two cohorts: families with children aged four to five years and those with children aged from birth to one year. It is investigating the contribution of children's social, economic and cultural environments to their adjustment and wellbeing. A major aim is to identify policy opportunities for improving support for children and their families and for early intervention and prevention strategies.
Longitudinal Surveys of Australian Youth (LSAY)	<http://www.lsay.edu.au/index.html> LSAY tracks young people as they move from school into further study, work and other destinations, providing a rich source of information to help better understand young people and their transitions from school to post-school destinations, as well as exploring social outcomes, such as wellbeing. Information collected as part of LSAY covers a wide range of school and post-school topics, including: student achievement, student aspirations, school retention, social background, attitudes to school, work experiences and what students are doing when they leave school. This includes vocational and higher education, employment, job-seeking activity, and satisfaction with various aspects of their lives.
Longitudinal Study of Indigenous Children — 'Footprints in Time' (LSIC)	<https://www.dss.gov.au/about-the-department/publications-articles/research-publications/longitudinal-data-initiatives/footprints-in-time-the-longitudinal-study-of-indigenous-children-lsic> LSIC covers a wide variety of topics on children's health, learning and development, their family and community. Children and their parents are interviewed yearly, with approximately 1680 families participating in the initial wave. The study includes two groups of Aboriginal and/or Torres Strait Islander children who were aged six to 18 months (B cohort) and three-and-a-half to five years (K cohort) when the study began in 2008.

Multi-Agency Data Integration Project (MADIP)	<http://www.abs.gov.au/websitedbs/d3310114.nsf/home/statistical+data+integration+-+madip> Currently in the evaluation stage, the MADIP brings important national datasets together to look at patterns and trends to help agencies and analysts address complex policy and service delivery questions facing Australia. The MADIP combines existing data on Medicare benefit claims, government payments, and income tax with the 2011 Census to create a linked dataset that provides a high-quality snapshot of Australia in 2011.
My School	<https://www.myschool.edu.au/> My School collects information such as performance data, financial information, population data and NAPLAN results on Australian schools. It also holds enrolment numbers and attendance rates. My School is managed by the Australian Curriculum, Assessment and Reporting Authority (ACARA).
National Assessment Program (NAP)	<https://www.nap.edu.au/> The NAP is aimed at measuring the educational outcomes of young Australians. It is run by the Education Council and, as well as administering the NAPLAN, it conducts occasional sample assessments in areas such as Information and Communication Technology (ICT) literacy. A small sample of randomly selected schools participates in NAP sample assessments, which occur annually on a rolling basis. In 2013, civics and citizenship was tested and, in 2014, ICT literacy was tested. Science literacy was tested in 2015. NAP encompasses three assessments: a. the National Assessment Program - Literacy and Numeracy (NAPLAN) b. three-yearly sample assessments in science literacy, civics and citizenship, and information and communication technology (ICT) literacy c. participation in international sample assessments.
National Assessment Program — Literacy and Numeracy (NAPLAN)	<www.nap.edu.au/naplan> NAPLAN is an annual assessment for all students in Years 3, 5, 7 and 9. It tests the types of skills that are essential for every child to progress through school and life. The tests cover skills in reading, writing, spelling, grammar and punctuation, and numeracy.
National Apprentice and Trainee Collection	<https://www.ncver.edu.au/data/collection/apprentices-and-trainees-collection> The National Apprentice and Trainee Collection holds information such as commencements, training rate and duration of training for all apprentices and trainees employed under training contracts. Data are collected via submission from state training authorities, through data contained on the apprenticeship/traineeship contract.

National Disability Insurance Scheme (NDIS)	<https://www.ndis.gov.au> The NDIS provides support to people with a disability, assisting them to access services and supports and funding these where appropriate. It stores records on all current and former clients, which, if de-identified, could be used for research.
National School Statistics Collection (NSSC)	<http://www.abs.gov.au/ausstats/abs@.nsf/products/6F7111FCBD0121C0CA256BD00027255B?OpenDocument> The NSSC is an annual Census conducted in cooperation with state, territory and Commonwealth education authorities and the ABS. It collects data on issues relating to schools, students and staff in primary and secondary schools in Australia.
National Student Outcomes Survey	<https://www.ncver.edu.au/data/collection/student-outcomes> The National Student Outcomes Survey collects information on VET students' reasons for training, their employment outcomes, satisfaction with training, and further study outcomes. Students included in the survey are those who completed their training in the previous calendar year and have an Australian address as their usual address. Since 1999, the survey has collected information on the outcomes of government-funded VET students, broadly defined as all activity delivered by government providers and government-funded activity delivered by community education and private training providers. In 2016, the scope of the survey was expanded to report on the outcomes of all graduates — referred to as total VET graduates — that is, those graduates whose training was Commonwealth or state-funded, as well as fee-for-service graduates (those who paid for the training or whose employer paid for the training).
National VET Finance Collection	<https://www.ncver.edu.au/data/collection/vet-finance> The National VET Finance Collection holds data such as revenues, expenditures, VET student loans, assets and liabilities in Australia's public VET system.
National VET in Schools Collection	<https://www.ncver.edu.au/data/collection/vet-in-schools> This National VET in Schools Collection provides data for vocational education and training undertaken by school students as part of their senior secondary certificate of education (SSCE) where the training is nationally recognised or delivered by schools or other training providers. Information on participation, students, courses and qualifications, and subjects relating to VET in Schools students of all ages is included.
National VET Provider Collection	<https://www.ncver.edu.au/data/data/total-vet-activity> The National VET Provider Collection data provide an estimate of the extent and nature of the vocational education and training delivered by Australian training providers. This picture of training activity has included since 2014 information from all types of providers and not only those in receipt of Commonwealth or state funding. Information is provided on the number of training providers, students, enrolments in programs, enrolments in subjects, hours of delivery and program completions.

Progress in International Reading Literacy Study (PIRLS)	<https://timssandpirls.bc.edu/> PIRLS is an international assessment of reading literacy, administered in Year 4 in participating countries every five years. PIRLS covered 50 countries in the 2016 sample, with 14 covered by the sister project, ePIRLS, which assesses literacy in an online context.
Programme for the International Assessment of Adult Competencies (PIAAC)	<https://www.oecd.org/skills/piaac/> PIAAC is a body for the administration of the Survey of Adult Skills, an international survey of information processing skills in those aged over 16 years. The survey is conducted in over 40 countries through face-to-face, computerised and pen-and-paper interviews and questionnaires.
Programme for International Student Assessment (PISA)	<https://www.oecd.org/pisa/> PISA is an international survey-based evaluation of education systems, conducted through standardised testing of 15-year-old students of participating countries.
Quality Indicators for Learning and Teaching (QILT)	<https://www.qilt.edu.au> QILT collects data from Australian higher education teaching institution surveys of students, and is maintained by the Australian Government. The data are available to students through the QILT website and to researchers upon request to the Social Research Centre.
SA–NT Data Link	<https://www.santdatalink.org.au/> SA–NT DataLink supports population-based data-linkage research in South Australia and the Northern Territory by providing access to information from government agencies and other relevant participating organisations in the form of linked administrative and clinical datasets. Relevant to the case studies outlined, the SA–NT Data Link includes: *South Australian Perinatal Statistics Collection* <http://www.sahealth.sa.gov.au/wps/wcm/connect/public+content/sa+health+internet/about+us/health+statistics/pregnancy+outcome+statistics> This collection of data relates to the health and wellbeing of mother and child, and is mandatory in South Australia via the Supplemental Birth Record, collected by the Perinatal Outcomes Unit of the South Australian Department of Health. It is made available to researchers through the SA–NT DataLink. *South Australian School Enrolment Census* This Census collects data on all students enrolled in government schools and participating in education programs in South Australia.

Survey of Employer Use and Views of the VET System	<https://www.ncver.edu.au/data/collection/employers-use-and-views-of-the-vet-system> This survey collects information about employers' use and views of the vocational education and training system and the various ways in which employers use the VET system to meet their skill needs. The information collected is designed to measure the awareness, engagement and satisfaction of employers with the VET system.
Teaching and Learning International Survey (TALIS)	<https://www.oecd.org/edu/school/talis.htm> TALIS is an international survey of teachers and school leaders and collects information in the areas of: learning environment; appraisal and feedback; teaching practices and classroom environment; development and support; school leadership; self-efficacy; and job satisfaction. In 2013, 34 countries and over five million teachers were surveyed.
Training and Youth Internet Management System (TYIMS)	<https://tyims.education.gov.au> TYIMS was previously the Australian Government's system for the administration of Australian Apprenticeships, including the processing and payment of incentive claims and storage of contracts of employment. It is in the process of being replaced by the Australian Apprentice Management System. (See Australian Apprentice Management System.)
Trends in International Mathematics and Science Study (TIMSS)	<https://timssandpirls.bc.edu> TIMSS is an international assessment of maths and science skills, conducted amongst Year 4 and Year 8 students every four years. In 2015, 57 countries participated, with more than 580 000 students assessed.

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>