Australian Bureau of Statistics

1351.0.55.044

Research Paper

# Assessing the Feasibility of Linking 2011 Vocational Education and Training in Schools Data to 2011 Census Data

New Issue

**Research Paper**

# Assessing the Feasibility of Linking 2011 Vocational Education and Training in Schools Data to 2011 Census Data

National Centre for Education and Training

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 9 DEC 2013

ABS Catalogue no. 1351.0.55.044



Produced by the Australian Bureau of Statistics

## INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact Mr Myles Burleigh, National Centre for Education and Training, on Canberra (02) 6252 6534 or email <education.statistics@abs.gov.au>.

# ASSESSING THE FEASIBILITY OF LINKING 2011 VOCATIONAL EDUCATION AND TRAINING IN SCHOOLS DATA TO 2011 CENSUS DATA

National Centre for Education and Training

## EXECUTIVE SUMMARY

As part of the Australian Bureau of Statistics' (ABS) investigations into integrating data, this study was conducted to assess the feasibility of linking 2011 Vocational Education and Training in Schools data to ABS 2011 Census of Population and Housing data. This project was undertaken by the ABS and funded by the Strategic Cross-sectoral Data Committee as part of the Transforming Education and Training Information in Australia (TETIA) initiative. The primary aim of the study was to establish the quality of a linked dataset where neither name and address nor a statistical linkage key was available for linkage; this is referred to as Bronze linkage.

Vocational Education and Training (VET) data at the enrolment unit record level were provided to the ABS by the National Centre for Vocational Education Research. The dataset used for this feasibility study covered the 2011 calendar year.

Preparation of the datasets for linkage involved a standardisation process to make the two datasets more compatible for linkage and comparison. A VET in Schools person level dataset was extracted from the VET datasets, which were supplied at the enrolment level and included non VET in Schools enrolments. Statistical geography codes were derived onto the VET data, by using the postcode and locality information provided, to match codes that were already present on the Census data set.

A deterministic method of linkage, also known as exact or rule-based linkage, was used to integrate the VET in Schools and Census datasets. This is where individual records from two datasets are compared on the basis of common linkage variables. Records which agree exactly or almost exactly, such as age within one year, on a subset of linkage variables are identified as matches. Reiterative linkage passes can be used to improve linkage rates using different linkage variables, and ranking and/or clerical review may be used to resolve instances where multiple match candidates arise. This report examines the data, methodology, and results of the linkage. Further exploration is recommended into the statistical and research outputs that would be supported by a successfully linked dataset.

The Bronze deterministic linkage method used in this study searched for records that linked across the VET in Schools and Census data on combinations of matching or near matching sex, date of birth, age, and various statistical geography codes. This process resulted in 49.4% of the VET in Schools person records being linked to an equivalent Census person record.

The quality of the linkage was assessed on the basis of four indicators:

1. expected links – the number of links expected to be made after taking the net undercount in the Census into account
2. missing data – a short analysis of the types of missing data on the linked and unlinked records
3. linking variables – an analysis of the accuracy of the linking variables and their usefulness for linkage
4. estimated link accuracy – a comparison of estimated link accuracy measures for different linking variables.

The results from this study indicate that linking VET in Schools data to the Census using a Bronze deterministic linkage method did not produce a dataset of acceptable quality for analysis or reporting. However, it did provide insights into the kind of data that would be required to make an acceptable link. In particular, the results showed that small population area geographic codes, such as Mesh Block, identified a much higher proportion of unique links than large area codes, such as Statistical Area 2.

With additional and improved linking variables, future integrated datasets could support a range of investigations, including analysis of pathways taken by people entering and leaving the VET system as part of their schooling, differences in reporting for key variables such as Indigenous status across the two collections, and the socio-economic and demographic characteristics of the participants of VET in Schools and their families.

Further linkage using more detailed geographic identifiers is recommended. This would require VET in Schools data with small area geographic identifiers, such as Mesh Block and Statistical Area 1, or address data from which these codes could be obtained. Alternatively, further work could be undertaken to determine ways of improving the accuracy and coverage of linkage from the existing data. However, as the existing VET in Schools data has only large population area information and the results of this study show that such data is not as effective for linkage as more detailed geographic information, this avenue is not as likely to be as successful in improving the quality of the linked dataset significantly.

# CONTENTS

# ABBREVIATIONS

| | |
|---|---|
| ABS | Australian Bureau of Statistics |
| ERP | Estimated Resident Population |
| NCVER | National Centre for Vocational Education Research |
| SA1 | Statistical Area 1 |
| SA2 | Statistical Area 2 |
| SCDC | Strategic Cross-sectoral Data Committee |
| SLA | Statistical Local Area |
| TETIA | Transforming Education and Training Information in Australia |
| VET | Vocational Education and Training |

# ASSESSING THE FEASIBILITY OF LINKING 2011 VOCATIONAL EDUCATION AND TRAINING IN SCHOOLS DATA TO 2011 CENSUS DATA

National Centre for Education and Training

## ABSTRACT

This study was conducted to assess the feasibility of linking 2011 Vocational Education and Training in Schools data to ABS 2011 Census of Population and Housing data, without the use of name and address or a statistical linkage key as linking variables. This initial paper details the methodology used in the linkage process, the outcomes of the project, the quality of the resultant linked dataset, and recommendations for future linkage projects.

The results from this quality study indicate that linking Vocational Education and Training in Schools data to the Census using a deterministic linkage method did not produce a dataset of acceptable quality for analysis using the available data. Improved data, particularly the availability of quality small area geographic codes, would likely result in a linked dataset that could be used for analysis and reporting. Further work could also be undertaken to improve linkage using the existing data. However, without improvements to the data, a change in quality sufficient to enable analysis and reporting is less likely to be achieved.

A link to the Census would add significant value to the information available for Vocational Education and Training. With this in mind, access to improved data and new linkage methodologies should be explored to link the data in the future.

# 1. INTRODUCTION

The ABS has been exploring ways to enhance the value of information assets held by government agencies and other organisations by integrating these data sources particularly with data from the Census of Population and Housing. Since 2006, the ABS has investigated various statistical linkage techniques to assess the linkage quality required to allow analysis and reporting to be performed on the resulting linked datasets. Opportunities to expand the information available from education data have been extensively explored through the 2011 suite of Census Data Enhancement Education Quality Study projects.

Integrating data provides the potential to gain more information from the combination of existing datasets than would be possible from those datasets alone; without the cost of enumeration or burden on providers associated with new collections (ABS, 2013b).

The ABS undertakes the integration of data strictly for statistical and research purposes only. These purposes include the description of characteristics of groups within a given population, and relationships that might exist between variables such as social, economic and environmental conditions, behaviours and outcomes. The ABS does not use integrated data for regulatory or compliance purposes. As an accredited Integrating Authority, the ABS adheres to the High Level Principles for data integration involving Commonwealth data (CPSIC, 2010).

The ABS, in consultation with the National Centre for Vocational Education Research, has continued investigations into enhancing education and training data through linking 2011 Vocational Education and Training (VET) in Schools data to the 2011 Census. This study investigated the feasibility of linking VET in Schools and Census data through a deterministic methodology, using de-identified data. If such a link is feasible, analysis of the integrated data may be explored in future studies.

Deterministic linking searches for records that match exactly or closely on common variables across two data sources. Data linkage without the use of identifying information, such as name and address, and without the use of a statistical linkage key is referred to as Bronze linkage. Previous studies have shown that the benefits of probabilistic linkage techniques[1] are less applicable for Bronze linkage; deterministic linkage is also less resource intensive than probabilistic methods and was seen to be the appropriate avenue for this study (ABS, 2013d).

1 Probabilistic linking compares records from two datasets using several variables common to both datasets and generates a single numerical measure of how well two particular records match. This allows ranking of all possible record pairs and assignment of the optimal link.

# 2. THE DATA

This section provides an overview of the two data sources brought together for linkage; that is 2011 Vocational Education and Training (VET) in Schools data and 2011 Census data. The data quality issues which impact on the compatibility of the two sources for linkage purposes are then discussed.

## 2.1 Vocational Education and Training data

For this study, VET data at the enrolment unit record level for all jurisdictions in Australia were provided to the ABS by the National Centre for Vocational Education Research (NCVER). While 2009, 2010, and 2011 data were submitted to the ABS; only 2011 data were used for this particular study.

The data received from the NCVER came from the VET in Schools collection and the Students and Courses collection. The datasets contained enrolment-level data with information about persons, training organisations, qualifications and enrolments. This data was processed and standardised to produce and prepare a person-level VET in Schools dataset for linkage.

Table 2.1 summarises the number of records received.

**2.1 Counts of person and enrolment records, by collection for 2011**

| | Enrolments | Persons |
|---|---|---|
| Collection | | |
| Students and Courses | 4,273,616 | 364,420 |
| VET in Schools | 3,297,417 | 236,461 |
| Total records | 7,571,033 | 600,881 |

Source: Vocational Education and Training enrolment records, 2011

## 2.2 Census data

The 2011 Census dataset used for this study consisted of 20,928,304 records, excluding imputed records. Imputed records are created to account for people for whom no Census form was returned – see the *Census Dictionary* (ABS, 2011b) for more information about imputation. The data was a person-level extract developed for the purpose of ABS Census Data Enhancements projects.

## 2.3 Preparing data for linkage

Preparing data for linkage involves several steps, taken to increase the compatibility of the original datasets and to identify and address any data quality issues. For this project, these steps included:

- ensuring that the population of interest is included in both datasets (Sections 2.3.1 and 2.3.2)
- ensuring the collection time frame of both datasets is comparable (Section 2.3.3)
- identifying variables suitable for linkage and ensuring these variables are collected and coded in a compatible way, or recoding variables where necessary to maximise the possibility for linkage to occur (Section 2.3.4)
- ensuring there is only one record per person on each dataset (Section 2.3.5).

Each step in the preparation process is explained in this section.

### *2.3.1 Scope of the Vocational Education and Training population*

The VET data received from the NCVER included persons aged 15–20 years who were undertaking a qualification, module, or unit of competency in 2011. Only VET enrolments delivered by registered training organisations, as recognised by the NCVER, were included in the data.

There were 236,461 people, approximately 39.4%, in the 2011 VET data who were undertaking their enrolments as part of a VET in Schools program. The remaining 364,420, approximately 60.6%, were undertaking their enrolments via other means, such as apprenticeships.

Only 2011 VET in Schools data was considered in scope for this study. Consequently, the relevant subset of the person-level data was extracted and used for linking.

### *2.3.2 Scope of the 2011 Census*

The scope of the Census data was in excess of the VET in Schools data, since the majority of persons represented in the Census data were not in the VET in Schools system at the time of collection. Consequently, a subset of the input Census file was derived for linking.

As age was of high quality on both files it was used to limit the scope of Census data. Age on Census night, 9 August 2011, was derived from the VET date of birth data item which confirmed only persons aged 15–20 years were included on the 2011 data received from the NCVER. Census data was limited to one year of age either side of this age range (14–21 years). This created a person-level dataset of approximately 2.2 million records, 10.5% of the input Census dataset, to be used for linking to the 2011 VET in Schools data.

### 2.3.3 Time frame comparability

The ability to link records from two datasets is maximised when the data is collected in the same time frame, however, this is rarely possible. Most data linkage occurs between sources that are collected in different time frames and a greater time difference reduces the likelihood of a successful linkage (ABS, 2013c). While the two datasets linked in this feasibility study were both extracted from data collected in 2011, there are differences in the time frames for when the data was collected.

The VET dataset is sourced from administrative data. For the VET in Schools collection, data are collected for enrolments for each unit of competency or module for a student during the calendar year of the collection. Data are collected via the senior secondary assessment authority, sometimes known as Boards of Studies, in each state or territory and reported through state training authorities or directly by the assessment authorities to the NCVER. Standardised data files are submitted to the NCVER by 31 March each year of the following year, for example the 2011 calendar year VET data was submitted to the NCVER by 31 March 2012.

While the VET in Schools collection comes from administrative data that are collected over a calendar year, the Census is sourced from persons, through the personal form, or their representatives, through the household, interviewer, or summary forms, and is a 'snapshot' – with the vast majority of Census forms being completed on Census night – 9 August 2011.

### 2.3.4 Selecting and standardising variables for linkage

The VET in Schools and Census datasets have few variables in common to facilitate linkage. Both data sources were de-identified, meaning that name and address were not available. However, less detailed data, such as date of birth and sex, were shared across the datasets. Additionally, both data sources use Australian standards for classifying Country of birth[2] and Main language spoken at home.[3] While high quality small area geography codes, such as Mesh Block, were available on the Census, the VET data was received from the NCVER with large area geography including suburb, postcode, and Statistical Local Area (SLA) derived by the NCVER from postcode. The ABS was able to use postcode to derive Statistical Area 2 (SA2) onto 86.7% of VET in Schools person records. Since only locality and postcode was available, small area geographic identifiers were only able to be derived for a very small proportion of records on the VET in Schools data: Statistical Area 1 (SA1) for 3.9% and Mesh Block for 0.7%.

---

2 *Standard Australian Classification of Countries* (ABS, 2011d).

3 *Australian Standard Classification of Languages* (ABS, 2011a).

While the geographic variables were derived in different ways, the codes themselves were directly comparable. However, some variables required adjustment to optimise the likelihood of linkage. For instance, all the date variables on the VET in Schools dataset were supplied with a timestamp that was removed so they would match the format of the Census date variables. Date of birth was then parsed into day, month, and year so that links could be made on each element. Reported age was supplied on VET in School data, however, age as of Census night was derived from date of birth information on the VET in schools data to account for the time difference compared to the collection of Census data. Sex and Indigenous Status were also supplied on VET in Schools data but had to be reclassified in accordance with ABS standards.

Not all the possible linking variables were used for linkage. The most common response for Country of birth recorded for students on the VET in Schools dataset was "Australia" – 35.6%, and similarly the most common response to Main language spoken at home was "English" – 68.1%. These proportions for "Australia" and "English" should be interpreted with caution as reporting information for Country of birth and Main language spoken at home was not mandatory for students. Approximately 60% of records having missing or invalid information for Country of birth and approximately 30% of records having missing or invalid information for Main language spoken at home. However, neither Country of birth nor Main language spoken at home was used for the final linkage as the validity of these variables on the VET in Schools data could not be verified by the ABS at the time of linkage. Indigenous status was also not used for linkage due to the issues in self-reporting for this characteristic, and to avoid bias in potential subsequent comparative analysis across the VET in Schools and Census datasets.

### *2.3.5 Multiple records*

In order for linkage to occur between two datasets, the records on each dataset must represent the same kind of entity. As most data linkage attempts to link persons across datasets, the aim is to have records that represent individual persons. Ideally, each person would be identified uniquely on the datasets to be linked, as multiple records for a unique person pose an issue when linking across datasets (ABS, 2013c). When duplicate records exist for a person on either dataset to be linked, the number of matched records becomes falsely inflated as it is difficult to extract matches between equivalent records on the original datasets, as the records are less distinct (ABS, 2013c). If there are unnecessary missed links, the overall linkage rate will be falsely deflated, and if there are unnecessary false links, the link accuracy will be decreased (ABS, 2013c). See Section 4.3 for more information on link accuracy.

While the VET in Schools data submitted to the ABS did not have duplicate records, it was composed of enrolment level data where the majority of persons had more than one record.  These multiple enrolment records were accounted for by the following circumstances:

- persons were undertaking a single qualification with more than one module and/or unit of competency within that qualification
- persons were undertaking more than one qualification with multiple modules and/or units of competency within those qualifications
- persons were undertaking one or more qualifications as well as one or more modules and/or units of competency outside of those qualifications
- persons were undertaking one or more modules and/or units of competency while not completing a qualification.

Individuals were identifiable by a unique person identifier developed by the NCVER. The only difference between the multiple records for an individual person was within the data related to their enrolments, such as the names of their modules or units of competency.  This is opposed to the demographic and geographic information for each person, which was replicated across all of their enrolment records.

The ABS derived a person-level file from the input VET in Schools data by choosing a singular enrolment record to represent each person.  Records were chosen from the enrolment with the highest hours of delivery within the primary qualification – as determined by the NCVER.

# 3. THE LINKAGE PROCESS

This section provides an overview of the work undertaken by the ABS to create a Vocational Education and Training in Schools to Census linked dataset.

## 3.1 Linking methodology – deterministic

Deterministic linkage is also known as exact or rule-based linkage. It involves assigning record pairs across the two datasets that match exactly or closely on common variables. Several passes of the datasets are undertaken to maximise the possibility that two matching records are compared, even when they do not match exactly on all the linking fields (ABS, 2013d).

Typically, deterministic linkage begins by using very stringent matching rules where the record pairs need to match exactly on as many linking fields as possible. It proceeds by dropping the requirement to match by one or more fields, tolerating greater differences in a field or expanding the geographic area in which a match can occur. Deterministic linkage also often involves a process where higher quality links found in the initial passes are removed from the pool of records needing to be linked and rules based on the analysis of the record pairs are used to assign link status (ABS, 2013d).

## 3.2 Implementation for this feasibility study

A modified deterministic linkage method, at a Bronze link level, was used for this study. Bronze linkages are performed where detailed identifying information, such as name and address or a common statistical linkage key, is unavailable. The linkages for this study utilised nine variables from the VET in Schools dataset and thirteen variables from the Census data set to form fifteen variable linking pairs; these are identified in table 3.1.

Combinations of these variable linking pairs were then grouped together to form 64 passes of comparisons of record pairs. These passes ranged from using more to fewer linking variable pairs, and from using variables that were more or less likely to discriminate between different people. For example, date of birth is more distinctive than sex.

The number of potential links in each pass was identified as the number of records in the VET in Schools data that linked to at least one record on the Census dataset. From this pool of potential links, each pass created a number of unique and duplicate links. Links were considered unique if only one VET in Schools record linked to only one Census record, while duplicate links occur when one VET in Schools record linked to more than one Census record.

**3.1 Vocational Education and Training (VET) in Schools and Census linking variables**

| *VET in Schools variable* | *Census variable* | *Match status* |
|---|---|---|
| Mesh Block (usual address) | Mesh Block (usual address) | Exact |
| | Mesh Block (usual address 1 year ago) | Exact |
| Statistical Area 1 (usual address) | Statistical Area 1 (usual address) | Exact |
| | Statistical Area 1 (dwelling address) | Exact |
| Statistical Area 2 (usual address) | Statistical Area 2 (usual address) | Exact |
| | Statistical Area 2 (dwelling address) | Exact |
| | Statistical Area 2 (usual address 1 year ago) | Exact |
| Statistical Local Area (usual address) | Statistical Local Area (usual address) | Exact |
| Day of birth | Day of birth | Exact |
| Month of birth | Month of birth | Exact |
| Year of birth | Year of birth | Exact, ± 1 year |
| Age | Age | Exact, ± 1 year |
| Sex | Sex | Exact |

The 64 passes were run iteratively and a certain number of unique and duplicate links were identified within each pass. A duplicate rate for each pass was calculated as the ratio of duplicate links to potential links within each pass. The unique links from each pass were then checked across all the passes to ensure they did not conflict with unique links from any other pass.

Unique links made from passes with lower duplicate rates were considered to be higher quality, therefore where unique links conflicted across different passes, matches were chosen from passes with lower duplicate rates. Since there will be more people in large geographic areas that share the same characteristics than in smaller areas, passes using a high level geographic identifier will have a high duplicate rate. As the majority of records were only coded to relatively high level geographies, the majority of unique links were made from passes with relatively high duplicate rates.

# 4. EVALUATION OF THE LINKAGE

The best way to evaluate the quality of a Bronze linked dataset is to compare it to an equivalent Gold linked dataset, as the Gold standard, where name and address is used, is assumed to identify all possible matches when the highest quality linkage methodology available is used (ABS, 2013c).

In the absence of a Gold linked dataset for comparison, the evaluation of the linkage included the following measures for this project:

- comparison of the expected number to the actual number of links between VET in Schools records and the Census to determine the impact of Census net undercount (ABS, 2011c) on the linkage (Section 4.1)
- analysis of the reporting quality of variables and their usefulness for linkage (Section 4.2)
- assessment of the estimated link accuracy (Section 4.3).

## 4.1 Comparing expected number of links to actual number of links

Initially, it is important to consider how many records might reasonably be expected to link. Persons on the VET in Schools dataset might be missing from the Census dataset for several reasons:

- they are temporarily out of the country on Census night
- they are missed by the Census, thus contributing to the Census undercount
- they emigrated from Australia before the Census
- they have died since their enrolment at school, but before the Census.

The last two of these reasons are less likely for the population represented in the VET in Schools data than for the population as a whole because the VET in Schools population is relatively young – with persons aged 15–19 years at the time of collection. Although direct estimation of the individual impact of each of these elements was not possible for this study, they are jointly taken into account in the calculation of Estimated Resident Population[4] (ERP) from Census counts (ABS, 2012). Therefore, the difference between ERP and the number of persons for whom a Census form was returned can be used to approximate the expected number of links possible for this study.

4 Further information on Estimated Resident Population is provided in the Explanatory Notes.

The first step in the estimation of how many VET in Schools records may have been available for linkage with the Census was to remove Residents Temporarily Overseas (RTO) from the ERP. The ratio of Census counts to the adjusted ERP was then applied to VET in Schools data to adjust the original number of students participating in VET in Schools by the estimated proportion of people in each state who completed a Census form. This adjustment factor is an estimate only as there was a lag between ERP which was estimated for 30 June 2011 and Census night on 9 August 2011.

Some demographic groups are more likely to be missed by the Census (ABS, 2011c). To ensure that the undercount adjustment factor was applied proportionately, the data was broken down by state and sex for the five year age group of persons aged 15–19 years. However, sex was not stated for 33 out of 236,363 persons aged 15–19 years on the VET in Schools data, and sex was proportionally imputed onto the not stated records to account for this. Figures were also calculated without incorporating sex as a factor. Additionally, ERP estimates that excluded RTOs were not available for 20 year olds, and as such 20 year olds have been excluded from all figures contributing to the calculations. The available breakdowns for each group were adjusted as follows:

$$\text{Persons }(\text{state , sex}) = \frac{\text{Census counts }(\text{state , sex})}{\text{ERP}}.$$

The expected links were then summed for each state. Table 4.1 shows the total number and expected number of VET in Schools person records aged 15–19 years available for linking. It also shows the linkage rates before and after adjusting for the expected number of links, to demonstrate the impact of Census net undercount on linkage.

**4.1 Linkage rates, adjusted for expected links**

| | *Bronze linked records* |
|---|---|
| VET in Schools data, ERP equivalent, persons aged 15–19 years | |
| Persons (no.) | 236,363 |
| Expected links, adjusted by sex (no.) | 228,711 |
| Expected links, not adjusted by sex (no.) | 228,699 |
| Number of persons linked (no.) | 116,700 |
| Proportion of persons linked (%) | |
| Pre-adjustment | 49.4 |
| Post-adjustment, adjusted by sex (%) | 51.0 |
| Post-adjustment, not adjusted by sex (%) | 51.0 |
| All VET in Schools data, persons aged 15–20 years | |
| Persons (no.) | 236,461 |
| Number of persons linked (no.) | 116,741 |
| Proportion of persons linked (%) | 49.4 |

As can be seen in table 4.1, Census net undercount had minimal impact on the success of the linkage process; with approximately 1.6% of 15–19 year old persons in the VET in Schools data not expected to link. Incorporating sex as a factor in the calculations for the number of expected links did not have noteworthy impact on the adjusted linkage rates.

## 4.2 Analysis of the linking variables

As discussed in Section 4.1, the lack of a corresponding Census record may have accounted for some of the VET in Schools records that were not linked. However, the major contributor to records not being linked was insufficient data on the VET in Schools records or corresponding Census records. Data were considered insufficient if:

- there was missing or incomplete information for the linking variables on either dataset
- the collected or derived data items were not reliable
- the variables used had low efficacy for identifying links.

### *4.2.1 Missing data*

Table 4.2 shows details of the linking variables which have missing values for the input and non-linked VET in Schools records. Input data records are all the records that were available for linkage – 236,461, while non-linked data records are all the records that did not match uniquely to a Census record – 119,720; proportions are calculated from these counts. Missing values are inclusive of not stated or invalid values.

The variables that were most impacted by missing information on the Census data used for linkage were day and month of birth with approximately 10% of records that had missing information.[5] All other Census linking variables had mostly complete data with a small amount of missing information, less than 5%, coming from imputed records.

5 Missing data for day and month of birth on the Census data can be mostly attributed to respondents being able to provide either their age or their date of birth.

**4.2 Missing or invalid values for input and non-linked VET in Schools records**

| | *Records with missing data* | |
|---|---|---|
| *Linking variable* | *Number* | *Proportion (%)* |
| | INPUT DATA | |
| Mesh Block, usual address | 234,844 | 99.3 |
| Statistical Area 1, usual address | 227,235 | 96.1 |
| Statistical Area 2, usual address | 31,541 | 13.3 |
| Statistical Local Area, usual address | 2,355 | 1.0 |
| Day of birth | – | – |
| Month of birth | – | – |
| Year of birth | – | – |
| Age, on 9 August 2011 | – | – |
| Sex | 33 | <0.1 |
| Total records | 236,461 | 100 |
| | NON-LINKED DATA | |
| Mesh Block, usual address | 119,058 | 99.5 |
| Statistical Area 1, usual address | 115,891 | 96.8 |
| Statistical Area 2, usual address | 21,582 | 18.0 |
| Statistical Local Area, usual address | 2,200 | 1.8 |
| Day of birth | – | – |
| Month of birth | – | – |
| Year of birth | – | – |
| Age, on 9 August 2011 | – | – |
| Sex | 22 | <0.1 |
| Total records | 119,720 | 100 |

### *4.2.2 Reliability of derived data items*

The largest known impact on the reliability of the Census data items stems from non-response. Non-response occurs when a person filling out a Census form does not enter a valid response for a question or questions. This may take the form of a complete lack of response, a response that cannot be interpreted or categorised, or the selection of more than one option where only one is required (ABS, 2013a). For the most part, non-response is addressed by categorising the response to 'not stated' or missing, or for a small number of selected variables, through imputation. As imputed values can create false links, the Census data used for this study had imputed values set to missing. Missing, invalid, and not stated responses are addressed in Section 4.2.1.

The VET in Schools data is also subject to non-response. However, the largest known accuracy issue for the VET in Schools data involves the geographic variables that were used for linking, specifically those that were derived from input data. The suburb and postcode of a person's residential address are the only geographic data items available for persons on the VET input data held by the NCVER. Prior to providing the VET data to the ABS, the NCVER used an ABS geography correspondence table to allocate SLAs to persons based on their postcode. The relationship between postcode and SLA in 2011 is considered to be of acceptable quality overall. However, caution needs to be applied in using SLA as the quality of the corresponding data will vary and may differ in parts from the actual characteristics of the geographic regions involved (ABS, 2013f).

While processing the data, the ABS used a geocoding process to allocate SA2s, SA1s, and Mesh Blocks to VET in Schools person records using postcode and locality. While the quality of this process is generally good, in some cases a person could be assigned a geographic code that does not represent their actual geographic location. Both the rate and the quality of the linkage will be affected by the quality of the geocoding. Further work could be undertaken to assess the quality of linkage using data that has used geographic correspondences or has been subject to geocoding. Notwithstanding these issues, the derived geographic codes were considered of sufficient quality to draw reliable conclusions from this feasibility study on the impact of geography on linkage rates and quality.

The presence of incorrect geographical information, which is likely to be the case for some records, means that some of the links made in this study have a chance of being false and that true links may not have been made. While this means that some of the links made using the derived geographic variables will not be true matches, all unique links have been retained to indicate the rate and quality of linkage that could be achieved with various geographic variables if they were accurate.

### *4.2.3 Efficacy of variables for linkage*

As referenced in the previous sections, the effectiveness of a variable for data linkage can be partially attributed to accuracy. Another key factor that contributes to the effectiveness of a variable for linkage is how well it discriminates between individuals within a group. This discriminative power is important for linking as it helps to find a common entity, or match, across different datasets. Although sex is a variable that usually has high reporting quality, it does not assist much with matching individuals across datasets as it only splits those datasets into two roughly equal groups. Date of birth, when combined with sex will break up these groups further and additional variables such as Country of birth, assuming they are accurate, can break up the groups further still. However, a geographical component on top of these other linking variables is almost essential to finding true matches.

Geographic data items that represent smaller populations, such as SA1 with 54,805 regions in Australia with population ranges between 200 and 800 persons (ABS, 2013e), are more effective at finding matching records between datasets. On the other hand, variables that represent larger populations, such as SA2 with 2,214 regions in Australia with population ranges between 3,000 and 25,000 persons (ABS, 2013e), are far less distinctive. This remains the case when other variables are introduced to the process; there are likely to be fewer people sharing the same date of birth and sex within a Mesh Block than within an SA2.

The efficacy of some of the variables used in this linkage study was considered to be of a moderate level of distinction. Date of birth and sex information provided reasonable distinction between individuals within the linking variable groups. As the derived geographic codes do not necessarily represent the exact geographic location of persons on the VET in Schools data, the distinction these variables provide between individuals cannot be reliably guaranteed.

## 4.3 Link accuracy

The best possible assessment of the link accuracy of a Bronze linked dataset comes from comparison with a Gold link using the same data (ABS, 2013c). However, as name and address were not available on either dataset, a Gold link was not able to be explored and was not available for comparison. This makes it difficult to measure link accuracy with absolute confidence; however, some useful estimates of link accuracy are available.

The first measure used to estimate link accuracy was the duplicate rate of a linkage pass. As mentioned in Section 3.2, the duplicate rate for each pass was calculated as the ratio of duplicate links to potential links within each pass. Passes with low duplicate rates were less likely to identify unique links by chance than passes with high duplicate rates. While the duplicate rate does not necessarily tell us how accurate an individual link is, it does tell us which variables are more likely to identify links accurately. As sex, date of birth, and age were used for most of the passes, they were treated as a constant and duplicate rates were compared for the derived geographic variables. Table 4.3 shows the duplicate rates for these derived geographic variables.

The second measure to estimate link accuracy rebases the duplicate rate using the amount of missing information for the linking variables used within a pass. A greater amount of missing data increases the likelihood of a false link being made. This is because links are only made on non-missing data and for any link that is accepted a more accurate link may have been accepted if all the data was available on the corresponding records.

The method of calculating the rebased duplicate rate is shown below; lower numbers for the rate indicate a lower likelihood that links within a pass were made by chance. As with the duplicate rate, these probabilities are shown for the derived geographic linking variables in table 4.3.

$$\text{Duplicate rate rebase} = \text{Duplicate rate} \times \left(1 - \text{Probability of non-missing data}\right)$$

$$\text{where} \quad \text{Probability of non-missing data} = \frac{\begin{pmatrix}\text{Number of records in Census} \\ \text{dataset with non-missing data} \\ \text{for linking variables of a pass}\end{pmatrix}}{\text{Total records in Census dataset}}$$

Link accuracy can also be estimated using the proportion of unique links that could have been identified within each pass. As mentioned in Section 3.2, the final pool of unique links that were accepted was influenced by the duplicate rate, meaning that some passes did not contribute unique links to the final pool. However, each pass may have been capable of creating unique links if the passes were run independently of the duplicate rate. This measure is represented as the proportion of potential unique links over all potential links for each derived geographic linking variable in the last column of table 4.3.

**4.3 Efficacy of levels of derived geography variables used for linkage**

| *Variable* | *Duplicate rate (%)* | *Duplicate rate rebased (a) (%)* | *Potential unique links (%)* |
|---|---|---|---|
| Mesh Block, usual address | 1.62 | 0.18 | 76.65 |
| SA1, usual address | 2.36 | 0.3 | 73.94 |
| SA2, usual address | 47.59 | 5.41 | 37.61 |
| SLA, usual address | 45.26 | 5.76 | 39.73 |

(a) Percentages for the rebase do not represent a true probability measure for calculating the chance of error. For example, 5% does not mean the pass made accurate links with 95% confidence, or that 95% of the links in the pass were accurate.

From table 4.3 we can identify that the smaller area geographic variables, namely Mesh Block and SA1, produced links with a higher estimated rate of accuracy than the larger geographic variables, SA2 and SLA.

The final measure of accuracy is uniqueness. By only accepting unique links, we can be confident that the links are accurate conditional on some assumptions. First, we have to assume that the population on the VET in Schools data is fully represented on the Census data. As mentioned in Section 4.2, there may be some people on the VET in Schools data that, for one reason or another, were not on the Census. However, the chance of this happening could be considered to be small enough to assume the same population was represented across the two data sources.

The second assumption we have to make is that the linking variables used to make the unique links are accurate. As indicated in Section 4.2.2, this may not be the case for the derived geographic linking variables used in this study.

However, if we are confident that the population to be linked is present on both datasets and the linking variables are accurate, then the likelihood of the unique links being accurate is significantly increased. This is because if a VET in Schools record linked uniquely to a Census record, then it did not link uniquely to any other Census record across the linkage passes. This means that if the linking variables are of a high quality, a unique link is more likely to represent a true match.

## 4.4 Results

The purpose of this study was to test the feasibility of linking VET in Schools data to Census data, and to provide information about the rate and quality of potential linkage. The linkage methodology used in this study provided sufficient information to evaluate the feasibility of a link, as such the possibilities within the methodology and alternate methodologies were not fully explored.

Initially, links were only accepted if they were unique, with no conflicting record pairs from either the Census or the VET in Schools datasets. Using this restriction, 116,741 or 49.4% of the 236,461 VET in Schools person records were linked to Census. This low linkage rate is mostly due to the limited number of compatible linking variables available and in particular, the relatively large area level of geography available on the VET in Schools data set. A further 113,492 or 48.0% of the 236,461 VET in Schools person records could have been linked to Census by allowing matches to be randomly assigned where there were duplicate links made, that is where one VET in School record linked to two or more Census records. However, as there was at least a one in two chance that these randomly assigned links would have been incorrect, they were not accepted as links. Further work could be done to extract unique links from this dataset through clerical review.

In consideration of the low reliability of key linkage variables, as discussed in Section 4.2.2, the links that were accepted should be deemed indicative only. Even though the links were unique, we have limited confidence that these are true links due to the reliability of the linking variables involved.

## 4.5 Recommendations for future improvements

As highlighted in Section 4.2, this study set out to test the feasibility of linking VET in Schools person records to the Census. The linked dataset produced from this study was not of an acceptable level of quality to be used for reporting and analysis. However, significant enhancements to linkage would likely be gained through improvements to data. Specifically, the availability of accurate small area geography, such as Mesh Block and SA1, on the VET in Schools data would not only increase the rate of linkage but also would result in vast improvements to linkage efficacy and accuracy.

If small area geography does not become available, more work could be done to identify potential approaches to finding greater commonality between the existing Census and VET in Schools data in the geographic variables used for linkage. While it is expected that this would lead to improved link accuracy, the rate of linkage would likely be relatively low due to the low efficacy of large area geography variables for linkage. See Section 4.2.3 for more on the efficacy of different linking variables.

Further to expanding the availability of quality geographic data, improvements to linkage could also be achieved by looking beyond the unique links. As mentioned in Section 4.2, only unique links were accepted for this study. This leaves a pool of duplicate links; from which matches may be found. Variables that were common to VET in Schools and Census data, but were not used for linking, could be used to source unique links from the pool of duplicate links through clerical review. These variables include Country of birth, Main language spoken at home, Indigenous status, and Highest level of education completed. Incorporating these additional variables into a link selection process could improve on the random selection undertaken to enhance the linkage coverage discussed in Section 4.4.

Finally, the exploration of new linking methodologies may also produce improvements in the rate and accuracy of linkage. Probabilistic linking in particular may account for missing or erroneous data for the linking variables better than a deterministic approach; as probabilistic linking ranks how well variable pairs agree rather than searching for exact, or near exact, agreements as deterministic linking does (ABS, 2013c). However, previous studies have found that deterministic methodologies are superior for Bronze linkage where the linking variables are not character strings prone to subtle differences across collections, such as school name (ABS, 2013d). While additional methodologies were not fully explored as part of this study, it is acknowledged that alternative methodologies would be likely to improve the linkage. However, the best possible improvements to the coverage and accuracy of linkage are more likely to result from additional information on the VET in Schools data.

# 5. CONCLUSIONS

The purpose of this study was to test the feasibility of linking Vocational Education and Training (VET) in Schools data to data from the 2011 Census of Population and Housing. Through exploring a recently developed deterministic linkage methodology, the study has been able to produce recommendations for ways of proceeding with linking VET in Schools data to deliver high quality linked datasets for research and statistical purposes.

This paper has detailed a Bronze deterministic method used to link 2011 VET in Schools person-level records to the 2011 Census. The method that was explored in this study was not successful in producing a dataset that adequately covered all of the possible links to a reasonable level of accuracy for reporting and analysis purposes.

However, similar linkage methods have proven successful in previous studies that have linked education data to the Census. These methods are reproducible and can be utilised to facilitate and benchmark future linkages. These previous studies have shown that deterministic Bronze methods proved both more effective in terms of approaching Gold quality, and more efficient in terms of requiring fewer calculations and less clerical review for the linkage process than Bronze probabilistic linkage (ABS, 2013d).

Census undercount had a small impact on the success of the linkage, with a very small gain in the linkage rate after adjusting for net undercount. However, the quality of linkage cannot be singularly attributed to the census undercount, as the measurement from this adjustment is an estimate and the impact was very small.

The largest impact on the success of the linkage came from the linking variables used. Insufficient levels of common information between the VET in Schools data and the Census, in particular, small area geographic information, negatively impacted on the coverage and accuracy of data linkage. Both the NCVER and the ABS were able to derive additional geographic data onto the VET dataset. However, due to the lack of detail in the geographic input data provided, the derived geographic data was not considered accurate enough to create true links for reporting and analytical purposes. It is anticipated that more detailed geographic information will become available on the VET data in the future as a result of changes to the Australian Vocational Education and Training Management Information Statistical Standard, under which VET data is submitted to the NCVER (NCVER, 2013).

However, while the rate and accuracy of linkage was low, the process created some useful information as to which linking variables produced more accurate links. Links were made with various levels of geography along with date of birth, sex, and age. While there were accuracy issues with the derived geographic variables, the unique links made were retained to give an indication of the linkage quality that would be achieved with more accurate and higher quality geographic linkage variables. The findings show that, with all other factors held equal, links made with smaller area geography codes, such as Mesh Block and SA1, were much more likely to be accurate than links made with large area geography codes, such as SA2.

Linking education data to the Census has proven useful in enriching the socio-economic and demographic information available for participants in the relevant education programs. Data integration has also proven useful for filling in some of the data gaps left by missing data, and also in identifying some data that has not been correctly collected or derived (ABS, 2013c). The VET datasets also provide better coverage of people within the VET system than does the Census, which has a relatively high proportion of missing data for type of educational institution. To gather the data that could be derived from an integrated dataset from persons, parents or caregivers, or training organisations involved in the VET system through a detailed survey or extended enrolment form would require an impracticable level of respondent burden and administrative work (ABS, 2013d).

Despite the fact that this study failed to produce a linked dataset of an acceptable level of quality for further analysis, linkage is likely to be feasible with improved data and linkage methodologies. Searching for opportunities to link the data in the future will contribute to the transformation of education and training data from discrete and somewhat fragmented data collections to an integrated and enhanced research base of participation, attainment and socio-demographic information (ABS, 2013d).

Work to integrate VET in Schools data with Census data is to be continued as part of the Transforming Education and Training Information (TETIA) in Australia initiative funded and overseen by the Strategic Cross-sectoral Data Committee.

# REFERENCES


Australian Bureau of Statistics (2011a) *Australian Standard Classification of Languages (ASCL)*, cat. no. 1267.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1267.0>

—— (2011b) *Census Dictionary, 2011*, cat. no. 2901.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/2901.0>

—— (2011c) Information *Paper: Measuring Net Undercount in the 2011 Population Census, 2011*, cat. no. 2940.0.55.001, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/2940.0.55.001>

—— (2011d) *Standard Australian Classification of Countries (SACC)*, cat. no. 1269.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1269.0>

—— (2012) *Australian Demographic Statistics, March 2012*, cat. no. 3101.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0>

—— (2013a) *2011 Census of Population and Housing > Data quality > Non-response rates*, ABS website content, ABS, Canberra. <http://www.abs.gov.au/websitedbs/censushome.nsf/home/nonresponserates>

—— (2013b) *ABS Statistical Data Integration*, ABS website content, ABS, Canberra. <http://www.abs.gov.au/dataintegration>

—— (2013c) "Assessing the Quality of Linking School Enrolment Records to 2011 Census Data", *Methodology Research Papers*, cat. no. 1351.0.55.041, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.041>

—— (2013d – forthcoming) "Assessing the Quality of Linking School Enrolment Records to 2011 Census Data: Deterministic Linkage Methods", *Methodology Research Papers*, ABS, Canberra.

—— (2013e) *Australian Statistical Geography Standard (ASGS): 2011 Structure and Summary*, ABS website content, ABS, Canberra <http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS)>

—— (2013f) *Australian Statistical Geography Standard (ASGS): Correspondences*, July 2011, cat. no. 1270.0.55.006, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.006>

Cross Portfolio Statistical Integration Committee (CPSIC) (2010) *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*, Australian Government, Canberra. <http://www.nss.gov.au/nss/home.NSF/pages/High+Level+Principles+for+Data+Integration+-+Content?OpenDocument>

National Centre for Vocational Education Research (2013) *AVETMISS 7.0 VET Provider Collection Specifications*, NCVER, Adelaide. <http://www.ncver.edu.au/publications/2578.html>

All URLs viewed 19 November 2013.

## ACKNOWLEDGEMENTS

This Feasibility Study was funded by the Strategic Cross-sectoral Data Committee as part of the Transforming Education and Training Information in Australia (TETIA) initiative 2013–14 work plan.

The ABS acknowledges assistance provided through a Memorandum of Understanding and ongoing consultation with the National Centre for Vocational Education Research. Special thanks go to Davinia Blomberg and Paul Foley.

This paper was prepared in the ABS National Centre for Education and Training Statistics, with assistance from the ABS Education and Training Census Data Enhancements team. The principal author and researcher was Luke Verghese.

The author would like to thank the ABS National Centre for Education and Training Statistics Administrative Data Centre, the ABS National Centre for Education and Training Statistics Census Data Enhancements team, the ABS Data Linkage Centre, ABS Analytical Services, ABS Geography, ABS Demography, and the ABS Data Integration National Statistics Centre. Special thanks go to Andrew Howie, Rita Scholl, Presley Peter, Noel Hanson, Brendan Kelly, Caitlin Szigetvari, Peter Rossiter, Myles Burleigh, Regina Kraayenbrink, and Carmen Kong.

## GLOSSARY

| | |
|---|---|
| Administrative data | Information that is collected for purposes other than that of a statistical nature. This type of information is often obtained from records or transactional data from government agencies, businesses or non-profit organisations which use the information for administrative purposes. |
| Bronze linkage | Linking data without the use of name and address or a statistical linkage key. |
| Competency standard | An industry-determined specification of performance, which sets out the skills, knowledge and attitudes required to operate effectively in employment. In Vocational Education and Training, competency standards are made up of units of competency, which are themselves made up of elements of competency, together with performance criteria, a range of variables, and an evidence guide. Competency standards are an endorsed component of a training package. |
| Deterministic linkage | Deterministic linking compares only record pairs that match exactly or almost exactly (e.g. age within one year) on a combination of variables, seeking unique matches wherever possible. |
| Gold linkage | Linking data with the use of name and address. |
| Locality | An area represented by the officially recognised boundaries of suburbs (in cities and larger towns) and localities (outside cities and larger towns). |
| Module (enrolment) | A self-contained block of learning which can be completed on its own or as part of a course and which may also result in the attainment of one or more units of competency. |

| | |
|---|---|
| Probabilistic linkage | Probabilistic linking compares records from two datasets using several variables common to both datasets and generates a single numerical measure of how well two particular records match. This allows ranking of all possible record pairs and assignment of the optimal link. |
| Registered training organisation | An organisation registered by a state or territory registering and accrediting body to deliver training and/or conduct assessments and issue nationally recognised qualifications in accordance with the Australian Quality Training Framework. |
| Statistical linkage key | A key that enables two or more records belonging to the same individual to be brought together. It can be derived from particular characters of a person's name as well as other data elements such as the day, month and year when the person was born and the sex of the person, concatenated together. |
| Unit of competency (enrolment) | A component of a competency standard (see above). A unit of competency is a statement of a key function or role in a particular job or occupation. |

# EXPLANATORY NOTES

*1. Australian Statistical Geography Standard (ASGS)*

The ASGS provides a common framework of statistical geography which enables the production of statistics that are comparable and can be spatially integrated. To assign statistical geography, statistical units such as households are first assigned to a geographical area in one of the ASGS structures. Data collected from these statistical units are then compiled into ASGS defined geographic aggregations which, subject to confidentiality restrictions, are then available for publication. The geographic aggregations used for the purposes of this study are given below.

***Mesh Blocks*** are micro-level geographical units for statistics and there are in excess of 300,000 Mesh Blocks covering the whole of Australia. A residential Mesh Block typically contains 30 to 60 dwellings. A street address can be coded to the appropriate Mesh Block, but Mesh Blocks cannot be coded back to a specific street address. Mesh Block is a useful linking variable when street address is not available.

***Statistical Area Level 1*** (SA1) is the second smallest geographic area defined in the ASGS after Mesh Block. The SA1 has been designed for use in the Census of Population and Housing as the smallest unit for the processing and release of Census data. SA1s are useful linking variables as they are still able to capture those who move within their local area without being so broad as to increase the possibility of matching different people who share similar characteristics, i.e. false links.

***Statistical Area Level 2*** (SA2) is an area defined in the ASGS, which consists of one or more whole Statistical Areas Level 1 (SA1s). Wherever possible, SA2s are based on officially gazetted State suburbs and localities. In urban areas, SA2s largely conform to whole suburbs and combinations of whole suburbs, while in rural areas they define functional zones of social and economic links. This level is broad enough to capture the majority of matching pairs, where geocoding to the locality (town or suburb) has been reasonably accurate.

*2. Estimated Resident Population (ERP)*

The ERP figures used in this paper are based on the 2011 Census of Population and Housing. ERP is an estimate of the Australian population obtained by adding to the estimated population at the beginning of each period the component of natural increase (on a usual residence basis) and the component of net overseas migration. For the states and territories, estimated interstate movements involving a change of usual residence are also taken into account. Estimates of the resident population are based on Census counts by place of usual residence, to which are added the estimated Census net undercount and the number of Australian residents estimated to have been temporarily overseas at the time of the Census. Overseas visitors in Australia are excluded from this calculation.

For more information, see *Australian Demographic Statistics, March 2012 – Explanatory Notes* (ABS, 2012).

## FOR MORE INFORMATION . . .

| | |
|---|---|
| *INTERNET* | **www.abs.gov.au** The ABS website is the best place for data from our publications and information about the ABS. |
| *LIBRARY* | A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries. |

### INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

| | |
|---|---|
| *PHONE* | 1300 135 070 |
| *EMAIL* | client.services@abs.gov.au |
| *FAX* | 1300 135 211 |
| *POST* | Client Services, ABS, GPO Box 796, Sydney NSW 2001 |

## FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

| | |
|---|---|
| *WEB ADDRESS* | www.abs.gov.au |