1351.0.55.047

Australian Bureau of Statistics

Research Paper

Assessing the Quality of Different Data Linking Methodologies Across Time, Using Tasmanian Government School Enrolment Data

www.abs.gov.au

New Issue

Research Paper

Assessing the Quality of Different Data Linking Methodologies Across Time, Using Tasmanian Government School Enrolment Data

National Centre for Education and Training Statistics

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 21 MAR 2014

ABS Catalogue no. 1351.0.55.047

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 QUALITY OF DIFFERENT DATA LINKING METHODOLOGIES ACROSS TIME, USING TASMANIAN GOVERNMENT SCHOOL ENROLMENT DATA

A Data Integration Feasibility Study

EXECUTIVE SUMMARY

Purpose

The purpose of this feasibility study is to assess the quality and accuracy of using the Statistical Linkage Key (SLK) 581 as a linkage method for education and training data integration projects. The Northern Territory feasibility study *Assessing the Quality of Different Data Linking Methodologies Across Time, Using Northern Territory Government School Enrolment Data*, provides an in-depth analysis of a number of different linking methods that could be used for data integration projects, including the SLK. This feasibility study aims to contribute to the findings from the Northern Territory study, by linking Tasmanian government school enrolment data over a longer period of time using the SLK linking method, and comparing analysis outcomes to data linked using the student identifier (Student ID).

Key findings

The results achieved by the two linking methodologies were fairly consistent overall and produced a very similar distribution of the population. For example, the majority of students were in the Leavers cohort (enrolled in 2006 and left before 2011) with results varying from 36.4% using the Student ID method to 37.1% using the SLK method.

Overall, this study found that the SLK provides a feasible alternative to linking using the Student ID, with a high level of linking accuracy (95.7%) being achieved by using the SLK linkage method. The ability to accurately link using the SLK is important because it allows linking between state and territory datasets, linking between government and non-government school sector data and also linking with other datasets where there is no comparable Student ID available. Using the SLK as a linking variable reduces the need to retain name and address information on datasets for linking purposes, which contributes to ensuring the privacy and confidentiality of student level data.

Future data integration projects

The outcomes from this report have highlighted a number of actions that would contribute to the quality of education and training data integration projects. These include:

1. Undertaking an additional feasibility study that replicates the SLK analysis for a larger state. This would provide a larger population on which to conduct additional testing of the quality of the SLK linkage methodology.
2. Continuing to work towards the improvement of data collected in the National Schools Statistics Collection (NSSC) to ensure the quality of data integration initiatives. In particular, options for including unit record level data from the non-government school sector should continue to be explored.

ACKNOWLEDGEMENTS

This feasibility study was funded by the ABS and the Strategic Cross-sectoral Data Committee in 2012–13, and is consistent with the directions of the Transforming Education and Training Information in Australia (TETIA) initiative. The TETIA initiative provides an overarching vision, and a strategic and staged approach to transform and coordinate continued improvements to cross-sectoral early childhood, education, higher education and training information in Australia over the next five to ten years.

The ABS acknowledges assistance provided through a Memorandum of Understanding with the Tasmanian Department of Education. Special thanks to Andrew Oakley and Leanne Males.

This paper was prepared in the ABS National Centre for Education and Training Statistics. The principal authors and researchers were Katherine Thomson and Megan Jobson. Special thanks go to Rita Scholl, Caitlin Szigetvari, Peter Rossiter, Regina Kraayenbrink, and Alan Herning.

CONTENTS

ASSESSING THE QUALITY OF DIFFERENT DATA LINKING METHODOLOGIES ACROSS TIME, USING TASMANIAN GOVERNMENT SCHOOL ENROLMENT DATA

A Data Integration Feasibility Study

ABSTRACT

This feasibility study contributes to the findings of the ABS Research Paper, *Assessing the Quality of Different Data Linking Methodologies Across Time, Using Northern Territory Government School Enrolment Data*, by providing additional evidence on the quality and accuracy of using the Statistical Linkage Key (SLK) 581 as a linkage method for education and training data integration projects. This study involved linkage of 2006 to 2011 Tasmanian school enrolment data using two deterministic linking methods:

- linking with Student ID, and
- linking with SLK.

Once the datasets were linked across years, educational pathways were created to represent when a student was enrolled in a Tasmanian government school. The overall results achieved by the two linking methodologies were fairly consistent overall and produced a very similar distribution of the population. For example, the majority of students were in the Leavers cohort (enrolled in 2006 and left before 2011) with results varying from 36.4% using the Student ID method to 37.1% using the SLK method.

This study confirms the accuracy of the SLK linking method for linking data longitudinally across six years of government school data. The high level of accuracy (97.5%) achieved using the SLK linking method indicates that the SLK would be a good quality linking method for education and training data integration projects, and would provide a feasible alternative to linking using Student ID.

1. INTRODUCTION

The purpose of this feasibility study is to compare the quality of data linkage results by using different linkage methodologies, specifically by comparing linkage results achieved by the Student ID and the Statistical Linkage Key (SLK).

Understanding the different linkage methodologies and the effect these methodologies may have on statistical output is important for moving forward with maximising the statistical and research value of administrative datasets, including for data integration projects. Whilst Student ID is a very good identifier to use within the government school system (where it is derived from a comprehensive student database), an identifier such as Student ID is not always available or consistently applied across different datasets, states or territories. Similarly, linkage using full name and address information, although considered a gold standard linkage methodology, is a methodology that comes with many confidentiality, privacy, legislative and data access considerations. Therefore, the SLK linkage method is a useful alternative that can be derived from any dataset that originally has full name, date of birth and sex information, with full name details being removed from the file once the SLK is derived. This study will compare analysis results using the Student ID as the benchmark linking method to the results achieved using the SLK method, to show how the different methodologies might affect the data output.

It is expected that the lessons learnt from this feasibility study will be transferable and applicable across all Australian student enrolments, which in the long term will assist in building a comprehensive picture of the educational pathways of all students in Australia.

2. BACKGROUND

Under a Memorandum of Understanding between the Australian Bureau of Statistics (ABS) and the Tasmanian Department of Education, it was agreed that student enrolment unit record level data files for the 2006 to 2011 school enrolment years would be used for the purposes of ABS data integration feasibility studies.

These files contained two student identifiers, the Tasmanian government schools unique Student ID and a student SLK, as used in many health linkage projects.

Other data items on the files aligned to the Data Collection Manual (DCM) standards for the National Schools Statistics Collection (NSSC)(ABS, 2013a), used for student enrolment unit record level data provision. As the NSSC DCM requirements have changed over the period from 2006 to 2011, some data items were not available for all six years of data. Please see Appendix A of this report for more information.

3. QUALITY ASSURANCE PROCESS

In 2011–12 a preliminary feasibility study was undertaken by the ABS to understand the student populations across 2006 to 2011, in Tasmanian government schools. The quality assurance included comparing the URL data supplied for that study to unconfidentialised pre-published *Schools, Australia* (ABS, 2013b) data, sourced from the National Schools Statistics Collection for verification of accuracy and consistency. Comparisons included aggregated counts of schools, counts of students by grade level and by characteristics including Indigenous status.

The apparent retention rates table and the table showing counts of enrolments, counts of students, and counts of students with multiple enrolments in Tasmanian government schools, produced as part of the preliminary study were then replicated for the present study. This ensured that the same data were being used for both studies. The tables produced for this quality assurance process can be found in Appendix C.

4. LINKING METHODOLOGY

Linking between the 2006 to 2011 datasets was carried out by applying deterministic linking methods. One of the deterministic linking methods that was used for this study involved the use of a unique identifier that was common to both datasets. A link is declared successful when the identifiers are exactly the same on both datasets. This type of matching can only occur when both datasets have unique and high quality identifiers that have been assigned consistently to all records on the datasets.

In the absence of a unique identifier, a linkage key can be used as the linking variable. A linkage key is created by consistently combining a number of variables into a string for each unit in the dataset. The linkage key is not unique as there is always a chance that more than one unit in the population may have identical responses for the variables used for linking. This concept will be discussed further later in this report.

For the purposes of this study, student level data with a high quality unique identifier (Student ID) attached to each student record was required. The Tasmanian Department of Education collects data for the NSSC via the Government Schools Administrative Computer System, which requires extensive validation and quality assurance checks on the data submitted from each school. Due to the nature of this administrative system, the ABS National Centre for Education and Training Statistics (NCETS) project team considered the Student ID to provide a high quality benchmark against which the SLK linkage method could be compared.

Prior to linking, the datasets were all prepared using the same process, which is described in Appendix A of this report. Figure A.1 in Appendix A shows how the files were constructed, and how they were linked using the SLK compared with linking using the Student ID.

The two types of deterministic linking conducted for this study included:

- linking with the Student ID (the benchmark method)
- linking with the SLK.

4.1 Student identifier (ID)

The Student ID is a unique identifier assigned to each individual student enrolled in the Tasmanian government school education system, from pre-year 1 to senior secondary. The Student ID remains with the same student throughout their schooling in the Tasmanian government school system, even when moving between different Tasmanian government schools. A student enrolled in more than one Tasmanian government school at the same time (e.g. a student that attends one of their classes at a different school) would have the same Student ID recorded for both enrolments.

Each Student ID is unique to a single student and therefore if two exact matches of a Student ID are found within a dataset, this would indicate that the student has multiple enrolments within a year.

Exact match linking was undertaken using the Student ID across the 2006 to 2011 datasets. Any records that did not have a corresponding match on the other file were not linked.

4.2 Statistical Linkage Key 581 (SLK)

The SLK 581 is a type of statistical linkage key that was developed by the Australian Institute of Health and Welfare (AIHW, 2013). The SLK 581 consists of a string of variables concatenated in the following order:

1. the second, third and fifth letters of a person's last name;
2. the second and third letters from a person's first name;
3. date of birth (ddmmyyyy);
4. an identifier for sex (1=male, 2=female).

The SLK does not produce a unique student identifier; it is a non-unique linking key and it is possible for different students to have the same SLK (if they have a similar name, same sex and same date of birth). It is also possible that a student could have more than one SLK over the course of their schooling. For example, a student's surname may change (due to a change in family structure), or their first name may change (due to personal or cultural reasons).

As the SLK is made up of components of student characteristics, the quality of the SLK is directly affected by the quality of these data items.

The SLK can be applied to any dataset where name, date of birth and sex information is available. In this feasibility study, linking with the SLK was achieved through exact matching techniques.

It is important to note that each record within a single year relates to an individual student, given the preparation process undertaken on both files to remove multiple enrolments using the Student ID. Therefore, all SLKs indicated unique students, despite the fact that there may have been multiple occurrences of the same SLK. Due to this, it was necessary to use a 'duplicate SLK flag' as part of the linking process to ensure that these records were only linked to the other dataset once. As a result of this flag, a number of records were identified as linking incorrectly. Figure A.2 in Appendix A shows the application of the 'duplicate SLK flag'.

If it is found that the SLK is a feasible linking variable, it will potentially allow linking of the Tasmanian government school data with data from other sources, such as early childhood education data, data from the non-government schools sector, school performance data – e.g. National Assessment Program–Literacy and Numeracy (NAPLAN) data (see ACARA, 2011) – or higher education data.

5. EVALUATION OF LINKAGE USING THE STATISTICAL LINKAGE KEY

A similar feasibility study undertaken for the Northern Territory analysed the suitability and quality of the SLK as a linking method. This included detailed analysis of the match-link rate and link accuracy of the SLK linked dataset compared with the Student ID linked dataset. This Tasmanian feasibility study provides a further evaluation of the SLK as a linking method, by comparing the analysis results with the Student ID linked dataset, which is used as the benchmark linking method.

For the SLK linking methodology, enrolment level data was initially linked within each year using the Student ID in order to create a student level file. This meant that students with multiple enrolments within a year were concatenated into one student record. The student's main school enrolment was then used for subsequent linking across 2006 to 2011 using the SLK linking method.

Once the datasets were linked across years, educational pathways in the form of binary strings of length six (representing the six years of linked data) were created to represent when a student was enrolled in a Tasmanian government school for a particular enrolment year.

This higher level analysis demonstrates that the SLK linkage method results in a slightly under-represented distribution for continuous students, and a slightly over-represented distribution for leaving students. The overall results achieved by the two linking methodologies are fairly consistent overall and produce a very similar distribution of the population.

5.1 Distribution of students by education pathway, SLK and Student ID linking methods, Tasmania, 2006–2011[(a)]

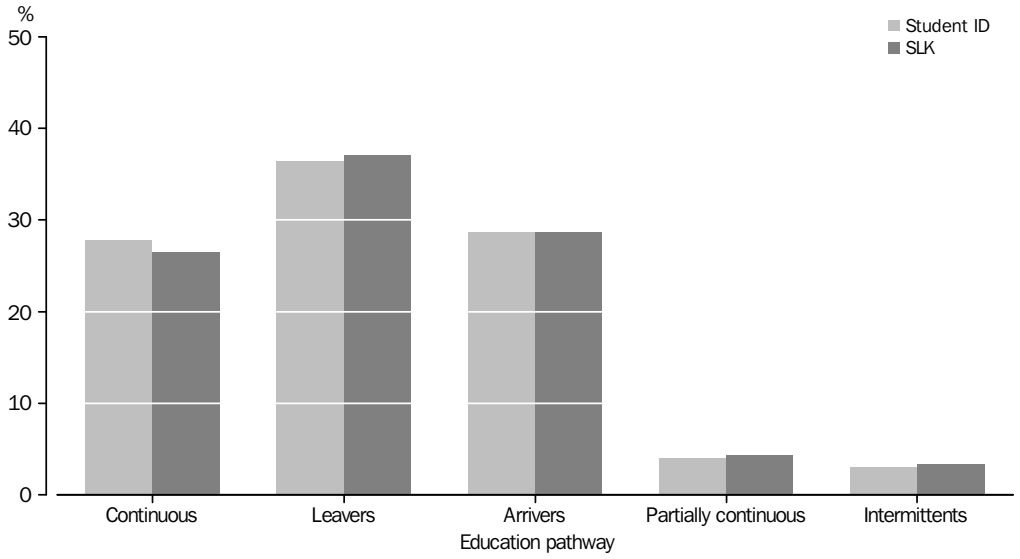
	Student distribution	
Cohort	*Student ID*	*SLK*
Continuous Students that were enrolled continuously from 2006 to 2011.	27.8%	26.5%
Leavers Students that were enrolled in 2006 and then left at some point before 2011. Includes students enrolled in only one year from 2007 to 2010.	36.4%	37.1%
Arrivers Students that arrived at some point after 2006 and remained until 2011.	28.7%	28.7%
Partially continuous Students that arrived after 2006 and left before 2011, but stayed continuously for more than one year.	4.0%	4.3%
Intermittents Irregular pathway patterns, constantly arriving, leaving and/or returning across 2006 to 2011.	3.0%	3.4%

(a) Polytechnic and Academy institutions use a different student administration system to the Tasmanian government schools. This has meant that some students could have been issued with a new Student ID when enrolled in these institutions, which would affect the linkage of these students across enrolment years when based on Student ID. The Tasmanian Department of Education has introduced processes to improve the consistent use and allocation of Student ID across the Tasmanian government education sector.

Figure 5.2 gives a graphical representation of the distribution of students across each education pathway for both linkage methods. The results demonstrate that the overall trends shown by the two linkage methods are very similar for all education pathways and that there is little difference between them at this level of disaggregation.

It is important to note that the term 'leavers' refers only to students that are leaving the Tasmanian government school system. These students may not in fact be leaving the school system altogether and could be transferring to a Tasmanian non-government school or moving interstate to complete their schooling. This is also true for the other transient education pathways. The inclusion of non-government school data in data integration projects would provide a better basis for classifying education pathways for students, which would in turn help to identify when students are leaving the school system or whether they are transferring to another sector.

5.2 Distribution of students by education pathway and linkage method, Tasmania, 2006–2011[(a)]



(a) Includes government school students only.

5.1 Match-link rates and link accuracy

The accuracy of a linking method can be more thoroughly evaluated by calculating the proportion of links in a given dataset that are matches (the link accuracy) and the proportion of possible matches that are actually linked in the dataset (the match-link rate).

The first step in calculating the match-link rate and the link accuracy of a dataset is to identify the match status and link status of the dataset, by comparing it with a benchmark method linked file. The Student ID linked file was the benchmark dataset for this analysis.

'Match status' is defined as the true status of a record pair. A match means that the two records belong to the same entity (i.e. the same student); whereas a non-match means that the two records belong to different entities (i.e. different students). 'Link status' is defined as the status assigned from a record linkage procedure, with record pairs assigned as links or non-links. Therefore, matches that are linked are called true links and non-matches that are linked are called false links. In the table below, total links includes the total of the true links and any false links. The total matches are the number of matches that should have been linked in the dataset. This total is obtained from the total matches found on the Student ID linked dataset.

It is important to note that students who did not have enrolment records in both datasets were not considered as possible matches, and are assigned a status of true non-matches. This includes students with an education pathway on the Student ID linked dataset of '100000', '010000', '001000', '000100', '000010' or '000001'. Table 5.3 compares the total links and true links achieved by using the SLK, with the total records matched using the Student ID (benchmark method).

5.3 True links by education pathway for the SLK linked dataset, Tasmania, 2006–2011[(a)]

Education pathway	*True links (SLK)*	*Total links (SLK)*	*Total matches (Student ID)*
	(no.)	(no.)	(no.)
Continuous	27,358	27,425	28,338
Leavers	21,662	22,654	22,247
Arrivers	20,606	21,795	21,312
Partially continuous	3,950	4,453	4,115
Intermittents	2,862	3,529	3,068
Total students	*76,438*	*79,856*	*79,080*

(a) Analysis does not include students with single year enrolments.

Linking using the SLK produced 79,856 links in total across the six years and 76,438 of these corresponded to benchmark method links (linking using the Student ID). The Student ID linked file produced 79,080 links, which are designated as total matches. There were more total links produced using the SLK linking method than total matches produced using the Student ID linking method because a number of student records linked incorrectly, which produced false links. This would be due to a coincidental matching of two SLKs that belong to two different students.

Once the match status and link status have been defined for a linked dataset, the match-link rate and link accuracy can be derived, and these are calculated below:

$$\text{Link accuracy} = \frac{\text{True links}}{\text{Total links}} = \frac{76,438}{79,856} = 95.7\%$$

$$\text{Match-link rate} = \frac{\text{True links}}{\text{Total matches}} = \frac{76,438}{79,080} = 96.7\%$$

The analysis shows that 95.7% of links in the SLK linked dataset were true matches. It also shows that 96.7% of possible matches were actually linked using the SLK method. This indicates that the SLK produces very high quality links and can be used successfully as an alternate linking methodology. This is particularly relevant given the number of linkages and the potential for reporting error that could occur over a six year period.

5.2 Replication of initial feasibility study tables using the SLK linking methodology

In order to obtain a more in-depth indication of the accuracy of the SLK linking method, more detailed analysis results have been produced for both SLK linked and Student ID linked datasets.

Unless otherwise stated, the following analysis of student characteristics is based on the characteristics of the student recorded at the main school of their last year of enrolment (most recently captured student characteristics).

The direct retention rate is an education statistic that can only be achieved when using a longitudinally linked student level dataset. The term 'direct' means that only students present in both cohorts (i.e. grade 7 and grade 9, for the grade 7–9 direct retention rate) are included in the resulting retention statistics. In comparison, apparent retention rates are calculated using the total count of full-time students at one enrolment year (e.g. grade 9 in 2011) and dividing by the total count of full-time students at another enrolment year (e.g. grade 7 in 2009). This method has many limitations, such as failing to take into account new or re-entry students, or students who have left the former cohort. Therefore, apparent retention rates often result in over 100%. The use of a longitudinally linked dataset where students can be linked over time allows a more accurate calculation of school retention.

Table 5.4 shows the percentage point difference between the direct retention rates when calculated after linking with the Student ID, compared with the rates calculated when linking with the SLK.

5.4 Differences between direct student retention rates using SLK and Student ID linkage methods, Tasmania, 2006–2011[(a)(b)(c)]

	SLK method minus Student ID method (percentage points)			
Grade range	*2008*	*2009*	*2010*	*2011*
		MALES		
Grade 7 – grade 9	–0.6	–0.5	–2.0	–1.3
Grade 7 – grade 10	n.a.	–0.8	–2.0	–2.1
Grade 10 – grade 12	–0.7	–0.4	–1.2	–0.5
Grade 7 – grade 12	n.a.	n.a.	n.a.	–0.5
		FEMALES		
Grade 7 – grade 9	–0.9	–0.6	–1.9	–1.6
Grade 7 – grade 10	n.a.	–1.6	–1.9	–1.5
Grade 10 – grade 12	–0.5	–0.4	–0.8	–0.9
Grade 7 – grade 12	n.a.	n.a.	n.a.	–1.3
		PERSONS		
Grade 7 – grade 9	–0.8	–0.6	–2.0	–1.5
Grade 7 – grade 10	n.a.	–1.2	–1.9	–1.8
Grade 10 – grade 12	–0.6	–0.4	–1.0	–0.7
Grade 7 – grade 12	n.a.	n.a.	n.a.	–0.9

(a) Grade 12 includes students in grade 12, grade 13 and ungraded secondary.
(b) Only includes students with a total full-time equivalent value of greater than or equal to 0.95.
(c) Applies the most recent sex recorded for the student.
n.a. Not available.

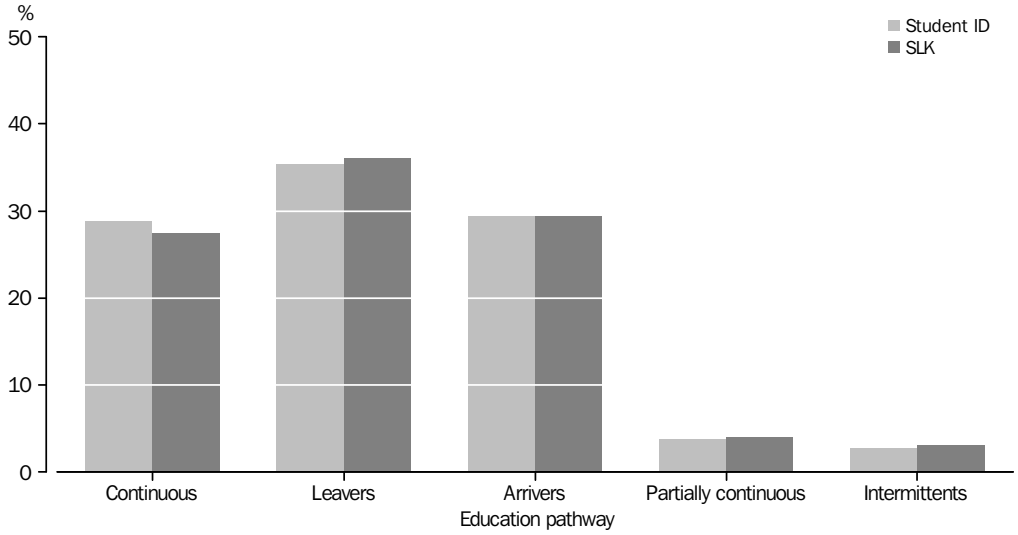
The size of the variation between the two methods is dependent on the grade cohort being analysed, however the overall results demonstrate that the SLK linking method is consistently under-reporting student retention in each grade range. This indicates that the SLK linking method was unable to link some students across years as successfully as the Student ID linking method. However, the results are not overly different and indicate that the SLK method can produce a good level of accuracy.

5.3 The impact of using different linkage methods to analyse typical sub-populations

The following analyses look at the outcomes achieved from using the different linkage methods to produce distributions of student sub-populations by each education pathway. The intention of this analysis is to demonstrate the effect of the different linkage methods on the resultant trends in the graphs.

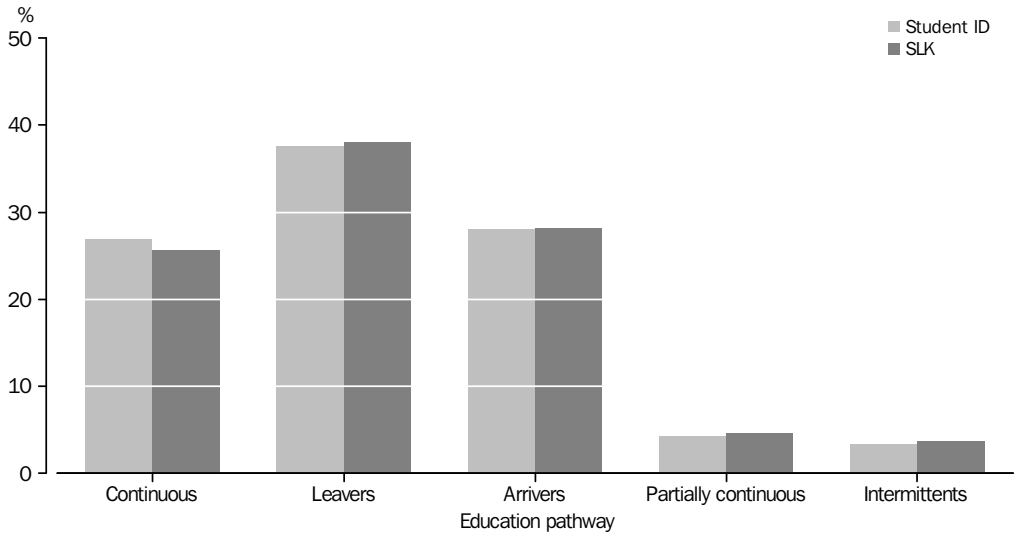
Figures 5.5 and 5.6 illustrate the effect of disaggregating the education pathways by sex. As expected for this sub-population, there were no particularly noticeable deviations from the original results presented in figure 5.2 for either male or female students when comparing the education pathways. Both linkage methods produced very similar results and were able to provide the same overall trend for both graphs. This means that the same conclusions could be drawn by using either linking method.

5.5 Distribution of male students, by education pathway and linkage method, Tasmania, 2006–2011[(a)(b)]



(a) Includes government school students only.
(b) Applies the sex captured on the student's most recent main school enrolment record.

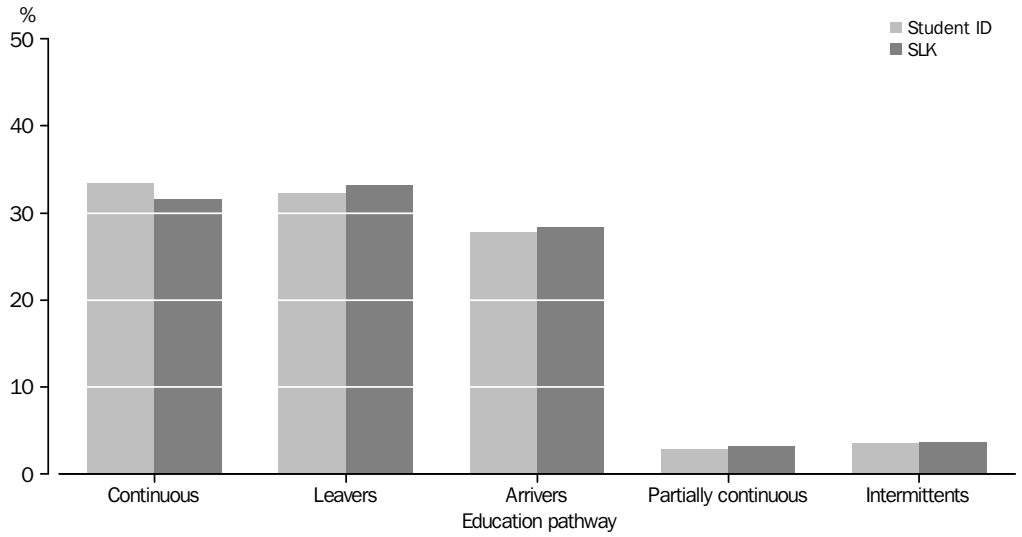
5.6 Distribution of female students, by education pathway and linkage method, Tasmania, 2006–2011[(a)(b)]



(a) Includes government school students only.
(b) Applies the sex captured on the student's most recent main school enrolment record.

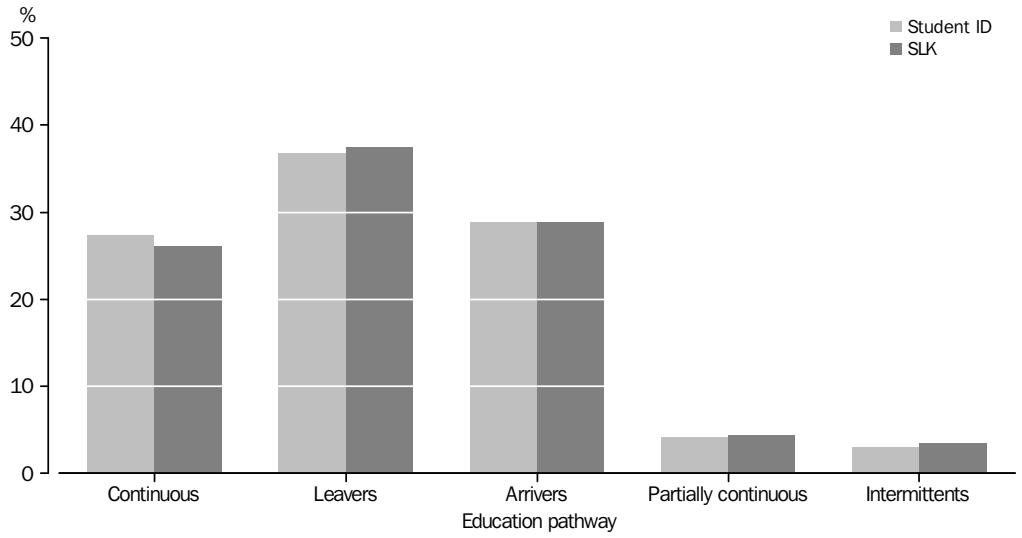
When the education pathway cohort categories are presented by Indigenous status, as shown in figures 5.7 and 5.8, both linking methods (Student ID and SLK) provide very similar findings. Therefore the same conclusions can be drawn from either linking method, with the actual proportions only differing slightly between the two methods.

5.7 Distribution of Aboriginal and Torres Strait Islander students, by education pathway and linkage method, Tasmania, 2006–2011[(a)(b)]



(a) Includes government school students only.
(b) Applies the Indigenous status captured on the student's most recent main school enrolment record.

5.8 Distribution of non-Indigenous students, by education pathway and linkage method, Tasmania, 2006–2011[(a)(b)]



(a) Includes government school students only.
(b) Applies the Indigenous status captured on the student's most recent main school enrolment record.

When looking at the difference between the SLK and the Student ID results for Aboriginal and Torres Strait Islander students, in figure 5.7, the Student ID results indicate that the majority of these students are categorised into the continuous cohort, followed closely by the leavers cohort. In contrast, the SLK results indicate that the majority of Aboriginal and Torres Strait Islander students are categorised into the leavers cohort, followed closely by the continuous cohort.

Figure 5.8 shows that when analysing the education pathways for non-Indigenous students, both Student ID and SLK resulted in the same pattern, with the majority of non-Indigenous students categorised into the leavers cohort, followed by the arrivers and continuous, for both linking methodologies.

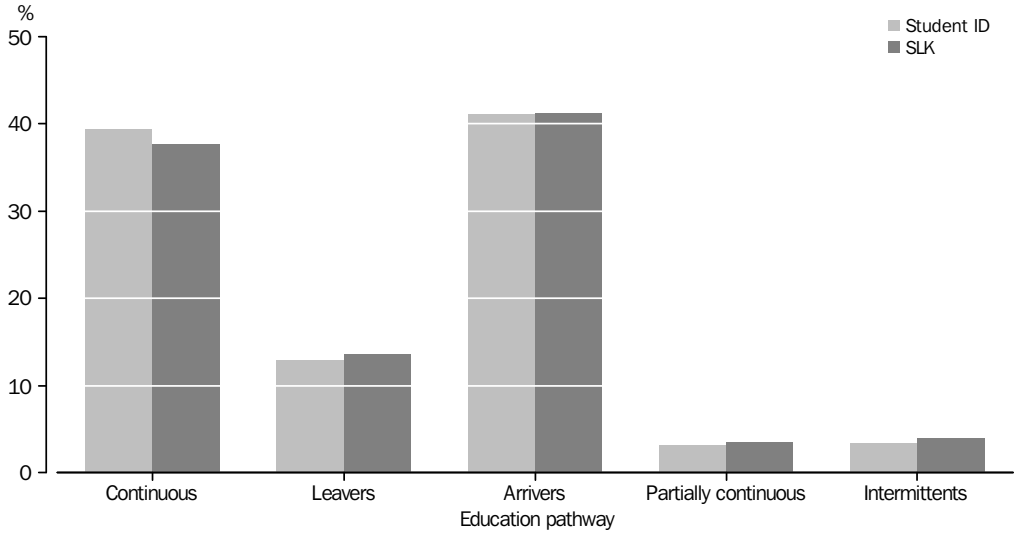
Figures 5.9 and 5.10 compare the distributions, using two different linkage methodologies, for students whose main language spoken at home is 'English' or 'not English'. The small proportion of 'leavers' in figures 5.9 and 5.10 is a significant difference from the overall trend in the distribution of the education pathways presented in figure 5.2. This is due to the fact that 'leavers', by definition, were not enrolled in 2011, and may not have been enrolled in 2009 or 2010 either. The difference partly reflects that the data were only collected from 2009 onwards, and should therefore be interpreted with caution. This highlights the importance of the continuing work towards improving the collection of NSSC data variables.

Figure 5.9 shows that students whose main language spoken at home is English had a higher proportion in the arrivers cohort, followed by the continuously enrolled cohort. In contrast, figure 5.10 shows that the majority of students whose main language spoken at home is not English are continuously enrolled, or are arrivers to the Tasmanian government school sector.

When comparing the two linkage methods and disaggregating by main language spoken at home, there is no significant difference between the results presented. Either could be used to draw the same analysis conclusions for this sub-population and the education pathways.

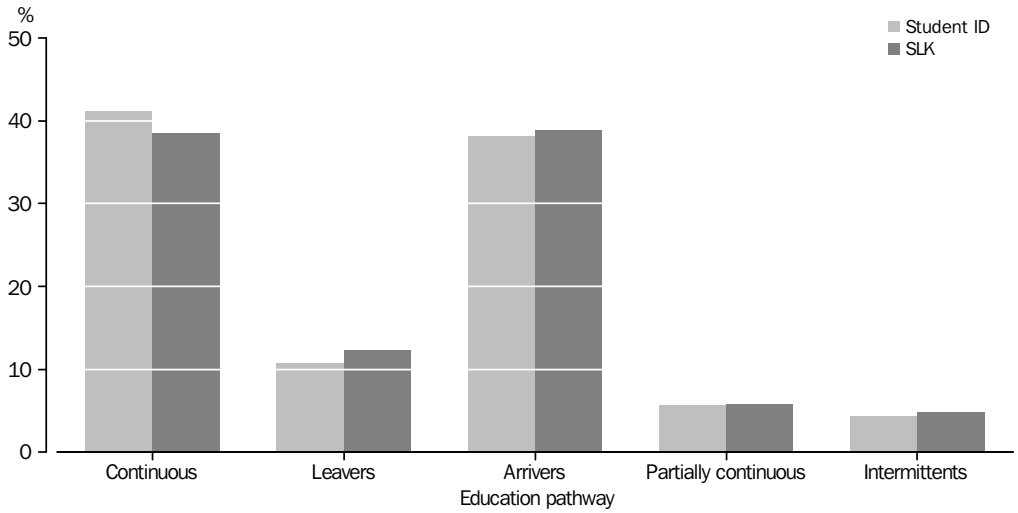
It is important to note that students with 'unknown' responses for the main language spoken at home variable were not included in these graphs. Full results for this analysis are available in tables D.5 and D.6 in Appendix D.

5.9 Distribution of students whose main language spoken at home is English, by education pathway and linkage method, Tasmania, 2009–2011[(a)(b)]



(a) Includes government school students only.
(b) Applies the Main language spoken at home captured on the student's most recent main school enrolment record.

5.10 Distribution of students whose main language spoken at home is not English, by education pathway and linkage method, Tasmania, 2009–2011[(a)(b)(c)]



(a) Includes government school students only.
(b) Applies the Main language spoken at home captured on the student's most recent main school enrolment record.
(c) Does not include not stated or missing dats.

6. CONCLUSION

This feasibility study has outlined two linkage methods that could be used for linking of education datasets. The study has confirmed the accuracy of the SLK linkage method for linking data longitudinally across six years of government school data. The SLK is an important linkage method as it allows de-identified data (data with no student name or address information) to be linked longitudinally, when a unique identifier such as the Student ID is not available.

The results of this analysis provide support that the SLK would be a high quality linking method for education data integration projects and would provide a feasible option when linking government school data to other data that does not have a consistent and unique Student ID, such as non-government school data and cross-jurisdictional data.

REFERENCES

Australian Bureau of Statistics (2013a) *NSSC Data Collection Manual (DCM)*, available on request from the ABS, Canberra.

—— (2013b) *Schools, Australia, 2012*, cat. no. 4221.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/4221.0/>

—— (2014) "Assessing the Quality of Different Data Linking Methodologies Across Time, Using Northern Territory Government School Enrolment Data", *Methodology Research Papers*, cat. no. 1351.0.55.046, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.046/>

Australian Curriculum, Assessment and Reporting Authority (ACARA) (2011) *National Assessment Program*, ACARA website content. <http://www.nap.edu.au/>

Australian Institute of Health and Welfare (2013) *Statistical Linkage Key 581 Cluster*, AIHW website content from METeOR Metadata Online Registry, AIHW, Canberra. <http://meteor.aihw.gov.au/content/index.phtml/itemId/349510>

All URLs viewed on 25 February 2014

APPENDIXES

A. EXPLANATORY NOTES

Prior to linking the files, preparation and cleaning of the datasets was undertaken to enable the best quality and accuracy in linking. Both of the linked datasets were created using the same files (2006–2009, 2010 and 2011) that were prepared by following the methods outlined below, which ensured that they were consistent. Therefore, the linked datasets can be compared against each other solely on the basis of their linking accuracy, without needing to take into account any other factors such as poor quality data or missing fields containing key linking variables.

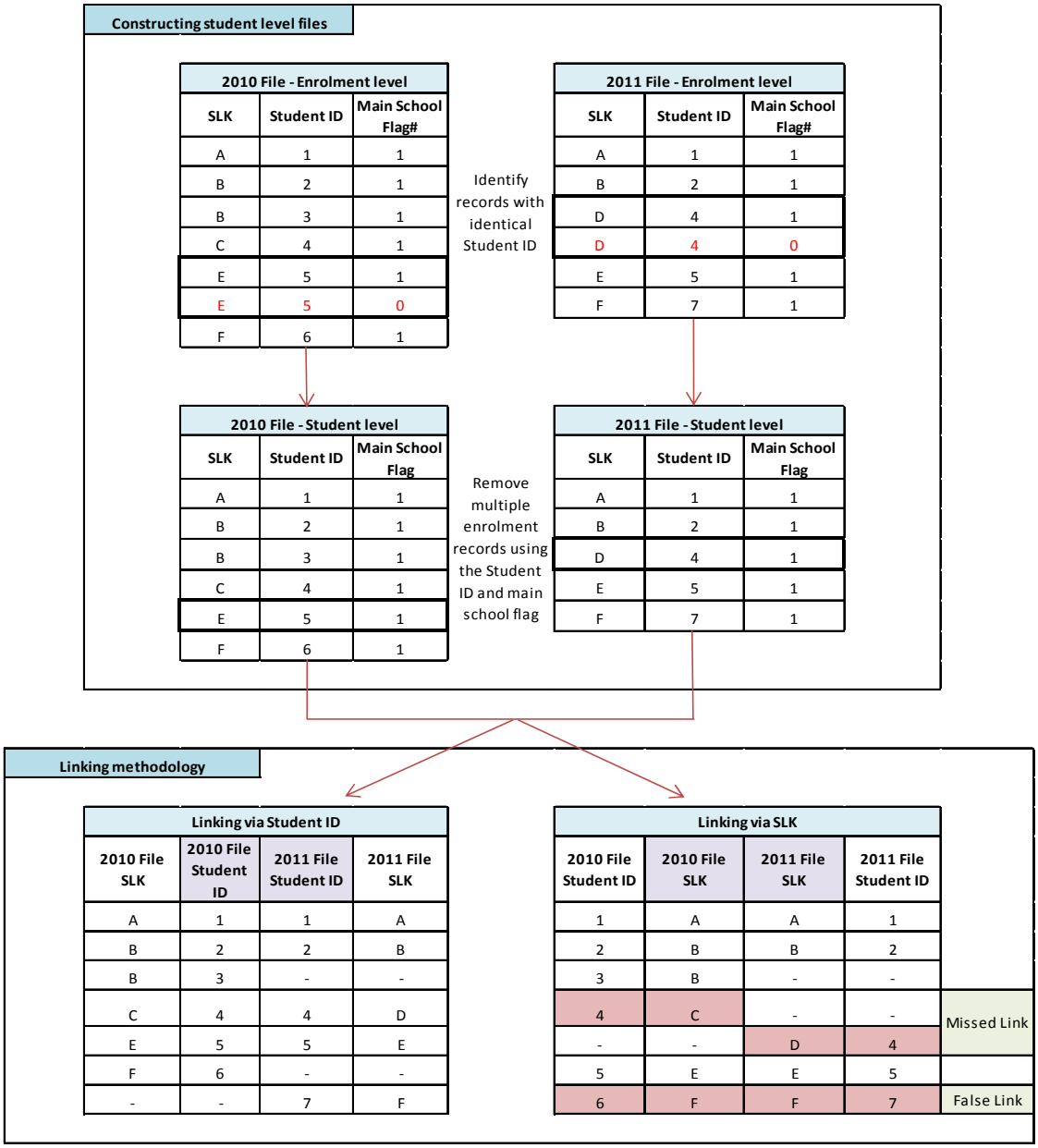
A.1 Creating a student level file

The Student ID was used to create a student level file. Figure A.1 demonstrates the process undertaken to create a student level file using the Student ID.

The Student ID may not always provide a unique identifier. Due to the Student ID being allocated at the school level, there may be instances where the Student ID is not unique to each student (e.g. through administrative error). A student may also be assigned more than one Student ID during the course of their schooling (e.g. where a student changes schools but their administrative information has not been accessed or found).

Despite this potential for error, for the purposes of this feasibility study the Student ID is considered to be of high enough quality to be used as a benchmark method against which the SLK linkage method could be compared. This is due to the overall quality of the Tasmanian Department of Education administrative system.

A.1 Process for construction and linking of files using a unique student identifier



See Section A.2 for information on how the main school flag was assigned for students with multiple enrolments.

A.2 Multiple enrolments and selecting student characteristics

Students may legitimately have more than one enrolment record within a given year for a number of reasons:

- many students legitimately enrol at multiple campuses/schools in order to complete courses that are not offered through their main school campus
- students who are being home-schooled are required in some states to enrol at a school campus, for the purpose of accessing resources or completing supervised examinations
- some students are highly mobile and may change schools without cancelling their prior enrolment
- some multiple enrolments may be due to clerical error or repair.

Student characteristics may differ across multiple enrolments, so it is important to select the appropriate student characteristic that matches the requirements of the particular research and analysis project being undertaken. It is important to note that there are many different algorithms that can be used to select individual student level characteristics when they differ across multiple enrolments for the one year, including selecting:

- the characteristics from the main school of enrolment (where the student spends the most number of hours enrolled at the school)
- the most recently captured characteristic
- the most commonly reported characteristics across the enrolments
- the characteristics based on random selection
- the characteristics based on attributes of the school enrolled (e.g. school size, school grade level, or school location).

For the purposes of this study, a main school flag was already applied to the datasets and therefore this variable was used to determine each student's main school enrolment.

Where student characteristics differed across years, the most recent characteristic was chosen for all analyses in this report. This was due to known data collection improvements in more recent years.

A.3 Assigning duplicate SLK flags

As it is possible and legitimate for more than one student to have identical SLKs, a duplicate flag was used to prevent records being linked more than once due to the linking process in the software.

After sorting the dataset by SLK, the flag was applied where a SLK occurred multiple times within a dataset (a single year). In most instances, the flag was assigned in the same order on all datasets, resulting in the records linking correctly. Due to sorting on one variable only, there were instances where the flag was assigned differently on the two files (see the example below), which resulted in a small number of false links. However, the use of the flag prevented unwanted replication of data and overall improved the link accuracy of the dataset. It was therefore considered appropriate and beneficial to include in the linking process.

A.2 Example of a false link using the SLK and the SLK duplicate flag

2010				2011		
Student ID	**SLK**	**SLK duplicate flag**		**Student ID**	**SLK**	**SLK duplicate flag**
ABC123	ABCDE121220032	1	Linked	XYZ789	ABCDE121220032	1
XYZ789	ABCDE121220032	2	Unlinked	–	–	–

A.4 Pre-existing data quality issues

As the NSSC data requirements have changed over the period from 2006 to 2011, some data items were not available for all six years of data. These data items included parental background information (such as occupation, school/non-school education) and main language spoken at home, which were only available on the 2009 to 2011 data files for students attending Tasmanian government schools. No parental information was provided from students enrolled in Polytechnic and Academy institutions. This information will be provided from 2012 onwards.

Data for student enrolments in Kindergarten grade level were only available on the 2006 to 2009 data files as these files were provided by the Tasmanian Department of Education via a separate data request. Data for the 2010 and 2011 years was attained from the NSSC submission, which does not include Kindergarten enrolment data. Therefore, all Kindergarten data was excluded from the analysis in this report.

It should be noted that the Tasmanian Polytechnic and Academy institutions use a different student administration system to the Tasmanian government schools. This has meant that some students could have been issued with a new student ID when enrolled in these institutions, which would affect the linkage of these students across enrolment years when based on Student ID. As a result, there is a chance that some students do not appear to have continued a Tasmanian government school education across 2006 to 2011, when they actually have, just under the assignment of a different Student ID. The Tasmanian Department of Education has introduced processes to improve the consistent use and allocation of the Student ID across the Tasmanian government education sector.

In terms of data quality, there were a number of Student IDs on the 2010 and 2011 datasets that were of an invalid length due to leading zeros being lost through data file conversion processes, with the appropriate length being 11. Student IDs with lengths of 7 were not amended as they contained letters and it was presumed that these were correct. However, Student IDs with a length of 9 or 10 were corrected by adding '0' or '00' to the beginning of these Student IDs to ensure that they were the correct length.

Linking between the datasets using the SLK linking method required a high quality SLK to be attached to each student record. The SLK needed to be created for the 2010 and 2011 datasets using the letters of surname, letters of first name, date of birth and sex variables. The SLK was already available on the 2006–2009 dataset however a number of these needed to be amended to remove dashes and hyphens.

B. KEY CONCEPTS AND DEFINITIONS

Main school enrolment

The main school enrolment for a student is the enrolment which is considered the primary enrolment for that student within a single year, and from which variables are used for analysis. See the *Explanatory Notes* section in this report for more information about a student's main school enrolment.

Education levels

The Tasmanian government school education levels are defined as:

- *Primary schooling:* Students enrolled in pre-year 1 through to grade 6.
- *Secondary schooling:* Students enrolled in grade 7 through to grade 10.
- *Senior secondary schooling:* Students enrolled in grade 11 and grade 12. This can include students in Year 13 and those classified in 'senior secondary other', which are students aged 21 years and over.

Education pathways

The educational pathways of all students can be categorised into the following continuous and transient student population cohorts:

- *Continuous:* Students that were enrolled continuously from 2006 to 2011.
- *Leavers:* Students that were enrolled in 2006 and left at some point before 2011. Includes students enrolled in only one year from 2007 to 2010.
- *Arrivers:* Students that arrived after 2006 and remained until 2011. Includes students that were only enrolled in 2011.
- *Partially continuous:* Students that arrived after 2006 and left before 2011, but stayed continuously for more than one year.
- *Intermittents:* Irregular pathway patterns, constantly arriving, leaving and/or returning across 2006 to 2011.

It is important to note that students can only be categorised into these pathways based on their government school enrolment information. It is likely that their education pathway would change with the inclusion of non-government school data and national data.

C. QUALITY ASSURANCE PROCESS

C.1 Counts of enrolments, students and students with multiple enrolments for Tasmanian government schools, by sex, 2006–2011[(a)]

	Number of enrolments	*Number of students*	*Number of students with multiple enrolments*
	No. of records on file	*No. of unique student identifiers*	
		2006	
Males	31,489	31,263	187
Females	31,193	30,511	508
Total	*62,682*	*61,774*	*695*
		2007	
Males	31,022	30,782	203
Females	30,370	29,770	463
Total	*61,392*	*60,552*	*666*
		2008	
Males	30,708	30,428	226
Females	29,897	29,360	431
Total	*60,605*	*59,788*	*657*
		2009	
Males	30,851	30,376	455
Females	29,809	29,103	647
Total	*60,660*	*59,479*	*1,102*
		2010	
Males	31,475	30,432	910
Females	30,521	29,056	1,187
Total	*61,996*	*59,488*	*2,097*
		2011	
Males	31,552	30,631	769
Females	29,945	28,905	863
Total	*61,497*	*59,536*	*1,632*

(a) Applies the sex captured on the student's most recent main school enrolment record.

C.2 Apparent retention rates, by sex, Tasmania, 2008–2011[(a)(b)]

	Apparent retention rates (%)			
Grade range	*2008*	*2009*	*2010*	*2011*
		MALES		
Grade 7 – grade 9	101.1	100.7	101.0	100.4
Grade 7 – grade 10	–	100.0	101.4	101.9
Grade 10 – grade 12	54.1	55.0	64.3	63.5
Grade 7 – grade 12	–	–	–	63.5
		FEMALES		
Grade 7 – grade 9	98.7	99.6	100.7	100.2
Grade 7 – grade 10	–	97.5	100.5	102.4
Grade 10 – grade 12	69.8	71.2	82.0	78.7
Grade 7 – grade 12	–	–	–	76.8
		PERSONS		
Grade 7 – grade 9	99.9	100.2	100.8	100.3
Grade 7 – grade 10	–	98.9	101.0	102.2
Grade 10 – grade 12	61.9	62.9	73.0	70.5
Grade 7 – grade 12	–	–	–	69.7

(a) Please refer to Section 5.2 for an explanation of retention rates.
(b) Applies the sex captured on the student's most recent main school enrolment record.

D. ACCOMPANYING TABLES

D.1 Distribution of students across each education pathway, by sex, using the Student ID method, Tasmania, 2006–2011[(a)(b)]

	Sex					
	Males		*Females*		*Total students*	
Education pathway	*no.*	*%*	*no.*	*%*	*no.*	*%*
Continuous	14,710	28.8	13,628	26.8	28,338	27.8
Leavers	18,024	35.3	19,047	37.5	37,071	36.4
Arrivers	14,992	29.4	14,228	28.0	29,220	28.7
Partially continuous	1,924	3.8	2,191	4.3	4,115	4.0
Intermittents	1,389	2.7	1,679	3.3	3,068	3.0
Total students	51,039	100.0	50,773	100.0	101,812	100.0

(a) Includes government school students only.
(b) Applies the sex captured on the student's most recent main school enrolment record only.

D.2 Distribution of students across each education pathway, by sex, using the SLK method, Tasmania, 2006–2011[(a)(b)]

	Sex					
	Males		*Females*		*Total students*	
Education pathway	*no.*	*%*	*no.*	*%*	*no.*	*%*
Continuous	14,256	27.4	13,169	25.5	27,425	26.5
Leavers	18,746	36.1	19,648	38.1	38,394	37.1
Arrivers	15,246	29.3	14,523	28.1	29,769	28.7
Partially continuous	2,087	4.0	2,366	4.6	4,453	4.3
Intermittents	1,618	3.1	1,911	3.7	3,529	3.4
Total students	*51,953*	*100.0*	*51,617*	*100.0*	*103,570*	*100.0*

(a) Includes government school students only.
(b) Applies the sex captured on the student's most recent main school enrolment record only.

D.3 Distribution of students across each education pathway, by Indigenous status, using the Student ID method, Tasmania, 2006–2011[(a)(b)]

	Indigenous status					
	Aboriginal or Torres Strait Islander		*Non-Indigenous*		*Total students*	
Education pathway	*no.*	*%*	*no.*	*%*	*no.*	*%*
Continuous	2,580	33.5	25,758	27.4	28,338	27.8
Leavers	2,491	32.3	34,580	36.8	37,071	36.4
Arrivers	2,141	27.8	27,079	28.8	29,220	28.7
Partially continuous	220	2.9	3,895	4.1	4,115	4.0
Intermittents	279	3.6	2,789	3.0	3,068	3.0
Total students	*7,711*	*100.0*	*94,101*	*100.0*	*101,812*	*100.0*

(a) Includes government school students only.
(b) Applies the Indigenous status captured on the student's most recent main school enrolment record only.

D.4 Distribution of students across each education pathway, by Indigenous status, using the SLK method, Tasmania, 2006–2011[(a)(b)]

	Indigenous status					
	Aboriginal or Torres Strait Islander		*Non-Indigenous*		*Total students*	
Education pathway	*no.*	*%*	*no.*	*%*	*no.*	*%*
Continuous	2,482	31.6	24,943	26.1	27,425	26.5
Leavers	2,609	33.2	35,785	37.4	38,394	37.1
Arrivers	2,228	28.3	27,541	28.8	29,769	28.7
Partially continuous	250	3.2	4,203	4.4	4,453	4.3
Intermittents	291	3.7	3,238	3.4	3,529	3.4
Total students	*7,860*	*100.0*	*95,710*	*100.0*	*103,570*	*100.0*

(a) Includes government school students only.
(b) Applies the Indigenous status captured on the student's most recent main school enrolment record only.

D.5 Distribution of students across each education pathway, by main language spoken at home, using the Student ID method, Tasmania, 2009–2011[(a)(b)]

	Main language spoken at home							
	English		*Other*		*Not stated*		*Total students*	
Education pathway	*no.*	*%*	*no.*	*%*	*no.*	*%*	*no.*	*%*
Continuous	25,814	39.5	2,483	41.2	41	0.8	28,338	36.9
Leavers	8,432	12.9	649	10.8	4,053	75.9	13,134	17.1
Arrivers	26,878	41.1	2,300	38.1	42	0.8	29,220	38.1
Partially continuous	2,057	3.1	342	5.7	794	14.9	3,193	4.2
Intermittents	2,236	3.4	258	4.3	410	7.7	2,904	3.8
Total students	*65,417*	*100.0*	*6,032*	*100.0*	*5,340*	*100.0*	*76,789*	*100.0*

(a) Includes government school students only.
(b) Applies the main language spoken at home captured on the student's most recent main school enrolment record only.

D.6 Distribution of students across each education pathway, by main language spoken at home, using the SLK method, Tasmania, 2009–2011[(a)(b)]

	Main language spoken at home							
	English		*Other*		*Not stated*		*Total students*	
Education pathway	*no.*	*%*	*no.*	*%*	*no.*	*%*	*no.*	*%*
Continuous	25,024	37.7	2,361	38.4	40	0.7	27,425	35.2
Leavers	9,031	13.6	753	12.2	4,143	75.9	13,927	17.9
Arrivers	27,343	41.2	2,385	38.8	41	0.8	29,769	38.2
Partially continuous	2,288	3.4	355	5.8	825	15.1	3,468	4.4
Intermittents	2,657	4.0	297	4.8	412	7.5	3,366	4.3
Total students	*66,343*	*100.0*	*6,151*	*100.0*	*5,461*	*100.0*	*77,955*	*100.0*

(a) Includes government school students only.
(b) Applies the main language spoken at home captured on the student's most recent main school enrolment record only.

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