Australian Bureau of Statistics

1351.0.55.046

Research Paper

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

www.abs.gov.au

New Issue

Research Paper

Assessing the Quality of Different Data Linking Methodologies Across Time, Using Northern Territory 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. 1351.0.55.046

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

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 NORTHERN TERRITORY GOVERNMENT SCHOOL ENROLMENT DATA

A Data Integration Feasibility Study

EXECUTIVE SUMMARY

Purpose

This feasibility study has assessed the suitability and quality of using different deterministic[1] linkage methods to link data without requiring personal characteristic information such as full name and address. The deterministic linking methodologies tested for this study included:

- linking with an unique student identifier (Student ID),
- linking with the Statistical Linkage Key 581 (SLK),
- linking with the SLK plus a geography variable (SLK+), and
- linking with a SLK that has been edited to improve its quality, plus a geography variable (SLK+ (edited)).

These linkage methods are explained further in Section 4 of this report.

Key findings

Both the Statistical Linkage Key (SLK) and the SLK+ geography (SLK+) linkage methods provided a very high match-link rate and link accuracy. SLK+ is the most likely method to be implemented for linking in data integration projects, as it has an extra linking variable that slightly improves the accuracy of the resultant dataset. If geography information is not available or not of a high enough quality, then the SLK method would still provide a satisfactory alternate linkage method to the SLK+ method.

Overall, while the SLK was considered to be of satisfactory quality, there is room to improve this data item for linkage purposes. The SLK+ (edited) method demonstrates the highly accurate linkage that can be achieved with a good quality SLK. The results for all analyses using the SLK+ (edited) method have shown that the quality of the data used to construct the linkage key directly influences the success of data linkage.

[1] 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.

An analysis of the integrated datasets undertaken for students that repeated a grade shows that all linking methods allow the same conclusions to be drawn for the analysis questions. While there were small differences between the results achieved for each method, these differences were small and only apparent at the more detailed disaggregations. For example, approximately 4.5% of Aboriginal and Torres Strait Islander students were repeating a grade in 2011, and across all four linkage methods about 57% to 57.7% of these repeating students were in the primary school grades.

Future data integration projects

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

1. Data integration authorities should note the importance of undertaking vigorous data quality assurance processes prior to linking. This is due to the positive effect that high quality data (particularly for constructing SLKs) can have on the resultant quality and accuracy of linking using deterministic linking methods.
2. Options for collecting and providing unit record level (URL) data from the non-government school sector should continue to be explored, since over a third of students attend non-government schools. The inclusion of non-government school data is vital to the creation of a complete picture of student education pathways and outcomes in Australia. For example, the analysis of repeating students may have unintentionally excluded students that were repeating grades, because their previous year of schooling was in a non-government school.
3. Estimates of linkage confidence or linkage error (i.e. the match-link rate and link accuracy) for linkages between various datasets should be made publicly available, to assist users of the linked datasets to evaluate the reliability of the linkage results. The most common quality measures used to assess the success of linkage methods include the calculation of link accuracy rates and match-link rates that directly compare the linked records to a benchmark linked file. In circumstances when a suitable benchmark file is not available for comparison, an iterative sample selection procedure can be used to estimate the false link rates.
4. Where student characteristics differ across multiple enrolments in the one year, where possible, outputs from data integration projects should include various characteristic data items that are based on different selection algorithms which would allow the researcher to decide which condition best suits the analysis task.

Future analysis ideas involving Northern Territory government school enrolment data

Due to time and resource restrictions, a number of additional analysis ideas could not be explored and are noted below in case this study is revisited, and/or for reference for future linkage studies.

1. Review the construction of the SLK in greater detail to determine the impact of:

   (a) when the geography variable in the SLK+ linkage method varies across enrolment records for highly mobile populations and over dispersed remote areas;

   (b) when date of birth on the enrolment form is an estimate or dump code (such as 1/1 or 1/7);

   (c) when students change their surnames; and

   (d) what component of the SLK changes most frequently for linked records with the same student id but different SLKs across enrolment years and/or multiple records.

   This review could potentially inform future improvements to the data cleaning process and data linkage activities.

2. Investigate missed and false links by particular characteristics to understand the impacts on under / over representation of sub-populations for the linked datasets.

3. Explore probabilistic[2] linking methods to determine if this linking methodology is more suitable compared to deterministic linking methods for small sub-populations who are more likely to be under / over represented, such as Aboriginal and Torres Strait Islander students, students enrolled in schools in very remote areas and students who mainly spoke a language other than English at home.

[2] 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.

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 Northern Territory Department of Education. Special thanks to Susan Kulda and Jeremy Howley.

This paper was prepared in the ABS National Centre for Education and Training Statistics. The principal authors and researchers were Megan Jobson and Katherine Thomson. 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 NORTHERN TERRITORY GOVERNMENT SCHOOL ENROLMENT DATA

A Data Integration Feasibility Study

ABSTRACT

This feasibility study assesses the suitability and quality of four different deterministic linkage methods, all of which link data without requiring full name and address.

Of the four linkage methods, both the Statistical Linkage Key (SLK) and the SLK+ geography (SLK+) linkage methods provided a very high match-link rate and link accuracy. SLK+ is the most likely method to be implemented for linking in data integration projects, as it has an additional linking variable that slightly improves the accuracy of the resultant dataset. A third linking method, the SLK+ (edited) method, demonstrates the highly accurate linkage that can be achieved with a good quality SLK. It also highlighted that the quality of the data used to construct the linkage key directly influences the success of data linkage.

The analysis undertaken using the integrated datasets for students that repeated a grade shows that all four linkage methods allow the same conclusions to be drawn for the analysis questions. While there were small differences between the results achieved for each method, these differences were very small and only noticeable at the more detailed disaggregations. For example, approximately 4.5% of Aboriginal and Torres Strait Islander students were repeating a grade in 2011, and across all four linkage methods about 57% to 57.7% of these repeating students were in the primary school grades.

This paper also details a number of considerations for improving the quality of future education and training data integration projects.

1. INTRODUCTION

This feasibility study evaluates the accuracy of different deterministic methods of linking student school enrolment records across years, without using name and address information. 'Gold standard' linking requires name, address, mesh block and other variables and is usually constructed in order to provide a benchmark dataset against which other methods can be compared. However, due to the need to protect privacy and confidentiality, name and address information is not usually available on datasets for research purposes.

This study analyses the quality of deterministic linkage methods using unique student identifiers and different types of statistical linkage keys. The results are compared to assist in determining which linking methods are of adequate quality to be used as linking solutions for data integration activities. This report also includes detailed analyses of sub-populations, including analysis of students that are continuously enrolled over a two-year period and those that are repeating a grade, to illustrate how analysis results might be affected by different deterministic linkage methods.

These linking methods represent the type of linkage that can be conducted in statistical studies to link educational administrative datasets with other administrative data sources, particularly when person level identifying characteristics are not available. It is therefore important that these methods are tested to ensure appropriate levels of quality in linkage projects.

The methods within this report are tested using Northern Territory Department of Education government school enrolment records for the 2010 and 2011 collection cycle years.

Due to time and resource constraints, probabilistic linking methods were not explored as part of this data linkage study. However, the Northern Territory Department of Education participated in the Census Data Enhancement (CDE) Education Quality Studies that linked 2010 and 2011 government school enrolment records together with the 2011 Census records using probabilistic linking methods. The results of this study are presented in the first CDE Education Quality Study publication, *Assessing the Quality of Linking School Enrolment Records to 2011 Census Data* (ABS, 2013).

2. BACKGROUND

The Northern Territory Department of Education provided the ABS with government school student level data files for the 2010 and 2011 school enrolment years. The data were sourced from the Northern Territory Department of Education under a data request for the purposes of the ABS data integration feasibility studies. The file structure was based on the National Schools Statistics Collection (NSSC) data specifications as outlined in the NSSC Data Collection Manual, which is available on request.

The files contained a unique student identifier (provided by the Northern Territory Department of Education), a student statistical linkage key (based on the SLK 581), student residential geographical information and school geographical information that were used for the purposes of linking.

Other data items on the files were provided as specified in the NSSC Data Collection Manual.

3. QUALITY ASSURANCE PROCESS

In 2011–12 a preliminary feasibility study was undertaken by the ABS to understand the characteristics of Northern Territory students enrolled in pre-Year 1 in 2011 at government schools and their preschool participation in 2010. The quality assurance included comparing the URL data supplied for that study to unconfidentialised pre-published National Schools Statistics Collection data. Comparisons included aggregated counts of schools, counts of students by grade level and by characteristics.

A selection of these tables were replicated for the current study, ensuring the same data was used for both studies. These tables are included in Appendix C of this report as tables C.1 and C.2 respectively. Comparisons to the data in table C.1 included aggregate counts of enrolments (records on file), counts of students (unique records on file) and counts of students with multiple enrolments. Table C.2 comparisons included counts of students by education level and numbers of students with multiple enrolments (including students with multiple enrolments across grade levels). The results from the quality assurance process showed matches for all comparisons and the data files were therefore considered appropriate for testing the linkage methodologies in this feasibility study.

4. LINKING METHODOLOGY

Linking between the 2010 and 2011 NSSC datasets was carried out by applying deterministic linking methods, which involves linking records from two different datasets that are exactly the same on both datasets.

One method of deterministic linking is to use a unique identifier that is common to both datasets. A link is 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 Northern Territory collects data for the NSSC via the Student Administration and Management System, which requires extensive validation and quality assurance checks on the data submitted from each school. Due to the nature of the Northern Territory administrative system, the project team considered the Student ID to provide a high quality benchmark against which the other linkage methods could be compared.

Prior to linking, the datasets were all prepared using the same process, which is described in Appendix A of this report. Following the data preparation process, just over 1% of records from both the 2010 and 2011 files were removed prior to linking, because they were identified as non-main school enrolment for students with more than one enrolment in a single year.

To clarify, each student remained represented on the data files. However, in order to reduce the complication of linking multiple enrolment records for one student in one enrolment year to a separate data file (considered enrolment level data files), a student level file was constructed where each record represented one individual student. When doing this, additional variables were created on the data files to capture the differing information from all enrolment records. New variables including number of enrolment records, address 1, address 2 etc, were created to ensure valuable information was not lost during the construction of a student level file. Retaining this information allows different engagement statistics to be calculated such as enrolment counts as opposed to student counts, as well as student workload to be calculated more accurately.

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

- linking with the Student ID (the benchmark method)
- linking with the SLK (SLK)
- linking with the SLK plus a geography variable (SLK+)
- linking with a SLK that has been edited to improve its quality, plus a geography variable (SLK+(edited)).

4.1 Student Identifier (ID)

The Student ID is a unique identifier assigned to each individual student enrolled in the Northern Territory government school education system, from pre-year 1 to senior secondary. The Student ID remains with the same student throughout their schooling in the Northern Territory government school system, even when moving between different Northern Territory government schools. A student enrolled in more than one Northern Territory 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 identical exact matches of a Student ID are found within a dataset, this would indicate that the student has multiple enrolments.

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

4.2 Statistical Linkage Key 581 (SLK)

SLK 581 is a type of statistical linkage key that was developed by the Australian Institute of Health and Welfare (AIHW, 2013). 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. a sex code (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). This is particularly relevant to

Aboriginal and Torres Strait Islander students and may lead to under representation. As the SLK is made up of components of student characteristics, the quality of the SLK is greatly 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 in the files relates to an individual student, given the preparation process undertaken on both files to remove multiple enrolments using the Student ID. Therefore, all SLKs on the files indicated unique students, despite the fact that there were many 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 small number of records were identified as linking incorrectly. Figure A.1 in Appendix A shows the application of the 'duplicate SLK flag'. Figure D.5 in Appendix D also shows how the files were constructed and how they were linked using SLK compared with linking using the Student ID. These linking errors are discussed in more detail in Section 5.2 of this report.

The use of a 'duplicate SLK flag' is common practice in some data linking exercises to improve the accuracy of the linking process. This linking methodology is used in the National Early Childhood Education and Care Collection to facilitate the correct handling of complex scenarios where several duplicate SLKs belong to two or more children. Such a flag could be provided on school enrolment data files where a suitable benchmark (such as the Student ID) is not provided.

4.3 SLK+ geography

SLK+ is a linking method that uses the SLK plus another variable to link the files. A number of possible variables were tested to use for the SLK+ technique, however the geography variables were found to be the most accurate when using Northern Territory NSSC data.

The SLK+ linking methodology involved two stages of linking. The first stage linked the 2010 and 2011 datasets by the SLK variable, the geography variable and a duplicate flag. A geography code was assigned to each student record using the rules outlined in Appendix A. The duplicate flag was applied to records to distinguish multiple occurrences of the same SLK and geography combination within a single dataset.

The second stage linked the files that were unlinked from the first stage. To do this, the unlinked records were extracted from the first stage file and linked using only the SLK and a duplicate SLK flag. The duplicate flag in this stage was assigned to multiple occurrences of the same SLK. Following the second stage of linking, the two resultant files were combined to form a single dataset.

4.4 SLK+ geography (edited)

While the SLK is a good non-unique linkage key, there are instances where data quality issues can lower the successfulness of the SLK as a linking variable and increase the likelihood that false matches or missed links will occur within a linked dataset.

To demonstrate the effect of having a high quality SLK used as a linking variable, the SLK+ (edited) methodology has been included in this study for testing. The SLK+ (edited) methodology involves updating two SLKs that conflict across years, so that they are recorded exactly the same on both datasets prior to linking. This ensures that the matching SLKs link across years, when they would not otherwise have linked if the SLKs had not been corrected.

For the 2010 dataset, the SLK was derived by the ABS from full name, date of birth and sex information. Therefore, the quality and accuracy of the SLK was high, though still dependant on the quality of the reporting on the dataset. The SLK for the 2011 dataset was already provided on the file and only the letters of the student's name that make up the component of the SLK were provided. This meant that there was no way to validate the accuracy of the SLK on the 2011 data file as full name was not available for comparison. Consequently, the SLK on the 2010 file was chosen as the benchmark SLK.

The SLK+ (edited) method first linked the 2010 and 2011 files together using the Student ID. This allowed determination of whether the SLKs matched across the files. For SLKs that were found to be mismatched for a record pair, the 2011 SLK was updated to match the 2010 SLK. 1,289 SLKs on the 2011 file were updated to match their record pair on the 2010 file. The dataset was then separated out into two individual files and linked in the same method as described above for the SLK+ linking method.

It is suggested that further analysis be conducted on the 1,289 records which had the 2011 SLK updated to match the 2010 SLK. This analysis could potentially inform improvements to data cleaning processes.

5. EVALUATION OF THE LINKAGE

The quality of a linked dataset can be evaluated in a number of ways. For this evaluation, the following points were considered:

- the quality of the Student ID linked dataset (i.e. the benchmark method)
- the match status and link status of the SLK, SLK+ and SLK+ (edited) linking methods compared with the benchmark method
- the under- or over-representation of sub-populations in the various linked datasets compared with the benchmark method
- the potential impact on analysis conclusions by using different linking methods.

It is important to note that the results presented in this report are based only on descriptive comparisons for indicative purposes, to demonstrate the analysis possible when linking NSSC data across years. No comparisons in this report are based on statistical significance. The results should be considered with care when drawing conclusions due to the small numbers of students involved in this study.

5.1 The quality of the benchmark method linked dataset

The aim of linking with Student ID was to provide a benchmark against which to test the accuracy and reliability of linking using the SLK and SLK+ methodologies.

Linking using the Student ID provided a comparable alternative to linking via student characteristic information, such as name and date of birth. This is because the Northern Territory Department of Education has a comprehensive student database in which all students enrolled in government schools are assigned their own individual Student ID. As mentioned previously, there are instances where the Student ID may not be unique to each student or where a student is assigned more than one Student ID during the course of their schooling, though this would only occur as a result of administrative error.

As discussed in Section 3, the enrolment files used for this data linking study were consistent with the data supplied to the ABS NCETS team for the purpose of the NSSC. This satisfied the quality assessment of the data files for this linkage study. Further analysis on a very similar Northern Territory Department of Education enrolment data file was undertaken as part of the CDE Education Quality Study. Please refer to Section 3.3 in the ABS publication, *Assessing the Quality of Linking School Enrolment Records to 2011 Census Data* (ABS, 2013) for analysis on the missing information on the Northern Territory enrolment data.

5.2 Match status and link status

For the purposes of this feasibility study, the dataset linked using Student ID was assumed to be linked at a 'gold standard' or the benchmark file. If the link status of the Student ID linked dataset is not an accurate reflection of the actual match status, then the rates of error for the SLK and SLK+ linked datasets will be biased. This must be taken into account when assessing the rates of accuracy outlined below.

The accuracy of a linking method can be 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 the benchmark method.

'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), an unmatch 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. It is important to note that students who did not have enrolment records in both datasets were not considered as possible matches, and were assigned a status of true non-links. The majority of these students would be the 2010 Year 12 leaving students and 2011 new students enrolled in kindergarten/pre-Year 1. Table 5.1 defines how to calculate the match status and link status for a linked dataset.

5.1 Method for calculating match status and link status for linked datasets

		Match status		
		Matches	Non-matches	Total
Link status	Links	True links *Matches that are linked* (n_{11})	False links (b) *Non-matches that are linked* (n_{12})	Total links $(n_{1.})$
	Non-links	Missed links (a) *Matches that are not linked* (n_{21})	True non-links *Non-matches that are not linked* (n_{22})	Total non-links $(n_{2.})$
	Total	Total matches $(n_{.1})$	Total non-matches $(n_{.2})$	Total record pairs $(n_{..})$

(a), (b) Figure D.5 in Appendix D shows a graphical example of a missed link and a false link.

Once match status and link status have been defined for a linked dataset, the match-link rate and link accuracy can be calculated. Table 5.2 outlines the calculation methods for the match-link rate and link accuracy measures, which are used to analyse the quality of a linked dataset.

5.2 Quality measures for linked datasets

Quality measure	*Definition*	*Formula*
Link accuracy	The proportion of assigned links in the linked dataset that are true matches.	$\text{Link accuracy} = \frac{\text{True links}}{\text{Total links}}$
Match-link rate	The proportion of possible matches that are actually assigned as links in the linked dataset.	$\text{Match-link rate} = \frac{\text{True links}}{\text{Total matches}}$

The match-link rate and link accuracy for each linked dataset is calculated below.

As described in Appendix A, students with sex or date of birth variables that were inconsistent across enrolment years were not linked using SLK or SLK+ methods. This is why the total number of record pairs displayed in the tables below are not the same for each linking methodology. Due to data discrepancies across enrolment records, some records were removed prior to linking as suitable SLKs could not be constructed.

5.2.1 Analysis of the Student ID linked dataset

The 2010 and 2011 Northern Territory government school datasets were linked using the Student ID. Table 5.3 shows the match and link status that resulted from the Student ID linkage methodology.

5.3 Student ID linkage – match status and link status

		Match status		
		Matches	Non-matches	Total
Link status	Links	22,364	N/A	22,364
	Non-links	N/A	12,973	12,973
	Total	*22,364*	*12,973*	*35,337*

5.2.2 Analysis of the SLK linked dataset

The 2010 and 2011 Northern Territory government school datasets were linked using the SLK variable. Table 5.4 shows the match and link status that resulted from the SLK linkage methodology.

5.4 SLK linkage – match status and link status

		Match status		
		Matches	Non-matches	Total
Link status	Links	21,053	208	21,261
	Non-links	1,311	12,589	13,900
	Total	*22,364*	*12,797*	*35,161*

Linking using the SLK produced 21,261 links in total and 21,053 of these corresponded to benchmark method links (linking using Student ID), where there were 22,364 links (designated as total matches).

Using the formulas outlined above in table 5.2, the match-link rate and link accuracy for the SLK linked dataset is calculated as follows:

$$\text{Link accuracy} = \frac{\text{True links}}{\text{Total links}} = \frac{21{,}053}{21{,}261} = 99.0\%$$

$$\text{Match-link rate} = \frac{\text{True links}}{\text{Total matches}} = \frac{21{,}053}{22{,}364} = 94.1\%$$

The analysis shows that of all the linked records in the SLK linked dataset, 99.0% are true matches. It also shows that 94.1% of all possible matches were actually linked using the SLK linking method.

5.2.3 Analysis of the SLK+ linked dataset

The datasets were also linked by the SLK+ method, which used a two-staged linking technique with the SLK and Geography variables. Table 5.5 shows the match and link status results for the SLK+ linking method.

5.5 SLK+ linkage – match status and link status

		Match status		
		Matches	Non-matches	Total
Link status	Links	21,068	193	21,261
	Non-links	1,296	12,602	13,898
	Total	*22,364*	*12,795*	*35,159*

Using this methodology there were 21,261 links in total and 21,068 of these corresponded to benchmark method links. This indicates a slight increase in the number of true links and a decrease in the total false links when compared with the SLK linking method, which indicates improved link accuracy.

$$\text{Link accuracy} = \frac{\text{True links}}{\text{Total links}} = \frac{21{,}068}{21{,}261} = 99.1\%$$

$$\text{Match-link rate} = \frac{\text{True links}}{\text{Total matches}} = \frac{21{,}068}{22{,}364} = 94.2\%$$

This linking method has produced a slightly improved result compared to the SLK linking method. A link accuracy of 99.1% and a match-link rate of 94.2% is a very high quality result. However, there were still links that were missed using this linking method.

5.2.4 Analysis of the SLK+ (edited) linked dataset

There were a number of SLKs on the 2010 file that were known to have matches on the 2011 file, that did not link due to quality issues with the SLK component variables. To demonstrate the potential of the SLK as a linking variable, SLKs that were deemed inaccurate on the 2010 file were edited to match the SLK of their record pair on the 2011 file. The match and link status for the SLK+ (edited) linkage methodology is outlined in table 5.6.

5.6 SLK+ (edited) linkage – match status and link status

		Match status		
		Matches	Non-matches	Total
Link status	Links	22,355	188	22,543
	Non-links	9	12,606	12,615
	Total	*22,364*	*12,794*	*35,158*

Using this methodology there were 22,543 total links and 22,355 of these corresponded to benchmark method links. The match-link rate and link accuracy for the SLK+ (edited) linked dataset are as follows:

$$\text{Link accuracy} = \frac{\text{True links}}{\text{Total links}} = \frac{22{,}355}{22{,}543} = 99.2\%$$

$$\text{Match-link rate} = \frac{\text{True links}}{\text{Total matches}} = \frac{22{,}355}{22{,}364} = 100.0\%$$

It is interesting to note from the analysis that editing the SLKs does not greatly improve the link accuracy of the SLK+ linking method, demonstrating the true effect of false links caused by students with identical SLKs. However, what is achieved is a reduction in the number of missed links, which thereby improves the match-link rate to just under 100%.

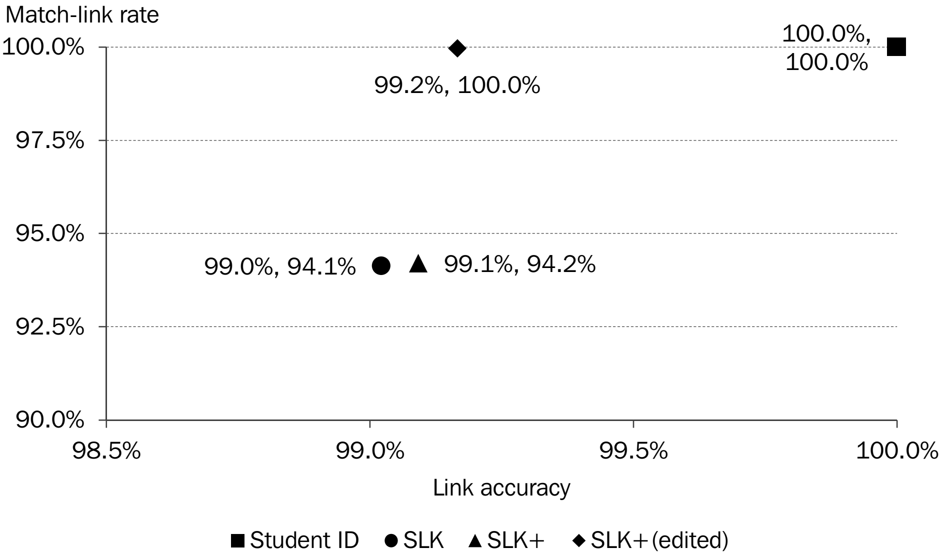
This means that of all the records that should have linked, almost all of them did link. The small number of remaining missed links can be attributed to the records that linked incorrectly by using the duplicate SLK flag. It should be noted that the benefit of using the duplicate SLK flag to improve link accuracy, especially for the SLK+ method, outways the very small number of missed links observed. Figure A.1 in Appendix A shows the application of the duplicate SLK flag.

5.2.5 *Comparison of match-link and link accuracy*

The results for the SLK, SLK+ and SLK+ (edited) linking methods are shown in figure 5.7. When compared with the SLK method, the SLK+ method demonstrated a slight improvement in link accuracy as the number of variables used for linking is increased. However, with the SLK+ (edited) method, the match-link rate and the link accuracy were both of a high standard.

This analysis shows that the SLK is a very effective linking variable, provided the data used in its construction are of a high quality. The analysis also demonstrates that the geography variable, when of a high quality, can assist in producing accurate links.

5.7 Match-link rate and link accuracy for each linkage method compared with the benchmark method (Student ID), Northern Territory, 2010–2011



While assessing each dataset's match and link status shows the accuracy of the linkage methods, it is also important to undertake analysis on the missed links and false links that occurred within each dataset. This will assist in determining caveats around analysis of social or demographic characteristics, particularly for those records with higher rates of missing links or false links.

5.2.6 *Analysis of missed links*

Missed links include those records that did not link with a record on the other dataset (SLKs did not match), but by comparing the links with the benchmark method, it is evident that these records in fact should have linked (Student IDs actually matched). A record can only be classified as a missed link if it is present on both the 2010 and 2011 datasets and it did not link during the linking process.

Table 5.8 is derived by comparing the number of missed links obtained from a linking method, with the total number of links from the benchmark method dataset (the total matches), for selected characteristics.

For both linking methodologies, ungraded secondary students resulted in a higher percentage of missed links. This may be because these students generally had limited information available on their records, which contributed to a high number of these record pairs being missed. Aboriginal and Torres Strait Islander students, students enrolled in schools located in very remote localities and students who mainly spoke a language other than English at home, are other sub-populations that had a higher percentage of missed links.

5.8 Percentage of missed links (b) using SLK, SLK+ and SLK+ (edited), by selected characteristics, Northern Territory, 2011

	Linking method		
Characteristic	*SLK*	*SLK+*	*SLK+ (edited)*
Sex (%)			
Female	5.8	5.8	0.1
Male	5.9	5.8	0.0
Indigenous status (%)			
Aboriginal or Torres Strait Islander	8.2	8.1	0.1
Non-Indigenous (a)	3.9	3.9	0.0
Education level (%)			
Primary	5.7	5.6	0.1
Secondary	6.6	6.5	0.0
Senior secondary	4.4	4.4	0.0
Ungraded secondary	21.2	21.2	0.0
Remoteness of school (%)			
Outer regional Australia	4.6	4.5	0.0
Remote Australia	4.5	4.5	0.0
Very remote Australia	9.1	8.9	0.0
Main language spoken at home (%)			
English	3.4	3.4	0.0
Other language	7.6	7.5	0.1
Total missed links (%)			
Total	*5.9*	*5.8*	*0.0*

(a) Includes responses of not stated.
(b) Compared with potential links when linking by Student ID (benchmark method).

Overall, the rate of missed links for both methods was not high, with fewer than 6% of links missed on either dataset.

Given the high proportion of missed links for the ungraded secondary education level, it is suggested that further analysis on the construction of the SLK for this sub-population be explored. Particular changes to the SLK components may be the cause of the high number of missed links.

5.2.7 Analysis of false links

False links occur when records relating to two different students are linked in error (based on matching SLKs). They are identified by comparing the linked records being assessed (e.g. SLK, SLK+ and SLK+ edited) to those in the benchmark method (Student ID) linked file.

Table 5.9 compares the number of false links obtained from a linking method, against the total number of links from the benchmark method dataset (the total links), for selected characteristics.

5.9 Percentage of false links[b] using SLK, SLK+ and SLK+ (edited), by selected characteristics, Northern Territory, 2011

	Linking method		
Characteristic	*SLK*	*SLK+*	*SLK+ (edited)*
Sex (%)			
Female	0.9	0.9	0.9
Male	0.9	0.8	0.8
Indigenous status (%)			
Aboriginal or Torres Strait Islander	1.5	1.4	1.3
Non-Indigenous (a)	0.5	0.4	0.4
Education level (%)			
Primary	1.0	0.9	0.9
Secondary	0.8	0.8	0.8
Senior secondary	0.8	0.8	0.7
Remoteness of school (%)			
Outer regional Australia	0.6	0.6	0.6
Remote Australia	1.0	1.0	1.0
Very remote Australia	1.6	1.4	1.3
Main language spoken at home (%)			
English	0.6	0.6	0.6
Other language	1.2	1.1	1.0
Total false links (%)			
Total	*0.9*	*0.9*	*0.8*

(a) Includes responses of not stated.
(b) Compared with potential links when linking by Student ID (benchmark method).

This table demonstrates that the proportion of false links for all of the linking methods is very low when compared with the number of links achieved by the benchmark method. False links appear to be slightly more prominent for Aboriginal and Torres Strait Islander students, students enrolled in schools in very remote localities, and students who mainly spoke a language other than English at home. However, due to the low numbers of false links, caution should be exercised when analysing these results.

While it is useful to look at match-link rate and link accuracy alone, it is also worthwhile to look at the potential under- or over-representation of sub-populations within each of the linked datasets.

5.3 Analysis of linkage: the under- or over-representation of sub-populations

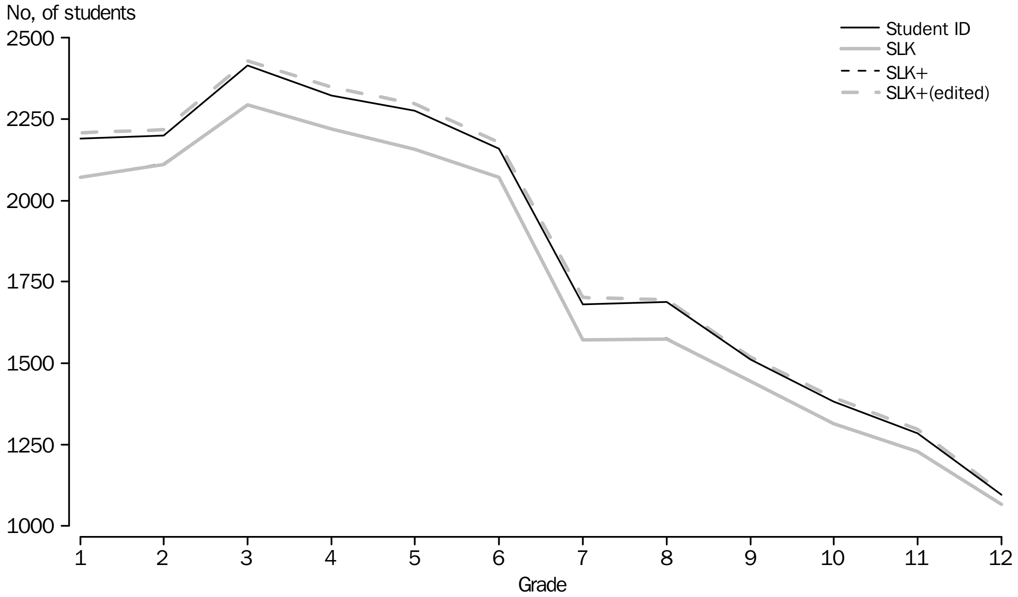
The following analyses compares the linked datasets to see how different criteria used in linking can affect the distribution of various sub-populations. In essence, this provides an analysis of the continuously enrolled population of students who have Northern Territory government school enrolments in both 2010 and 2011.

Analysis of student characteristics on the linked datasets was undertaken and the relative distributions and frequencies compared to assess the quality of each linking method. The analysis indicated that no sub-populations were highly affected by the different linking methods. However, some populations do seem to be more difficult to link resulting in some degree of under-representation on the linked files.

Unless otherwise stated, the following analysis of student characteristics is based on the characteristics as reported for the 2011 student enrolment. This attempts to capture the most recent student characteristics for analysis. It is also important to note that the majority of continuous students enrolled in pre-year 1 are students that are repeating this grade or re-enrolled due to age entry restrictions (four year old students re-enrolled in the five year old pre-Year 1 (kindergarten) program), therefore the numbers for pre-year 1 are very small and have been excluded from the analysis. Ungraded secondary students are also excluded from the analysis due to insufficient data.

Figure 5.10 shows the distributions achieved for each linking method, for the number of continuous students by grade. The SLK and the SLK+ linked files follow the same distribution line and are both under-representing continuous students for all grades when compared with the benchmark method. This indicates that the observed relationship is slightly different for these datasets. It also suggests a reduced number of linked record pairs spread evenly across all grades in the dataset. The level of under-representation slightly decreases towards the higher grades indicating that students in these grades were more accurately linked.

5.10 The number of continuous students for each linking method, by grade, Northern Territory, 2011



The SLK+ (edited) method produced a very similar distribution to the benchmark method. However, it is slightly over-representing all grades, which is more noticeable for the lower grades. This is consistent with having a higher number of false links, coupled with a low number of missed links. This can be compared to the SLK and SLK+ methods that had similar numbers of false link numbers but a much higher level of missed links, indicating a lower number of total links achieved. Despite this, all linkage methods follow a very similar trend line as the Student ID method, which suggests that the overall proportions would produce very similar results for all linking methods.

The following tables show the relative frequencies for selected sub-populations. This demonstrates the over- or under-representation of these populations for each linking method as a proportion of the total.

5.11 Relative frequencies of continuous students, by Indigenous status and linkage methodology, Northern Territory, 2011

	Linking method			
	Student ID	*SLK*	*SLK+*	*SLK+(edited)*
	%	%	%	%
Indigenous status				
Aboriginal or Torres Strait Islander	44.8	44.0	44.0	45.0
Non-Indigenous	55.2	56.0	56.0	55.0

5.12 Relative frequencies of continuous students, by remoteness of school and linkage methodology, Northern Territory, 2011

	Linking method			
	Student ID	*SLK*	*SLK+*	*SLK+(edited)*
	%	%	%	%
Remoteness of school				
Outer regional	54.6	55.2	55.2	54.5
Remote	16.7	16.9	16.9	16.7
Very remote	28.7	27.9	27.9	28.8

5.13 Relative frequencies of continuous students, by main language spoken at home and linkage methodology, Northern Territory, 2011

	Linking method			
	Student ID	*SLK*	*SLK+*	*SLK+(edited)*
	%	%	%	%
Main language spoken at home				
English	41.5	42.5	42.5	41.4
Language other than English / Not stated	58.5	57.5	57.5	58.6

Overall, all methods produced very similar results. Some of the sub-populations that are slightly under-represented on the SLK and SLK+ linked datasets were Aboriginal and Torres Strait Islander students, students enrolled in schools in very remote areas, and students who mainly spoke a language other than English at home. Interestingly, the SLK+ (edited) method tended to slightly over-represent these sub-populations, whereas the alternate sub-populations, such as non-Indigenous students, were under-represented. However, the SLK+ (edited) method produced a much more consistent frequency to the Student ID method and the rate of error is relatively small (approximately 1%) for each sub-population.

5.4 Typical analyses and the impact on conclusions by using different linkage methods

While it is important to consider the effect the linkage methods have on the outcomes for different sub-groups in a population, it is also interesting to see what effect discrepancies in representation and the quality measures (link accuracy and match-link rate) are likely to have on typical analyses.

The following analysis focuses on students that are repeating grades, which is a cohort of students that can be extracted from a dataset that is linked across at least two years. Students repeating grades are defined as those students enrolled in 2011 that are repeating the grade that they were enrolled in for 2010.

Potential research questions about students repeating grades that could be answered with an integrated dataset of longitudinal school enrolment data, and that are analysed within this report, include:

1. Which school grades are the most frequently repeated?
2. Are there higher rates of repeating for students with certain characteristics, such as sex, Indigenous status or remoteness of school?

This analysis will also focus on identifying the impact on the quality of the linkage for sub-populations of students that are repeating a grade in 2011. The analysis relates to the characteristics as reported for the student's 2011 enrolment.

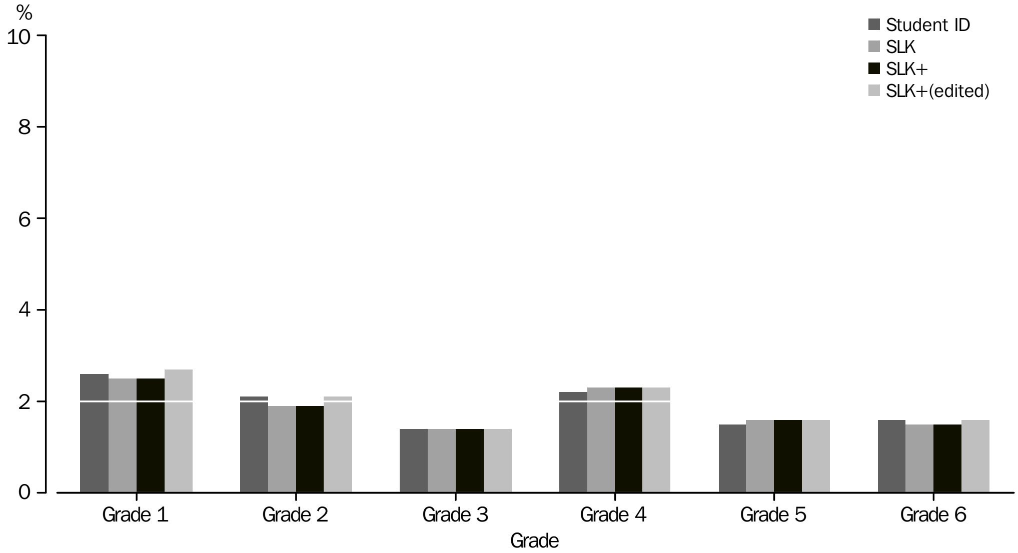
It is important to note that the analysis below excludes any students that are repeating a grade in 2011 and were not in the Northern Territory government school system in the previous year (2010). These students are not included in the counts of students repeating grades or the counts of continuous students. The analysis on students repeating grades also excludes students listed as being in ungraded secondary, as the data does not show whether these students are actually repeating a grade.

Overall, approximately 3% of all continuous students in Northern Territory government schools for 2010 and 2011 repeated a grade in 2011[3]. Figures 5.14 and 5.15 show the proportion of continuous students who were repeating a grade in 2011, for primary and secondary students. The graphs show that the grades most likely to be repeated by students are the higher secondary school years (grades 10, 11 and 12), with over 8% of continuous students repeating grade 12. It should be acknowledged that many senior year students choose to complete Year 12 part-time over two years, so these students are not necessarily repeating but are part-time completing the Year 12 workload over two years. The results for the lower grades tend to be more consistent, with approximately 2% of students repeating each grade.

The graphs also show the relationship between the different linkage methods. There are slight over- and under-representations between each linking method when compared to the benchmark Student ID method. However, the differences between the proportions are very small.

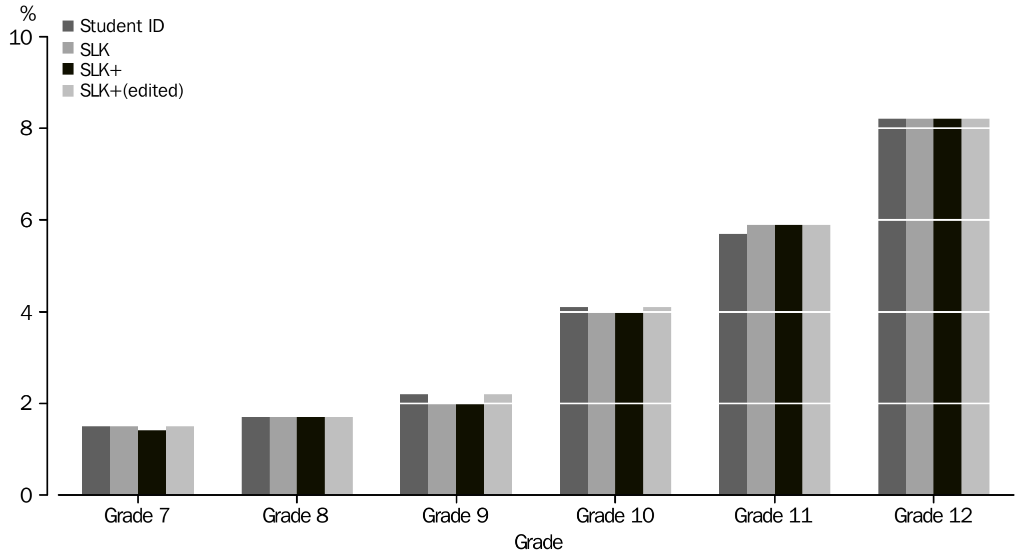
[3] Calculated using the Student ID linked dataset.

5.14 Percentage of continuous[(a)] primary school students who repeated a grade, by grade and linkage methodology, Northern Territory, 2011



(a) Student enrolled in 2010 and 2011.

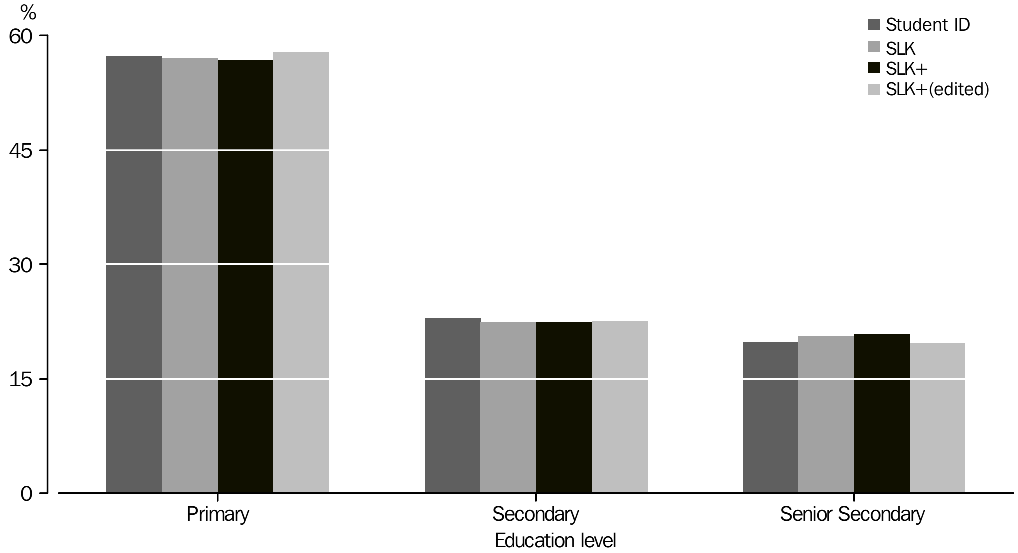
5.15 Percentage of continuous[(a)] secondary school students who repeated a grade, by grade and linkage methodology, Northern Territory, 2011



(a) Student enrolled in 2010 and 2011.

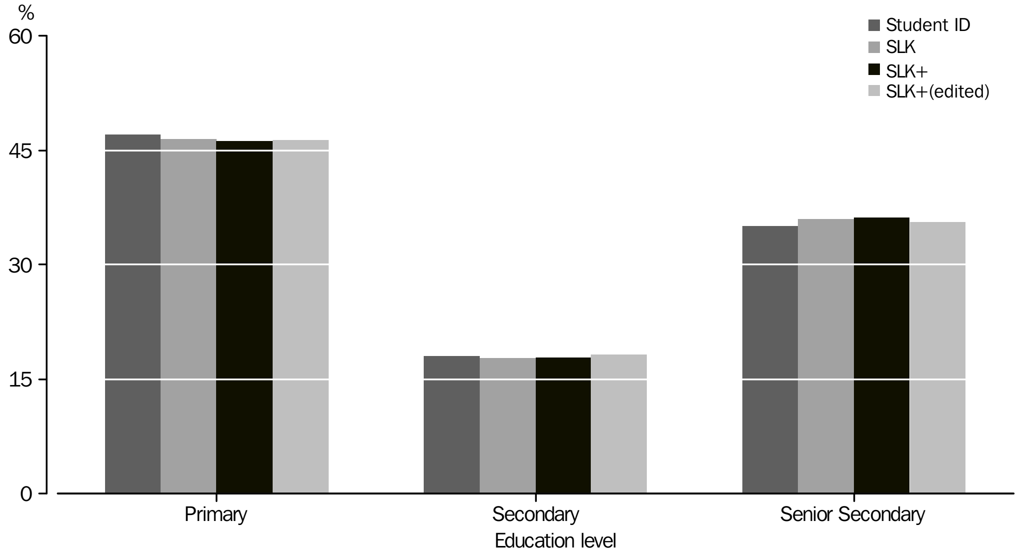
Figure 5.16 summarises the education levels of Aboriginal and Torres Strait Islander students who repeated a grade in 2011, as derived from the four linkages. Figure 5.17 provides the same summary for other Australian students. In 2011, approximately 4.5% of Aboriginal and Torres Strait Islander students were repeating a grade, and about 57% of these repeating students were in the primary school grades. Approximately 1.6% of other Australian students were repeating a grade in 2011, with the majority (about 54%) repeating secondary school grades.

5.16 Education level of Aboriginal and Torres Strait Islander students who repeated a grade[(a)], by linkage methodology, Northern Territory, 2011



(a) Does not include ungraded secondary students, as they cannot be confirmed as repeating a grade.

5.17 Education level of non-Indigenous students who repeated a grade[(a)], by linkage methodology, Northern Territory, 2011

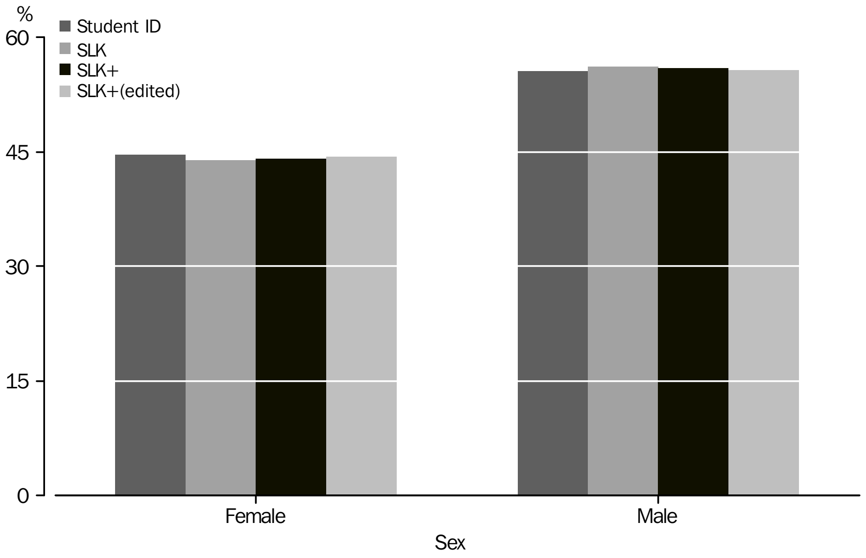


(a) Does not include ungraded secondary students, as they cannot be confirmed as repeating a grade.

When looking at the difference between the four linkage methods, the SLK and SLK+ linkage methods are both slightly under-representing Aboriginal and Torres Strait Islander students and over-representing other Australian students when compared with the Student ID method. The SLK+ (edited) method provides a similar proportion to the Student ID method.

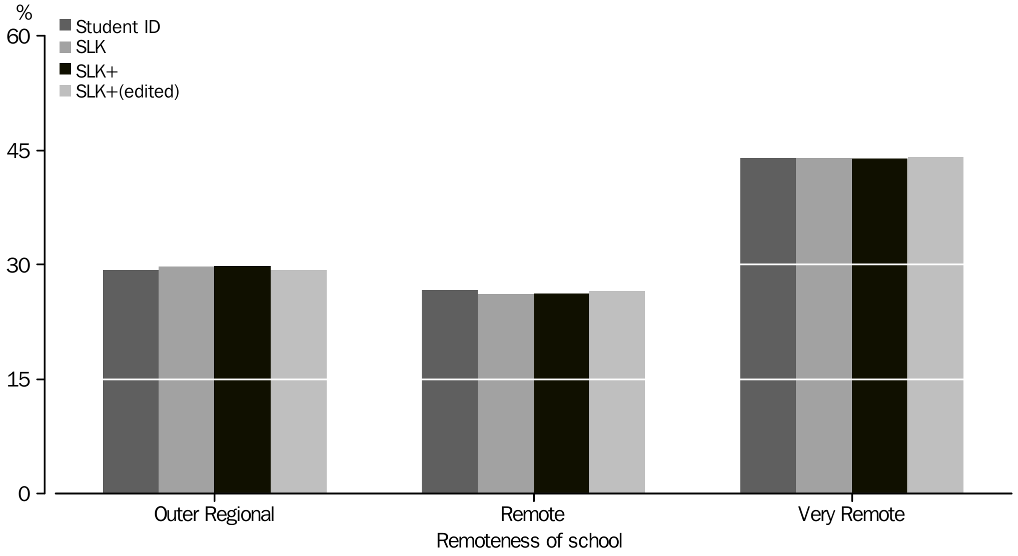
While there is a slight difference between the results for each linkage method, the differences are not substantial. The proportions displayed are very similar and all methods are able to demonstrate to the same degree that Aboriginal and Torres Strait Islander students are repeating grades more than other Australian students.

5.18 Sex distribution of students who repeated a grade[(a)], by linkage methodology, Northern Territory, 2011



(a) Does not include ungraded secondary students, as they cannot be confirmed as repeating a grade.

5.19 Remoteness of school attended by students who repeated a grade[(a)], by linkage methodology, Northern Territory, 2011



(a) Does not include ungraded secondary students, as they cannot be confirmed as repeating a grade.

Figure 5.18 shows the difference in students repeating a grade for male students compared with female students. The results show that a higher percentage of male students (55.5%) repeated a school grade in 2011 than female students (44.5%).

In terms of the differences between the linking methodologies when producing these analysis results, it is interesting to note that all methods are slightly under-representing female students and over-representing male students. However, overall the same conclusions can be drawn from this data no matter which linking method is used.

Figure 5.19 highlights the proportion of repeating students for each remoteness classification in the Northern Territory. The results show that of all students enrolled in schools in very remote areas, approximately 44% repeated a grade in 2011. Interestingly, students enrolled in schools in remote localities have a slightly lower rate of repeating a grade (26.7%) compared to those in outer regional schools (29.3%).

When analysing remoteness of school data, the results achieved for the Student ID dataset are very close to the other linkage methods and therefore unlikely to adversely affect the analysis of these sub-populations. The research questions could be answered quite accurately using any of the linkage methods.

REFERENCES

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

—— (2012) *Information Paper: Converting Data to the Australian Statistical Geography Standard*, cat. no. 1216.0.55.004, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1216.0.55.004>

—— (2013) "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>

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. All linked datasets (Student ID, SLK, SLK+ and SLK+ (edited)) were created using the same 2010 and 2011 files 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 prior to linking

The Student ID was used to create a student level file. 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 is not known or accessible).

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 other linkage methods could be compared.

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 record 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 (the school where the student has the greatest workload)
- selecting the most commonly reported characteristics across the enrolments
- selecting the characteristics based on random selection
- selecting 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, the following criteria were applied to both the 2010 and 2011 datasets, to identify the main school enrolment for students who had multiple enrolments within the same year:

(a) Identify students with multiple enrolment records using the Student ID data item.

(b) Select the enrolment record with the highest full-time equivalent (FTE) value.

(c) If the FTE is the same across the enrolment records, select the enrolment record with the *lowest grade*.

(d) If the FTE and grade are the same across the enrolment records, select the enrolment record with the *lowest age*.

(e) If FTE, grade and age are all the same across the enrolment records, *randomly* select an enrolment record.

This criterion is a similar procedure to that used for the NSSC to construct student level data and select the student characteristics to be used for analysis and reporting. The construction of student level data files enables the count of students as opposed to counts of enrolments to be calculated. The criteria is essentially a pragmatic way to identify the "main" enrolment for those students enrolled across multiple schools. As previously explained in Section 4, new variables were created on the data files to capture the differing information from all enrolment records. New variables including number of enrolment records, address 1, address 2 etc, were created to ensure valuable information was not lost during the construction of a student level file.

The random selection of an enrolment record (criteria 'e' above) is simply one algorithm that can be used to select which student characteristics to use for analysis and reporting when the previous selection steps (criteria 'a' to 'd') are not satisfied. Retaining the information from all enrolment records by creating new variables allows researchers to use different algorithms to select which characteristics to use for analysis and reporting. It is important that data systems used to analyse linked data incorporate flexible system functionality (for example, when using the Remote Execution Environment for Microdata (REEM)) which allows researchers to choose the algorithms for selecting student characteristics that best suit the needs of their analysis.

A.3 Preparation of files prior to linking

Linking between the 2010 and 2011 datasets using the Statistical Linkage Key (SLK) required a high quality SLK to be attached to each student record. For the 2010 dataset, the SLK was constructed by the ABS using detailed name, date of birth and sex information, and therefore there were no poor quality SLKs that needed to be removed prior to linking. The SLK was already present on the 2011 dataset and therefore could not be validated for accuracy against full name information. A small number of SLKs on the 2011 file were found to have a length of 15 instead of the standard length of 14. These SLKs were able to be corrected prior to linking to allow a more accurate evaluation of the SLK linkage method.

There were also a very small number of Student IDs on the 2011 file that had a length of 14 instead of 13. These student records were removed from the dataset during the file preparation phase so that they would not negatively affect the results of the linkages.

A.4 Data discrepancies across years

Issues with the quality of the data due to apparent administrative or reporting errors were also identified during the file preparation stage. There were a small number of instances where the sex or date of birth variable was mismatched when student records were linked across years using the Student ID. For example, a student may have been recorded as a female in 2010, but was then recorded as a male on the 2011 file. Alternatively, there may have been an administrative error in recording the date of birth consistently across years.

For the purposes of this feasibility study, the most recent variable recorded for the student was chosen in all analysis work on the linked datasets (i.e. the characteristics from the 2011 file).

These discrepancies affected the quality of the SLK linkage as the sex and date of birth variables are components of the SLK. The discrepancies were not corrected prior to linking with the SLK to allow a true analysis of how these types of errors affected the quality of the SLK linkage methodology. As a result, no students with sex or date of birth variables that were inconsistent across years were linked by SLK or SLK+. This highlights the importance of the quality of the data used to construct the SLK for deterministic linking in data integration projects.

The results of linking using a high quality SLK have been demonstrated in this report using the SLK+ (edited) methodology. For this methodology, SLKs that were inconsistently reported across years were amended prior to linking. The aim of including this methodology was to demonstrate the potential linking benefits that can be achieved with a high quality and accurate SLK.

A.5 Student geography data

For linking with the SLK+ methodology, coded geography information was required for each student. As the Northern Territory has a high rate of student mobility and students that live in very remote localities, accurate address information was not always available for all students. It should be acknowledged that while Northern Territory students may be highly mobile across Communities (home address changes), they may actually remain enrolled at the same school.

If student address was available, the ABS converted it to a geography code in accordance with the Australian Statistical Geography Standard (ASGS). Information was coded to Mesh Block, Statistical Area Level 1 or Statistical Area Level 2, depending on the quality of the address information provided.

If no student level address information was available, the geography code provided for the student's school was used. This meant that these records could be retained for the purposes of trialling the SLK+ linking method. The geography for schools was based on the previous ABS geography standard, the Australian Standard Geographical Classification (ASGC), which included either Collection District or Statistical Local Area. This information had already been derived by the Northern Territory and was included on both files.

When selecting a geography code for use in the SLK+ linking methodology, the lowest geography code available was applied, following the criteria below:

Using student geography:

1. Select the *Mesh Block* in the first instance if available.
2. If Mesh Block is not available, select the *Statistical Area Level 1*.
3. If Statistical Area Level 1 is not available, select the *Statistical Area Level 2*.

Using the student's school geography:

4. If Statistical Area Level 2 for the student is not available (thereby no information is available on the student's address), select the *Collection District of the school*.

5. If the Collection District of the school is not available, select the *Statistical Local Area of the school*.

6. If the Statistical Local Area of the school is not available, assign a *random observation number* to the student as a pseudo-geography code to remove instances of missing linking variables.

There were only a small number of records which had no geographic information available and for which a random observation number was assigned.

A.6 Assigning duplicate Statistical Linkage Key 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. In most instances, the flag was assigned in the same order on the 2010 and 2011 file, resulting in a SLK on the 2010 file and a SLK on the 2011 file 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 and missed 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, especially for the SLK+ linking method where additional variables are included in the linkage key.

A.1 Example of a false link using the SLK and the SLK duplicate flag

2010					2011			
Student ID	Geography	SLK	SLK duplicate flag		Student ID	Geography	SLK	SLK duplicate flag
ABC123	7000112	ABCDE121220032	1	Linked	XYZ789	7000113	ABCDE121220032	1
XYZ789	7000113	ABCDE121220032	2	Unlinked	–	–	–	–

B. KEY CONCEPTS AND DEFINITIONS

B.1 Australian Statistical Geography Standard

The Australian Statistical Geography Standard (ASGS) is the ABS's geographical framework (see ABS, 2012). For the purposes of linking with the SLK+ methodology, student address information provided by the Northern Territory Department of Education was geocoded to the lowest possible region:

- Mesh Block (MB): the smallest area geographical region, with approximately 340,000 covering the whole of Australia. Residential MBs usually contain 30 to 60 households.
- Statistical Area Level 1 (SA1): built from whole MBs with approximately 55,000 covering the whole of Australia. They have an average population of about 400 persons.
- Statistical Area Level 2 (SA2): have an average population of about 10,000 persons, with a minimum population of 3,000 and a maximum of 25,000. There are about 2,200 SA2s, built from whole SA1s.

B.2 Continuous students

Continuous students are those students that were enrolled continuously over the period of 2010 and 2011 in Northern Territory government schools. Continuous students appear in both the 2010 and 2011 datasets. For this report, continuous students are those students that could be linked across both years, hence the different numbers of continuous students that were achieved for each linking method.

B.3 Education levels

The Northern Territory 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.
- *Ungraded secondary schooling:* Students enrolled in any secondary grade from grade 7 through to grade 12.

B.4 Link vs match

The term 'match' is used in this report to indicate that two linked records do belong to the same entity, whereas the term 'link' is used to indicate that two records have been linked together by a data linking process. Linked records may or may not belong to the same entity.

B.5 Main language other than English spoken at home

This variable is defined as the main language other than English, spoken in the home by the student. The classifications presented in this report are either ‘English’ or ‘Other language not further stated’ according to the *Australian Standard Classification of Languages (ASCL)* (ABS, 2011).

B.6 Remoteness of school

The Remoteness structure is based on the ASGS and it classifies Australia into large regions that share common characteristics of remoteness. The Remoteness structure is classified as follows (in order of least remote to most remote):

- Major cities of Australia
- Inner regional Australia
- Outer regional Australia
- Remote Australia
- Very remote Australia
- Migratory

In the Northern Territory, the capital city (Darwin) and immediately surrounding areas are classified as ‘outer regional’ areas due to the Northern Territory’s location and population size. As a consequence, the classifications ‘major cities’ and ‘inner regional’ are not applicable to the Northern Territory. It should also be noted that a considerable share of the Northern Territory population live in areas classified as ‘remote’ or ‘very remote’.

B.7 Students repeating grades

Students repeating grades are those students that were enrolled in the same grade over the period of 2010 and 2011. These students therefore appear in both the 2010 and 2011 datasets and have the same school grade recorded for both years. Ungraded secondary students are not included in the numbers of students repeating grades as there is no method to determine whether these students are actually repeating a grade. There is also no way to determine if students new to the Northern Territory government school system in 2011 are repeating the grade they were previously enrolled in at another school.

C. QUALITY ASSURANCE PROCESS

C.1 Enrolment and student numbers within government schools, Northern Territory, 2010–2011

	No. of enrolments	*No. of students*	
	Number of records on file	*No. of unique records on file*	*Number of students with multiple enrolments*
		2010	
Sex			
Female	14,028	13,877	151
Male	15,101	14,909	191
Total	*29,129*	*28,786*	*342*
		2011	
Sex			
Female	14,139	13,967	169
Male	15,204	14,999	205
Total	*29,343*	*28,966*	*374*

C.2 Distribution of government school students with multiple enrolments across education level, Northern Territory, 2010–2011

	Education level			
	Primary	*Secondary*	*Senior Secondary*	*Number of students with multiple enrolments*
		2010		
Sex				
Female	92	51	11	154 (a)
Male	110	68	20	198 (b)
Total	202	119	31	352 (c)
		2011		
Sex				
Female	112	44	22	178 (d)
Male	131	59	28	218 (e)
Total	243	103	50	396 (f)

(a),(b),(c),(d),(e) & (f) Counts do not match those in table C.1 as there are students with multiple enrolments across different grade levels and consequently different education levels.

D. SUPPORTING INFORMATION AND TABLES

Table D.1 shows the composition of the 2010 and 2011 datasets prior to linking, following the file preparation methods outlined in Appendix A.

D.1 Number of students enrolled in government schools, by selected characteristics (a), Northern Territory, 2010 and 2011

	Year of enrolment	
	2010	*2011*
Education level		
Primary (b)	18,285	18,363
Secondary (c)	7,564	7,584
Senior secondary (d)	2,821	2,919
Ungraded secondary	89	76
Sex		
Female	13,873	13,964
Male	14,886	14,978
Indigenous status		
Aboriginal or Torres Strait Islander	12,773	12,906
Non-Indigenous (e)	15,986	16,036
Remoteness of school		
Major cities	n.a.	n.a.
Inner regional	n.a.	n.a.
Outer regional	15,395	15,449
Remote	5,000	5,047
Very remote	8,364	8,446
Migratory	n.a.	n.a.
Total	*28,759*	*28,942*

(a) Data in this table is at the student level. Additional records for students with multiple enrolments have been removed from these numbers.
(b) Includes students enrolled in pre-year 1 to grade 6.
(c) Includes students enrolled in grade 7 to grade 10.
(d) Includes students enrolled in grade 11 and grade 12.
(e) Includes responses of not stated and inadequately described.

D.2 Number of students enrolled in government schools, by grade, Northern Territory, 2010 and 2011

	Year of enrolment	
	2010	*2011*
Grade		
Pre-year 1	2,646	2,796
Grade 1	2,577	2,585
Grade 2	2,843	2,543
Grade 3	2,611	2,724
Grade 4	2,658	2,662
Grade 5	2,505	2,584
Grade 6	2,445	2,469
Grade 7	2,093	1,988
Grade 8	1,885	1,996
Grade 9	1,772	1,799
Grade 10	1,814	1,801
Grade 11	1,673	1,709
Grade 12	1,148	1,210
Ungraded secondary	89	76
Total	*28,759*	*28,942*

D.3 Number of missed links for each linkage method, by selected characteristics, Northern Territory, 2011

	Linking method		
Characteristic	*SLK*	*SLK+*	*SLK+ (edited)*
Sex			
Female	624	620	6
Male	687	676	3
Indigenous status			
Aboriginal or Torres Strait Islander	826	811	9
Non-Indigenous (a)	485	485	0
Education level			
Primary	783	772	n.p.
Secondary	412	408	n.p.
Senior secondary	105	105	0
Ungraded secondary	11	11	0
Remoteness of school			
Outer regional	557	555	0
Remote	169	168	n.p.
Very remote	585	573	n.p.
Main language spoken at home			
English	316	314	0
Other language	995	982	9
Total missed links			
Total	*1,311*	*1,296*	9

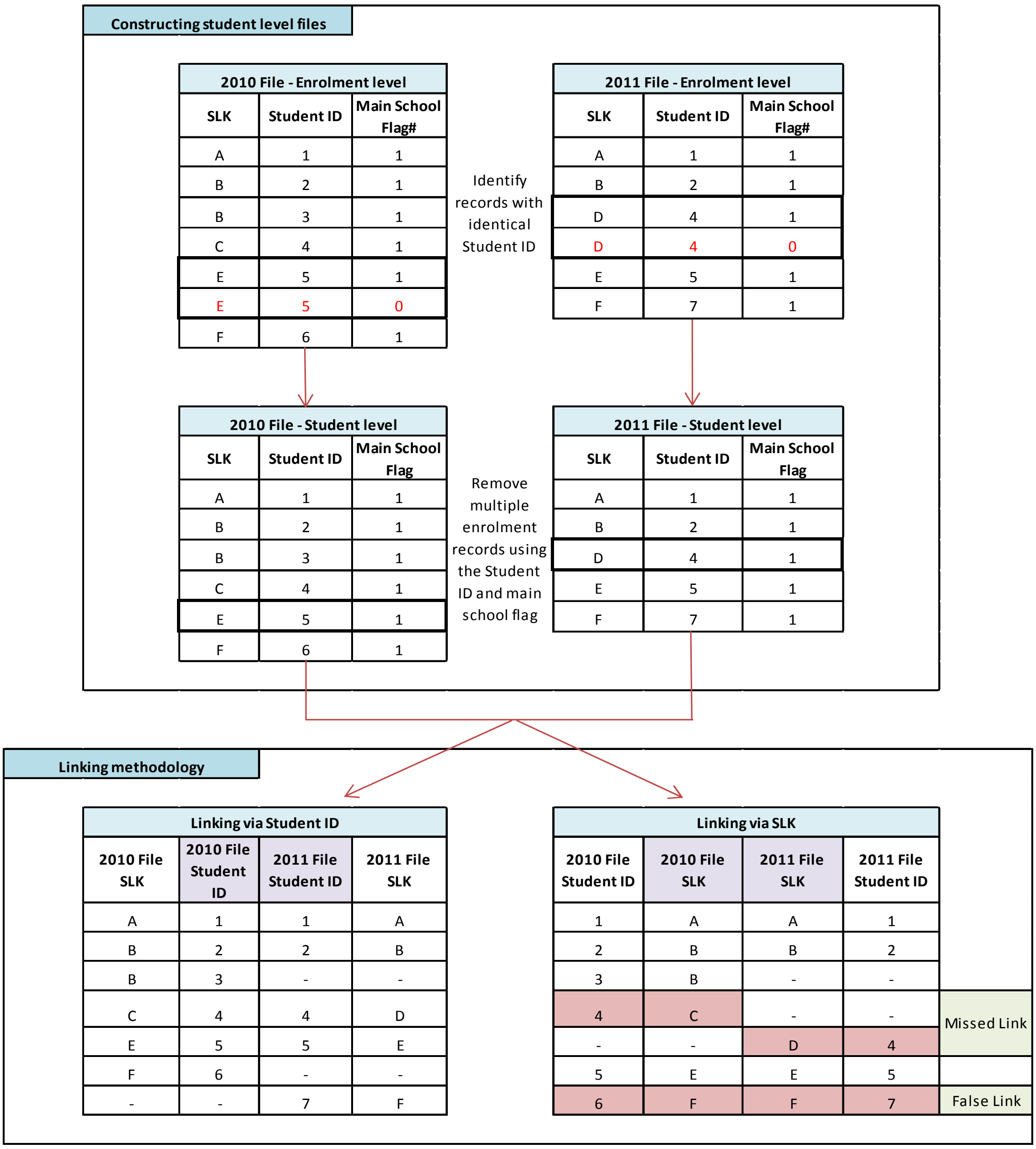
(a) Includes responses of not stated.
n.p. Not for publication.

D.4 Number of false links for each linkage method, by selected characteristics, Northern Territory, 2011

	Linking method		
Characteristic	*SLK*	*SLK+*	*SLK+ (edited)*
Sex			
Female	100	96	93
Male	108	97	95
Indigenous status			
Aboriginal or Torres Strait Islander	152	138	133
Non-Indigenous (a)	56	55	55
Education level			
Primary	137	125	123
Secondary	53	50	49
Senior secondary	18	18	16
Remoteness of school			
Outer regional	70	68	68
Remote	37	37	37
Very remote	101	88	83
Main language spoken at home			
English	53	52	52
Other language	155	141	136
Total false links			
Total	*208*	*193*	*188*

(a) Includes responses of not stated.

D.5 Process for construction and linking of files using a unique student identifier



Constructing student level files

2010 File - Enrolment level

SLK	Student ID	Main School Flag#
A	1	1
B	2	1
B	3	1
C	4	1
E	5	1
E	5	0
F	6	1

Identify records with identical Student ID

2011 File - Enrolment level

SLK	Student ID	Main School Flag#
A	1	1
B	2	1
D	4	1
D	4	0
E	5	1
F	7	1

2010 File - Student level

SLK	Student ID	Main School Flag
A	1	1
B	2	1
B	3	1
C	4	1
E	5	1
F	6	1

Remove multiple enrolment records using the Student ID and main school flag

2011 File - Student level

SLK	Student ID	Main School Flag
A	1	1
B	2	1
D	4	1
E	5	1
F	7	1

Linking methodology

Linking via Student ID

2010 File SLK	2010 File Student ID	2011 File Student ID	2011 File SLK
A	1	1	A
B	2	2	B
B	3	-	-
C	4	4	D
E	5	5	E
F	6	-	-
-	-	7	F

Linking via SLK

2010 File Student ID	2010 File SLK	2011 File SLK	2011 File Student ID	
1	A	A	1	
2	B	B	2	
3	B	-	-	
4	C	-	-	Missed Link
-	-	D	4	
5	E	E	5	
6	F	F	7	False Link

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

D.6 Number of students enrolled in school, by selected characteristics, Northern Territory, 2011

	Linking method			
Characteristic	*Student ID*	*SLK*	*SLK+*	*SLK+ (edited)*
	REPEATING A GRADE (a)			
Sex				
Female	298	276	275	302
Male	372	353	349	378
Indigenous status				
Aboriginal or Torres Strait Islander	470	437	433	477
Non-Indigenous (c)	200	192	191	203
Remoteness of school				
Outer regional	196	187	186	199
Remote	179	165	164	181
Very remote	295	277	274	300
Main language spoken at home				
English	179	168	168	179
Other language	491	461	456	501
Total				
Total students repeating grades	*670*	*629*	*624*	*680*
	CONTINUOUS (b)			
Sex				
Female	10,733	10,209	10,209	10,820
Male	11,631	11,052	11,052	11,723
Indigenous status				
Aboriginal or Torres Strait Islander	10,024	9,350	9,351	10,148
Non- Indigenous (c)	12,340	11,911	11,910	12,395
Remoteness of school				
Outer regional	12,214	11,727	11,727	12,281
Remote	3,728	3,596	3,597	3,764
Very remote	6,422	5,938	5,937	6,498
Main language spoken at home				
English	9,289	9,026	9,027	9,341
Other language	13,075	12,235	12,234	13,202
Total				
Total continuous students	*22,364*	*21,261*	*21,261*	*22,543*

(a) Excludes students recorded as being in ungraded secondary.
(b) Includes students recorded as being in ungraded secondary.
(c) Includes responses of not stated.

D.7 Number of repeating and continuously enrolled students, by grade (a), Northern Territory, 2011

	Linkage method			
	Student ID	*SLK*	*SLK+*	*SLK+ (edited)*
	REPEATING A GRADE (a)			
Grade in 2011				
Pre-year 1 (b)	104	94	94	105
Grade 1	58	52	51	59
Grade 2	46	41	40	47
Grade 3	35	33	33	35
Grade 4	51	52	51	53
Grade 5	35	34	34	36
Grade 6	34	32	31	34
Grade 7	26	23	22	26
Grade 8	29	27	27	29
Grade 9	33	29	29	33
Grade 10	56	53	53	57
Grade 11	73	72	72	76
Grade 12	90	87	87	90
Total				
Total students repeating grades	*670*	*629*	*624*	*680*
	CONTINUOUS			
Grade in 2011				
Pre-year 1 (b)	105	95	95	106
Grade 1	2,190	2,072	2,071	2,209
Grade 2	2,199	2,112	2,112	2,217
Grade 3	2,416	2,293	2,294	2,430
Grade 4	2,322	2,219	2,219	2,347
Grade 5	2,276	2,159	2,159	2,296
Grade 6	2,160	2,072	2,071	2,179
Grade 7	1,681	1,571	1,571	1,702
Grade 8	1,689	1,575	1,576	1,695
Grade 9	1,511	1,443	1,443	1,519
Grade 10	1,382	1,315	1,315	1,394
Grade 11	1,284	1,228	1,228	1,295
Grade 12	1,097	1,066	1,066	1,102
Total				
Total continuous students	*22,312*	*21,220*	*21,220*	*22,491*

(a) Excludes students recorded as being in ungraded secondary.

(b) The majority of continuous students in pre-year 1 are repeating that grade, and therefore the numbers are low.

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

1351.0.55.046 • RESEARCH PAPER: ASSESSING THE QUALITY OF DIFFERENT DATA LINKING METHODOLOGIES ACROSS TIME, USING NORTHERN TERRITORY GOVERNMENT SCHOOL ENROLMENT DATA

Produced by the Australian Bureau of Statistics