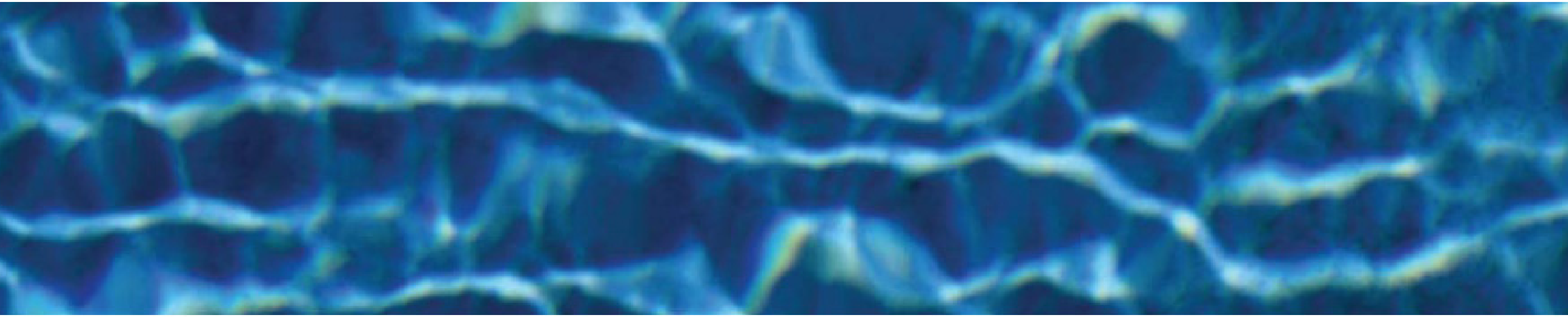
MINISTRY OF EDUCATION
Te Tāhuhu o te Mātauranga

Linking students

Review of the methods used to link students in historical New Zealand tertiary education data



New Zealand Government

Learners in tertiary education

This report forms part of a series called *Learners in tertiary education*. Other topics covered by the series are access, pathways, support, participation, retention and qualification completions.

Author:
David Scott
Email: david.scott@minedu.govt.nz
Telephone: 04-463 8052
Fax: 04-463 8526

Published by:
Tertiary Sector Performance Analysis and Reporting
Strategy and System Performance
MINISTRY OF EDUCATION

This report is available from the Ministry of Education's Education Counts website:
www.educationcounts.govt.nz

November 2007

ISBN 978-0-478-13764-4
ISBN 978-0-478-13765-1 (web)

Linking students, review of the methods used to link students in historical New Zealand tertiary education data

Contents

Glossary of terms

The following terms and acronyms are used commonly in this report:

False positive

Where records are linked as the same student, but are, in fact, different students. Also known as mismatches.

False positives lead to an undercount of students, and act to understate rates of completion, and overstate rates of attrition.

False negative

Where records are not linked as the same student, even though they are the same student. Also known as non-matches.

False negatives lead to an overcount of students, and act to overstate rates of completion, and understate rates of attrition.

Qualification completion rate

The percentage of students starting a qualification that go on to complete a qualification. Often expressed as completion after some defined period of time, eg five years.

First-year qualification attrition rate

The percentage of students starting a qualification that do not complete a qualification that year, and do not enrol in the following year.

Institution and system rates

An institution completion or attrition rate considers only those students who enrol or complete at the same institution. A system rate includes students who re-enrol or complete at a different institution to the one they started at. eg System rates include transfers, whereas institution rates count transfer as non-completion. System rates will therefore generally always be higher than institution rates.

Commonly used acronyms in this report:

NSN	National Student Number
SN	The unique student number derived from the matching
TEC	Tertiary Education Commission
BMR	Baseline Monitoring Report
SDR	Single Data Return. The name of the survey used to collect tertiary education enrolments and completions data.
TEO	Tertiary Education Organisation
ITP	Institute of Technology or Polytechnic
PTE	Private Training Establishment

The terms "institution", "provider" and "Tertiary Education Organisation" or "TEO" are used interchangeably in his report, as are the terms "linking" and "matching".

Summary

This report looks at the methods used by the Ministry of Education to link the records of tertiary education students across time and across providers. The ability to link students helps us to answer important questions about the performance of New Zealand's tertiary education system, such as what proportion of students complete their qualification, how long do people study for, or how much funding does not result in some form of recognised achievement.

Up to 2003, the lack of a unique national student number limited our ability to answer such questions in a systematic or complete way. In 2003, two things happened to change this. National student numbers (NSNs) were introduced, and a statistical matching process was developed to estimate unique student numbers (which were called 'SNs'), that would link students enrolled without NSN before 2003.

The development of this derived linking variable has been used for a range of subsequent research and analysis. National estimates of the rates of retention and completion were published for the first time in 2004, along with the rates that students progress to further study after completion. A number of additional longitudinal pathway studies have been undertaken using SN which have provided further new perspectives on the performance of the tertiary education system in New Zealand.

This analysis has been mainly at the system level. An assessment of the original methodology concluded that this matching was robust for statistical reporting purposes at an aggregate level, but not always robust at an institution level, or for data before 1998. The assessment found that the matching was more likely to under-report linkages than to make false links, resulting in a slight over-estimation of student counts and attrition rates, and an under-estimation of qualification completion rates.

There were two reasons that led to a review of the matching. Firstly, since the initial matching code was developed, there have now been four years of enrolments data with NSN available (2003 to 2006). Differences were beginning to emerge between completion and attrition rates before 2003 (based on SN) with rates from 2003 (based on NSN). For longer qualifications, in particular, such as bachelors degrees, and doctorates, a clear discontinuity was appearing in rates primarily based on SN, with those mostly using NSN.

Secondly, as part of the tertiary reforms introduced by the government in 2006, the Tertiary Education Commission began developing a new system of institution monitoring to support the government's 'investment in a plan' approach to funding tertiary education organisations. Key indicators of performance for each institution included measures of attrition, completion and progression. The initial round of Baseline Monitoring Reports developed during 2007 used SN as part of the first-year attrition and completion rate indicators[1]. While the Baseline Monitoring Reports clearly needed to be accurate at the institutional level, it was known that SN did not always provide an ideal matching for some institutions. There was a need to have better information on this quality for each individual institution.

The needs of the TEC baseline monitoring process as well as the emerging discontinuity in system completion and attrition rates provided the imperative for the Ministry of Education review the matching methodology. The availability of four years of NSN data was able to provide a powerful independent means to measure the accuracy of the matching, and to revise the match weights and processes used. The methodology was extensively reviewed in 2007 and SNs were regenerated for all existing enrolments and completions data from 1994 to 2006.

[1] Monitoring reports use the SN data from after the review.

This report documents this review, the quality of the revised matching, and its subsequent impact on student counts and rates of completion and attrition. The review also provides a resource for agencies and providers to inform discussions on the quality of institution and system level indicators based on SN, and to illustrate the rigour of the dataset.

The review has led to a measurable improvement in the quality of the linking. This quality was assessed by comparing what would have happened to counts and rates in the absence of NSN, with counts and rates based on NSN.

In all this analysis, the assumption is made that there are no NSN errors. While the level of NSN errors is considered small enough to make this assumption, NSN errors definitely still exist in the data, particularly for some institutions. When SN and NSN differences were examined in the 2004 assessment of the old matching, it was estimated that SN was likely to be in error 75 percent of the time, while NSN was likely to be incorrect 25 percent of the time.

However, four years of improvements in NSN allocation and validation, and the progressive identification and centralised recording of students with more than one NSN, or NSNs assigned to more than one student have further reduced the frequency of NSN error. While this review highlights that some NSN errors still exists[2], the level was considered small enough to assume that any differences with SN were more likely to be matching errors. Therefore, the error rates presented in this report represent a maximum potential error rate in SN, ie assuming there were no NSN errors.

In line with the need for improved institution-level information, the review measured the quality of matching both within-institution and across institutions. Over 60 percent of students re-enrol with the same provider, while almost all students completing a qualification, do so at an institution that they are, or have been enrolled in. Linking students who enrol in the same institution was eight times more accurate than linking students enrolling in different institutions. Resulting institution attrition and completion rates are therefore also more accurate than system rates (ie which include transfers).

SN error rates

Matching	Enrolments			Completions		
	False positives	False negatives	Total False	False positives	False negatives	Total False
Within institution	0.3%	0.3%	0.5%	4.1%	0.5%	4.7%
Across institution	1.4%	2.6%	4.0%	*Negligible student numbers*		
Total	2.2%	1.4%	3.6%	5.9%	1.9%	7.8%

The revised matching had a total error rate of 3.6 percent in enrolments data and 7.8 percent in completions data. That is, for every 1,000 qualification enrolments 36 should have been assigned a different SN, while for every 1,000 qualification completions, 78 should have been assigned a different SN. For those students enrolling in the same institution the error rate was just 0.5 percent (5 errors in every 1,000 enrolments).

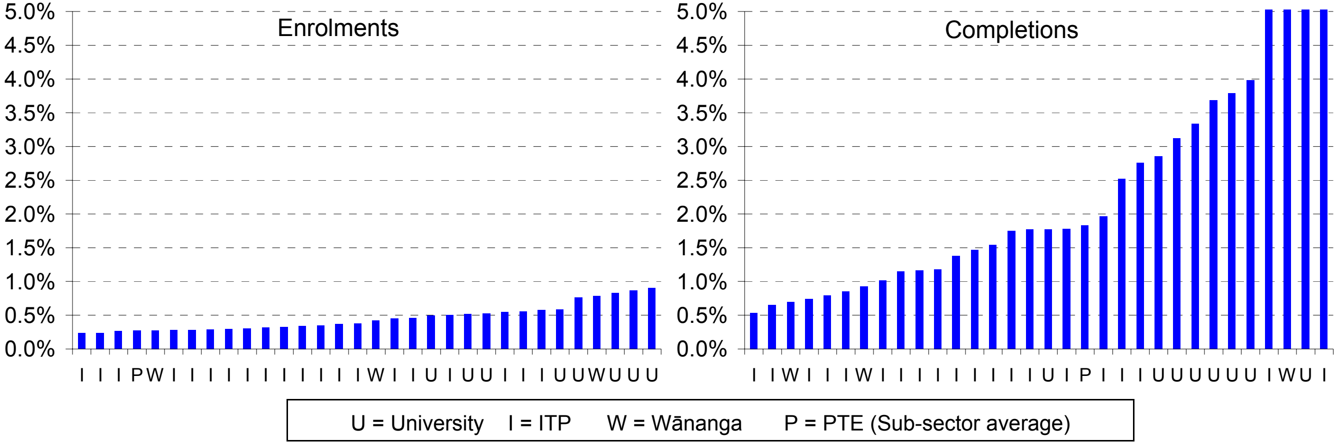
1.4 percent of enrolments were not assigned the same SN from another enrolment belonging to the same student (false negative). This was an 81 percent improvement on the old matching method. A further 2.2 percent of enrolments were incorrectly assigned the same SN from another enrolment (false positive). This was a 21 percent improvement on the old matching method. Overall, the review was successful in increasing the ability to link to other students, without increasing the level of false links.

[2] Possibly significant in some cases for some large institutions.

Surprisingly, the review found that the error rate in completions was higher than for enrolments – where the matching is significantly more complex. In contrast to enrolments, nearly all (over 98 percent of) completions match to an enrolment at the same institution. The error rate for matching completions to an enrolment at the same institution was 4.7 percent compared with 7.8 percent overall. Both these rates however, rate are affected by the highly skewed distribution of error rates across institutions. In fact, just four institutions (including 2 polytechnics, one university and one wānanga) account for 66 percent of potential errors. Put another way, if these four institutions are removed the error rate reduces from 4.7 percent to 2.2 percent. Nevertheless, the level of error in completions was still higher than for enrolments, and not as low as was expected from the review.

The matching largely works the same for all institutions. Therefore one might expect a relatively uniform distribution of error. The highly skewed distribution of error rates in the completions data indicates potentially significant quality issues with the data submitted for those institutions whose error rates differ significantly from the median. By contrast, the distribution of error rates in the enrolments data was more uniform.

Distribution of potential error rates by institution



The three-year institution completion rates that would have resulted from the revised matching in the absence of NSN were compared with those based on NSN. Average differences were less than one percentage point across sub-sectors and levels of study, except for masters and doctorate levels where they were between one and two percentage points. Percentage differences were greater for some smaller Private Training Establishments (PTEs). Three Tertiary Education Institutions (TEIs) including one university, two polytechnics, and one wānanga had differences of more than two percentage points for particular levels of study. Over 70 percent of institutions had no difference in SN-based or NSN-based rates. The matching tended neither to understate nor overstate completion rates, with 15 percent with higher SN-based rates, and 15 percent with lower SN-based rates.

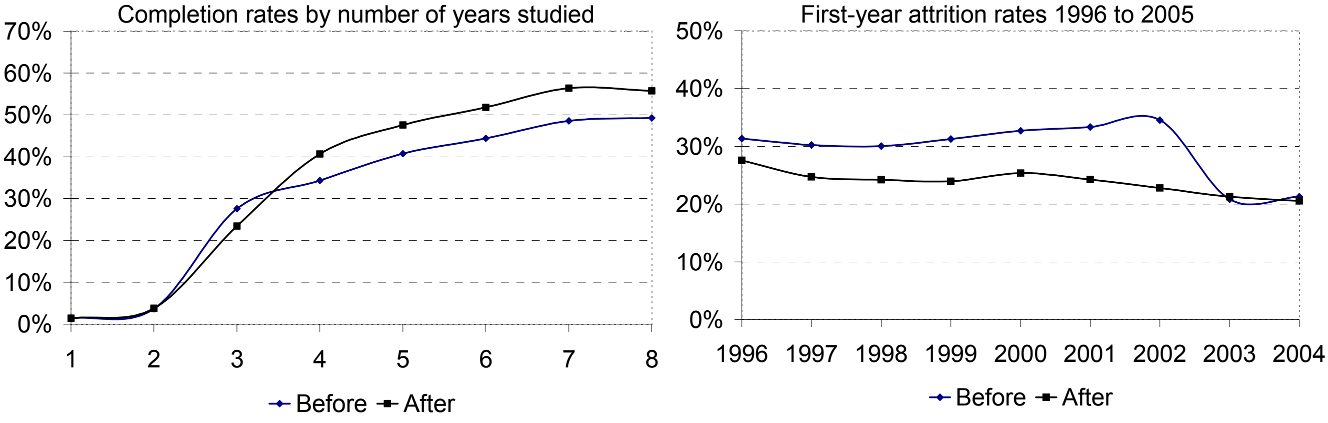
Average difference in institution rates between SN and NSN

Level of study	Difference in three-year completion rate	Difference in first-year attrition rate
Level 1-3 Certificates	0.6%	0.5%
Level 4 Certificates	0.9%	0.7%
Diplomas	0.9%	0.7%
Bachelors	0.3%	0.4%
Level 8 Postgraduate	0.8%	0.2%
Masters	1.9%	0.2%
Doctorates	1.8%	0.4%
Any level	0.7%	0.6%

Note: These figures are the average institution difference in absolute percentage point terms between rates based on SN and the same rate based on NSN

The impact of the revised matching on attrition and completion rates can also be seen by comparing the rates based on the old SN with rates based on the new SN. The new SN has resulted in improved linkages, and therefore higher estimates of completion rates than previously published. The major differences can be seen in long-term rates at bachelors and doctorate levels where for example, the estimated eight-year completion rate has gone from 49 percent to 56 percent for bachelors degrees, and from 42 percent to 56 percent for PhDs. Similarly, first-year attrition rates have also improved most noticeably at bachelors and PhD level, where previous discontinuities have been removed and rates pre- and post-2003 are now more comparable.

Bachelors degree completion and attrition rates before and after the review



The review has also reduced discontinuities in linking that occurred in 2000, when the survey used to collect data was significantly changed, and in 1997 when abbreviated names were first collected. Previously, rates based on starters before 1998 were not considered robust enough to be published, and the review had not been expected to change this significantly. However, the rates for pre-1998 starters now appear more in line with those starters from 1998 on. This provides an unexpected benefit of the review in terms of being able to provide new information on long-term rates of completion. In particular, it highlights the longer time to completion for PhD students, where the rate of completion increases from around 30 percent after five years to around 60 percent after nine and ten years of study.

1 Introduction

This report looks at the methods used by the Ministry of Education to link records of tertiary education students across time and across institutions.

The ability to link students allows us to answer some important questions about New Zealand's tertiary education system. What proportion of students complete their qualification? What proportion leave without completing a qualification? How long do people study for? How many change what or where they study? How do these things vary from institution to institution, or between men and women, or between school leavers and those who entered tertiary education from the workforce or from unemployment?

Up to 2003, the lack of a unique national student number limited our ability to answer such questions in a systematic or complete way. In 2003, two things happened to change this. National student numbers (NSNs) were introduced, and a statistical matching process was developed to estimate unique student numbers (called SNs), that would link students enrolled before 2003.

The development of the derived linking variable has been used for a range of subsequent research and analysis to improve our understanding of the performance of the tertiary education system in New Zealand. National estimates of the rates of retention and completion were published for the first time in 2004, along with the rates at which students progress to further study after completion. These rates have been updated annually since then in reports and tables available on the Ministry of Education's website. A number of longitudinal pathway studies have been undertaken using SN. The derived SN also forms the basis of student counts before 2003. Previously, a student enrolled in more than one institution had been counted twice, whereas SN provided the basis for a good estimate of the number of unique individuals participating in tertiary education. SN has been used in research to assess the impact of the Performance-Based Research Fund on PhD retention. Some of these studies are listed in the bibliography at the end of this report.

The initial matching code was first developed and run prior to the availability of NSN. It derived a unique student identifier, called 'SN' for all qualification enrolments and completions between 1994 and 2002. The matching has been run annually since then to link latest year's data with previous years. The matching has had minor changes each year, but has largely used the same methodology and match-weights as used in 2003. With the introduction of NSN, the matching was modified to utilise NSN directly where available, otherwise to use existing SNs from previous year's enrolments. For data from 2003 on, each SN has only one corresponding NSN, and conversely each NSN has only one corresponding SN.[3]

In 2004, a quality assessment was done on the matching. This utilised the 2003 enrolments data, which was the first year that NSNs were available. The matching routine was run independently of NSN, and records linked by this method were compared with records linked using NSN. The study found that the matching programme had an error rate of 3.7 percent. That is, for every 1,000 pairs of enrolment records being compared, 37 pairs would be incorrectly matched or not matched. The matching was more likely not to match two records that were the same student, than to mismatch two distinct records. The false positive error rate was 0.3% and the rate of false negative matching was 3.4%. A higher level of false negatives means higher student counts, higher attrition rates, and lower completion rates. This report "Assessment of TSPAR Matching (SNs and NSNs)" can be found on the Ministry's Education Counts website.

[3] For full documentation of this initial matching process as run in 2003 see "Retention, Completion and Progression in Tertiary Education 2003, Technical Documentation" on the Ministry of Education's Education Counts website.

Since the initial matching was developed, there have now been four years of enrolments data with NSN available (2003 to 2006). Emerging trends in completion and attrition rates were now beginning to show some differences between rates before 2003 (based on SN) with rates from 2003 (based on NSN). For longer qualifications, in particular bachelors degrees, and doctorates, a clear discontinuity was appearing in rates primarily based on SN, with those mostly using NSN.

Latest five and six-year completion rates based on 2000 and 2001 starters were also now starting to show discontinuities with previously published rates based on 1998 and 1999 starters. The reasons for this highlighted the extent of change that occurred in 2000 when the current Single Data Return (SDR) collection of tertiary enrolments and completions was introduced. The data systems used to collect and report tertiary enrolments and completions were significantly updated to accommodate the new SDR and its increased reporting and validation requirements. These changes had a significant impact on the quality of information held by institutions and collected by the Ministry of Education. Key matching variables, such as student ID along with other demographic information were updated, meaning that there was a greater chance that the record would not match to historical records.

Completion and attrition rates are affected in two ways by the matching – how well the matching can trace a student forward in time (to whether they re-enrolled or completed) – but also how well the matching can trace back in time – to determine whether a student has previously been enrolled at a particular level of study[4]. The determination of 2000 and 2001 starters was impacted by these changes. In particular, for smaller subgroups such as PhD starters, there was a discontinuity in completion rates for pre- and post-2000 starters.

Another key change occurred in 1997 when the first four characters of student surnames were collected for the first time. This key matching variable significantly improved the ability to link students. Its effect is noticeable when pre 1998 rates are compared with rates from 1998 on. This factor, along with other differences in the collection and quality of data in these early collection years was such that before this review, only rates from 1998 on were considered robust enough to be published.

Over the last four years, since the original matching was developed, there has been substantial work aimed at improving the quality, completeness and consistency of underlying historical enrolments and completions data, and ongoing improvements also to the SDR collection itself. This includes improvements to many of the variables used directly in the matching process. It was expected that even in the absence of any review, a re-running of the matching should improve resulting SNs.

The original driver for the development of unique student identifiers was to enable the Ministry of Education to estimate system rates of attrition and completion. On the first publication of completion rates in 2004, there was an immediate interest in rates at the institution level. While rates for many individual institutions were considered robust, not all were, and in the absence of a fully validated NSN, assessing the quality of individual institution rates based on SN was difficult. However, at aggregated levels, the matching was considered robust. Because of the nature of the matching, and the obvious need to have some degree of confidence in SN-based institution rates, a decision was taken not to publish at institution level.

In 2006, the Government released its Tertiary Education Strategy for 2007 to 2012. The priorities included in that document place increased emphasis on qualification completion. Alongside the new strategy, the government has reformed the approach to funding, shifting to an investment approach. This requires Tertiary Education Organisations (TEOs) to commit to the achievement of outcomes in investment plans approved by the Tertiary Education Commission (TEC).

[4] This determines whether a student belongs in the starting cohort, ie whether they are counted in the denominator used for calculating the rate.

A new system of institution monitoring was also introduced to support this new "investing in a plan" approach, and to ensure alignment with wider tertiary education priorities. In the initial phase of the development of this monitoring, baseline performance information in the form of Baseline Monitoring Reports (BMRs) was produced in 2007. These were used to inform discussions between the TEC and TEOs on the development of plans. The monitoring reports included indicators on rates on qualification attrition, completion and progression. Limited weight was placed on this performance information in this initial round, to recognise that there were quality and coverage issues with the data at a TEO level which the TEC and MoE are working with the sector to resolve.

Indicators such as attrition, completion and progression require the use of a unique student identifier. NSN is considered the most robust identifier and is used in some aspects of the BMR such as participation and successful course completion. However, NSN is only available from 2003 and so is not yet able to provide information on longer-term rates of qualification completion, such as for bachelors degrees or PhDs. In addition, to ensure data quality and minimise future changes it would be useful for the methodology adopted to allow for the use of the NSN going forward. So, for the initial round of BMRs, the TEC decided to use SN data from Ministry of Education to provide the linking for years before 2003[5].

SN is derived from a complex statistical matching algorithm and was created to support research and policy development, rather than monitoring. And to a degree, some of the underlying data, had also been collected for different purposes (especially, for funding, research, and policy development). However, the use of SN provided some significant benefits for monitoring, which were not possible with NSN. While SN incorporates NSN from 2003 on, it is available back to 1994. This allows longer-term rates of completion to be calculated. The statistical matching used to derive SN takes account of cases where institutional ID numbers have changed over years and thus provides greater accuracy for calculating rates, than by using ID numbers alone. Also, the SN number is applicable across providers and thus allows for progression indicators to be produced, and will also assist other supplemental information, such as transfer rates. SN will have less impact as years pass, and the number of years of available NSN data increases.

While SN matching was considered robust at aggregate levels, and for many individual institutions, it had not been shown to be robust for every individual TEO. Part of the driver for this review was to investigate and gain a better understanding of the quality of SN as it impacted on TEOs to support the interpretation of BMR information and future work on data quality.

The needs of the baseline monitoring process as well as the emerging discontinuity in system completion and attrition rates provided the imperative for the Ministry of Education to review the matching methodology. The availability of four years of NSN data provided a powerful independent means to measure the accuracy of the matching algorithm used for the years before the introduction of the NSN. By re-running the matching independently of NSN, and then comparing resulting SNs with NSNs, the results could be used to revise the match weights and processes, with a view to minimising differences. The revisions and improvements to underlying data were also expected to improve the quality of the matching, independently of any changes to the methodology. The revised matching code used in 2007 was the first full matching run since it was first run in 2003. Every SN was derived from scratch using the new code and weights, and the improved historical data.

The purpose of this report is to document this review, by describing the changes made, measuring their subsequent impact on student counts and on rates of completion and attrition, and to assess the quality of revised matching methodology.

[5] Initial performance data based on the old SN data was subsequently replaced by the revised SN data from after the review, once it became available.

2 Matching process

This section provides an overview of the matching process. For more complete documentation of the methods used please contact the author.

The Ministry of Education collects data on every individual enrolment in tertiary education. From 1994 to 1999 these were at qualification level. From 2000, the collection operated at course level. Summarised data from this collection holds one record for every qualification a student is enrolled in during the year. It is at this level that each SN is assigned. The Ministry of Education also collects data on the number of qualifications completed each year in tertiary education organisations. In conjunction with enrolments, these are used to tell us what the rate of qualification completion is, ie what percentage of students complete qualifications. The matching also generates an SN for each qualification completion, so that it can be linked back to the enrolments data.

The matching code comprises three main matching programmes. These are:

- Match
- Join the dots-Weakest link
- Match with Completions

The first two matching programmes are used to derive the unique student number SN for enrolments data. The last programme, Match with Completions, is used to derive SN for completions data.

Match is the first programme. This compares enrolment records for one year with enrolment records for another year. If it considers that the two records represent the same student, it outputs this pair of record identifiers to a matches file. All enrolment files for each year 1994 to 2006 are compared with each enrolment file from 1994 to 2002 (ie where no NSN is present). For matching 2003 to 2006 data with 2003 to 2006, NSN is used. The final set of all matches containing pairs of matching records and their match score is the output of this programme, and the input for the second stage of the matching, Join the dots-Weakest link.

There are two stages within Match. The first stage, loosely called 'exact matching', uses an institution code and the student ID code assigned by institutions to link the majority of students who re-enrol at the same institution and have the same student ID from year to year. Additional demographic information from both enrolments is compared to help validate the match. For most institutions these matches are very reliable (see section 5.1).

The second stage, loosely called 'statistical matching', is used to link those students enrolled across different institutions, or in institutions where student IDs have changed over time. Up to 14 fields are compared between the two enrolments records. Each contributes a different positive weight if they are the same, and a different negative weight if they are different. Any fields with missing values, neither add nor subtract anything to the score. The amount each field adds or subtracts differs for each field, and each of these weights has been carefully calibrated so that a score of one or higher is considered to represent the same student. Every enrolment record is compared against every other record until a match is found. So for two files with 300,000 enrolments each, there are potentially 90 billion comparisons.

The following table shows the 14 fields that are used in the statistical matching part of the match process. It shows one example of a possible combination of match results when the 14 fields between two different enrolments are compared. In this particular case, they are linked as the same student even though not all the fields match.

Linking field	Match status
Institution	different
First 4 characters of surname plus first initial	different
4 character surname only	same
Year of birth	same
Month of birth	same
Day of birth	same
Gender	same
Last secondary school	same
Highest school qualification	different
Last year of secondary schooling	same
Ethnic group	different
Country of citizenship	same
Disability status	same
First year in tertiary	different

In the absence of NSN, no one can know for sure for any given combination, whether they are the same student or not. Even if all 14 fields are identical on the two records being compared, there will be a few cases when these records relate to two different students. In the case above, for example, when 2003 enrolments are compared with 2006 enrolments, in all but two cases the two records had the same NSN. By linking all, we allow some to be falsely linked. Conversely, by allowing none to link, means some students may end up with different SNs. Therefore, it is not possible to develop a methodology, using the data alone, which is 100% accurate. In the matching developed here, the setting of the weights is critical for ensuring that the level of mismatches and non-matches is minimised. The use of NSN in the settings of these weights and measurement of subsequent error is discussed further in Section 3.

Even if two records are linked as the same student from this process, the final decision on whether two records end up with the same SN or not is determined only after the next of the matching is run, that is the Join the dots-Weakest link stage. This programme provides an important quality validation check on the initial matches. In this programme, all pairs of matched enrolment records are logically grouped. For example, if record A matches with record B, and record B matches with record C, then records A, B and C are all grouped – regardless of whether A matches to C or not. This part of the process is called joining the dots. A specific algorithm then tests to see if this grouped set of records are likely to represent one student, or whether a spurious match (a weak link) has caused two or more students to be falsely linked. If a weak link is detected, the group is split accordingly. SNs are allocated to each final group of linked enrolment records. So while the first Match programme determines links on a record by record basis, the Join the dots-Weakest link assesses and confirms these matches by using the information from all the matched pairs.

The last matching programme, Match with Completions, is used to link a qualification completion record with a student in the enrolments data. The linking variables used here are institution, student ID, date of birth, gender and qualification code. Where a match is determined, the SN value from the enrolments record is copied to the matching completions record. The linking with completions is more straightforward as over 90 percent of completions have a corresponding enrolment in the same institution at the same year, or in the immediate one or two years prior. Here, we are interested in finding a corresponding student in the enrolments file, in order to assign an SN. The qualification completed and any linking qualifications enrolled in do not necessarily need to match.

The matching is designed so that each NSN can only have one SN value, and so that each SN has only one NSN. Where the matching has correctly identified two different students, but these students have incorrectly been assigned the same NSN, the code will assign only one SN, so that one to one correspondence with NSN is maintained. While NSN errors can occur, the error rate is very small, and such cases are more likely to represent SN errors.

3 How the review was done

3.1 Using NSN to refine match-weights

At the time of the review in 2007, NSN data was available for four full years of enrolment data, 2003 to 2006. The matching uses NSN to directly link enrolments involving these years. However, by removing this use of NSN, and running the matching independently of NSN for the years 2003 to 2006, we can then compare those records which are considered the same student based on SN, with those considered the same student based on NSN.

This approach formed the basis of the review. Those records where the SN-based and NSN-based linkages were not consistent were grouped according to the frequency with which combinations of the 14 fields used matched or did not match. Those combinations with the highest differences were examined further. The match-weights for particular fields were then altered to the point that those comparisons that should have linked did so, or those that should not have linked, failed to do so.

For example, in one of the early comparisons between 2003 and 2004 enrolments, the following was a relatively frequently occurring combination that did not link, whereas according to NSN they should have linked.

Linking field	Match status
Institution	different
First 4 characters of surname plus first initial	same
4 character surname only	not used
Year of birth	same
Month of birth	same
Day of birth	same
Gender	same
Last secondary school	different
Highest school qualification	different
Last year of secondary schooling	different
Ethnic group	same
Country of citizenship	same
Disability status	same
First year in tertiary	different

Even though the secondary school details are different between the two enrolments being compared, they did in fact represent the same student according to their NSN. In this case, by reducing the value of the weight that is subtracted when last secondary school is different, to the point that a score of one is achieved, any two enrolment records with this combination of matches will then link.

In this case, every time this combination occurred, the NSNs were the same, indicating they were in fact the same student. However, for any set of enrolments with the same combination of matching fields there will be a mixture, some the same NSN, and some with different NSNs. Even for combinations where every single linking field matched, there will be a few cases where they are not the same student. Twins are often an example of this.

But in the absence of an NSN, we are not able to tell which are the same student, and which are different, and so we have to make a decision to either link all – or link none. When the original matching was done, we had to guess whether we thought particular combinations were more likely to represent the same student or not. With the availability of NSN we are now to tell more accurately what proportion of a combination of linking fields represents the same student, and what proportion represent different students.

In the case below for example, the initial match-weights were such that any two enrolment records with this combination of match fields resulted in a link. That is, on balance they were considered more likely to be the same student rather than different students.

Linking field	Match status
Institution	different
First 4 characters of surname plus first initial	different
4 character surname only	same
Year of birth	same
Month of birth	same
Day of birth	same
Gender	same
Last secondary school	same
Highest school qualification	same
Last year of secondary schooling	different
Ethnic group	same
Country of citizenship	same
Disability status	same
First year in tertiary	different

However, now we have NSN, we know that 50 percent of the time they had the same NSN, and 50 percent of the time they had different NSNs. By allowing all of these to link also allows half to falsely link. By allowing none of these to link, allows half to falsely not link. We can never therefore reduce errors to zero. However, we can be very sure of the level of error for a particular set of match-weights. Relaxing the weights allows more records to link, some of which will be false links. By tightening the weights we reduce false links, but increase the number of false negatives, ie when two records relating to the same student are not linked.

From the 2004 assessment, we know that the original matching had more of a tendency to under match, than to match incorrectly. In this review then, we aimed to reduce the level of false negatives without increasing false positives. A false negative is where the matching fails to identify two records that are the same student. A false positive is where the matching incorrectly identifies two different records as the same student. False negatives lead to an overcount of students, higher attrition rates, and lower completion rates, while false positives lead to an undercount of students, lower attrition and higher completion rates. Weights were adjusted iteratively in this way, ie until the level of false negatives failed to reduce, and the level of false positives failed to rise. This process saw the level of false negatives drop by nearly 80 percent, while the level of false positives dropped by over 20 percent. The level of false positives and false negatives before and after the review is discussed further in Section 5.

One of the primary drivers for the review was to reduce the level of discontinuity in attrition and completion rates that appeared in 2003 (when NSN was introduced), and also in 2000 (when many institutions changed systems for the introduction of the SDR), and in 1997, (when a student's name was first collected). Therefore, a second approach to the review was to assess the impact on any changes on attrition and completion rates.

In general, rates do not change much over time, or at least, the rate of change is fairly slow and constant. There is certainly evidence to support this premise[6]. However, the tertiary education system

[6] Eg See Tinto, V. (1982). Limits of Theory and Practice in Student Attrition, *Journal of Higher Education, 1982, Vol. 53, No. 6, 687-700.*

in New Zealand is much smaller than many overseas, and was undergoing quite dramatic change in the first part of this decade. This included the dramatic rise and fall in wānanga and international students, funding policy changes impacting in particular, on sub-degree level provision, the signalling of shifts to outcomes-based funding, the introduction of the Performance-Based Research Fund (PBRF), and the introduction of institution-level monitoring. These changes may be expected to impact on rates around this time, independent to data collection changes, or matching methodology changes.

However, the premise used was that any change in rates was likely to be gradual, and the review should act to nearly remove the existing discontinuities in attrition and completion rates in 2003, and if possible around 1999-2000. It was not expected the review would address the quality of matching pre-1998 (ie before names were available as a match field). However, the fact that it did was a bonus. The degree to which the impact on rates was achieved is discussed in Section 4.

In all this analysis, the assumption is made that there are no NSN errors. This assumption may be a reasonable starting point, but is not in fact true, particularly for the first year of NSN in 2003, where in some institutions, new NSNs were assigned for students that already had NSNs (at another institution). Sometimes the same NSN was issued to two different students.

In the 2004 assessment – which was based on 2003 data, differences between SN and NSN were assessed to gauge whether it was the matching in error, or whether NSNs were likely to be in error. The study found that there was a possible error rate of 1.1 percent with the NSN. The false positive error rate was 0.3 percent and the rate of false negative matching was 0.8 percent. That is, in three cases out of 1,000, two different individuals would have the same NSN, while in eight cases out of 1,000, the same individual would have more than one NSN. A higher level of false negatives means higher student counts, higher attrition rates, and lower completion rates. This was beginning to be seen in attrition rates, where rates based on SN were in general higher than those based on NSN.

2004 Assessment of SN and NSN error rates (based on 2003 data)

	Correct	Incorrect
SN	96.4%	3.7%
NSN	98.9%	1.1%

Put another way, the 2004 assessment found that whenever SN and NSN differed, the matching was likely to be in error 75 percent of the time, while NSN was likely to be incorrect 25 percent of the time. However, it is suspected that NSN errors are very highly skewed, in that a relatively small minority of institutions are likely have the large majority of NSN errors.

Also, since 2003, a register has been progressively updated as known cases are found of students with more than one NSN, or NSNs assigned to more than one student. This master-slave register holds the true NSN for each case. There are now over 170,000 entries in this register. This register is used to reassign all NSNs supplied by institutions to their true NSN. With four years of updates now, this is likely to have further reduced the rate of NSN error.

However, some NSN error still exists, especially when large differences at individual institutions are considered. The issue of possible NSN errors is discussed further in Section 5. However, unlike the 2004 assessment, it was decided not to include an assessment of NSN error within the scope of this review. The scope of this review is focused on the quality of the SN. For the purposes of this review, it was considered that the level of NSN errors was, in general, small enough to make the assumption that any differences between SN and NSN would likely be due to matching. Therefore, the error rates presented in this report represent a maximum potential error rate if there were no NSN errors.

3.2 Code changes

The process of comparing SN and NSN gave rise to a number of specific coding changes, over and above the adjustment of match-weights.

The major of these were to do with the way the code handled institutions where student ID numbers had changed. For the majority of students who re-enrol in the same institution, their student ID is the primary and most accurate way to link them. However, it was not uncommon for institutions to change systems over the years and introduce new or modified student ID numbers for the same student. The pattern of change was examined for the major institutions where this occurred. Typically this involved the addition or change of a prefix or suffix to an existing ID. The code comparing student ID was accordingly altered to better detect such changes.

The other major change affected the Join the dots-Weakest link code. The join the dots part of this code logically associates all linked pairs of enrolments into one SN group. For example, if record A matches with record B, and record B matches with record C, then records A, B and C are all grouped – regardless of whether A matches to C or not. These grouped set of records are then tested to see if they all do in fact represent one student, or whether because of a spurious match (a weak link) this group in fact represents two or more students. If a weak link is detected, the group is split accordingly. The logic to assess for any such spurious links was substantially tightened. In particular, more use of NSN was made to help identify if and where any weak links may have occurred. The result was that a number of previously falsely connected records were now correctly split into different groups.

3.3 Data changes

In the four years since the initial matching was developed, there has been substantial activity aimed at improving the quality, completeness and consistency of underlying historical enrolments and completions data, and ongoing improvements also to the SDR collection itself. This included improvements to many of the variables used directly in the matching process. These changes would be expected to improve the quality of the matching independent of any changes to methodology.

Many of these changes related to pre-2000 data where the level of validation was not as highly developed as in the current SDR. Key revisions resulted in improvements for example, to ethnic group and country of citizenship which in the early years used different classifications, and so reduced the overall likelihood of matching. Errors in gender, date of birth and ethnic group were also corrected. Other matching fields that underwent revision included first year in tertiary and last year in secondary school.

The other significant change occurred with the definition of which records were in scope for matching. Data on non-formal enrolments and enrolments in programmes of less than one week’s full-time duration are only available from 2000 on. These record types are not included in the matching and derivation of SNs. A change to the definition of short courses in 2003 saw a large number change from being in scope to being out of scope. This would impact largely on counts and rates for level 1-3 certificates.

4 Impact of changes

One of the primary drivers for the review was to reduce the level of discontinuity in attrition and completion rates that appeared in 2003 (when NSN was introduced), and also in 2000 (when many institutions changed systems for the introduction of the SDR), and in 1997 (when name was first collected).

The impact of the review is assessed by comparing student counts based on the old SN with the counts based on the new SN. In addition to before and after counts, rates of attrition and completion before the review are compared with rates based on the new SN.

4.1 Number of students enrolled in qualifications

The following table shows the number of students enrolled from 1994 to 2006, based on SNs from the old matching method, and SNs based on the new matching method. Both in the old matching and the new matching, the SNs were constrained so that there was a one to one correspondence with NSN from when NSN became available in 2003. Hence, there is no difference in student counts before and after the review for years 2003 onwards.

In this table, students are counted once, regardless of whether they have additional qualification enrolments at the same or different institutions. While the review has changed this count of unique students, the number of qualifications enrolled in before and after the review is unaffected.

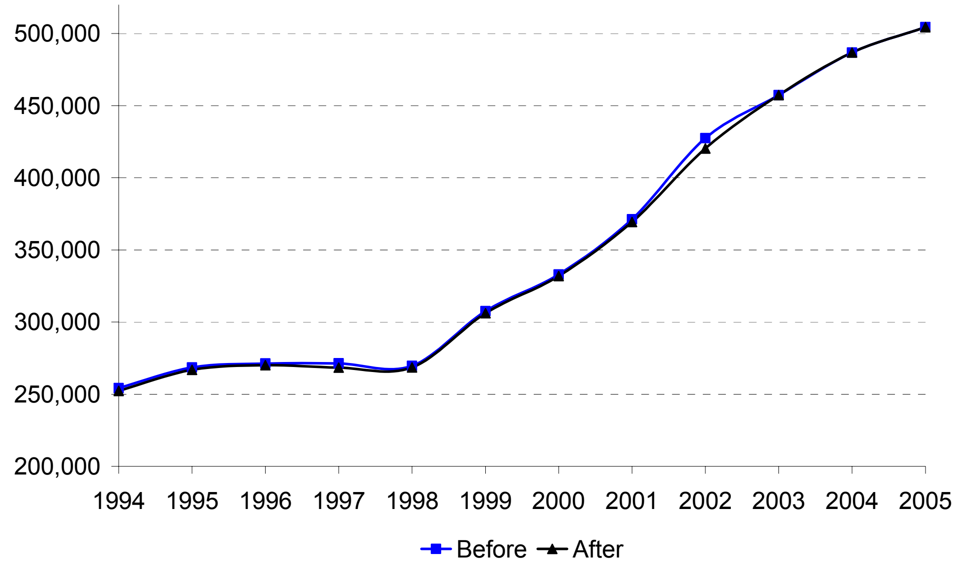
Count of students enrolled – based on SN before the review and after the review

Year	Students before review	Students after review	Difference	
			Number	Percent
1994	254,148	252,269	-1,879	-0.7%
1995	268,517	266,846	-1,671	-0.6%
1996	271,221	270,091	-1,130	-0.4%
1997	271,325	268,496	-2,829	-1.0%
1998	269,687	268,518	-1,169	-0.4%
1999	307,627	306,163	-1,464	-0.5%
2000	333,000	331,875	-1,125	-0.3%
2001	371,310	369,322	-1,988	-0.5%
2002	427,467	420,335	-7,132	-1.7%
2003	457,311	457,311	0	0.0%
2004	486,806	486,806	0	0.0%
2005	504,434	504,434	0	0.0%

The revised matching has resulted in around 1,500-2,000 fewer students each year. With the exception of the 2002 year, this represents a reduction in students enrolled each year of between 0.3 percent and 1.0 percent. This reduction is consistent with one of the aims of the review which was to reduce the level of false negatives, and thereby increase linkages – including those enrolled in different institutions in the same year. In Section 5.1 we examine the quality of these linkages, and will see that the level of false negatives has reduced significantly, while the level of false positives has also decreased.

The 2002 year shows a much higher change in student count. This change is due largely to a single large institution that changed student ID numbering during this year. The revised matching has correctly identified these now as the same student whereas previously they were counted as different students. With this institution removed, the effect is similar as previous years.

Comparison of students enrolled – based on SN before the review and after the review



Count of students enrolled by level of study– before and after the review

Year	Level 1-3 Certs	Level 4 Certs	Diplomas	Bachelors	Level 8 postgrad	Masters	Doctorate	Total	Level 1-3 Certs	Level 4 Certs	Diplomas	Bachelors	Level 8 postgrad	Masters	Doctorate	Total
1994	-121	-76	-198	-141	8	4	0	-1,879	-0.1%	-0.4%	-0.4%	-0.1%	0.1%	0.1%	0.0%	-0.7%
1995	-31	-30	-122	-123	8	6	0	-1,671	0.0%	-0.2%	-0.2%	-0.1%	0.1%	0.1%	0.0%	-0.6%
1996	-87	-38	-142	774	23	31	0	-1,130	-0.1%	-0.2%	-0.3%	0.7%	0.2%	0.3%	0.0%	-0.4%
1997	-233	-49	-375	84	16	26	-1	-2,829	-0.3%	-0.4%	-0.8%	0.1%	0.2%	0.3%	0.0%	-1.0%
1998	-171	-93	-276	198	9	15	1	-1,169	-0.2%	-0.7%	-0.6%	0.2%	0.1%	0.2%	0.0%	-0.4%
1999	-468	-115	-258	217	13	5	2	-1,464	-0.4%	-0.8%	-0.5%	0.2%	0.1%	0.0%	0.1%	-0.5%
2000	-631	-53	-200	521	-17	-3	2	-1,125	-0.5%	-0.3%	-0.4%	0.4%	-0.1%	0.0%	0.1%	-0.3%
2001	-1,289	-46	-172	557	-23	-7	0	-1,988	-0.8%	-0.2%	-0.3%	0.4%	-0.2%	-0.1%	0.0%	-0.5%
2002	-3,599	-91	-219	410	-30	-12	0	-7,132	-1.9%	-0.3%	-0.4%	0.3%	-0.2%	-0.1%	0.0%	-1.7%
2003	0	0	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%	0%	0%
2004	0	0	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%	0%	0%
2005	0	0	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%	0%	0%

In most cases, the revised matching has reduced student counts, by linking more students enrolled in different institutions. If more students are correctly linked, the level of false negatives is lower, and the number of students is also lower.

The 'Total' difference is much greater than the sum of differences at each level. This reflects the increased matching across institutions where the same student may be enrolled but at a different level. For example, in the old matching, a student enrolled in a certificate qualification at one institution and a bachelors degree at another say, has been incorrectly assigned two different SNs. They were counted as one student at certificate level and one student at bachelors level, and two students in total. The new matching has correctly assigned them the same SN. They are still counted as one student at certificate level and one student at bachelors, but now only as one student in the total, instead of two. Hence the differences at each level will generally always be lower than the total difference.

An exception to the pattern of reduced counts occurs at bachelors level, where there has been an increase in students of between 0.1 percent and 0.7 percent. This suggests that unlike other levels, the level of false positives has reduced more than the level of false negatives. It is perhaps less likely that bachelors students, especially those enrolled full-time, are enrolled in other institutions also, and the increase in numbers, may reflect the fact the revised matching and match-weights are making fewer false positive linkages to other bachelors students.

Another way to assess the impact is to look at the number of qualification enrolments per student. This is shown in the table below. The table shows some increase in enrolments per student, in particular from 2000 on. These are now much closer to the enrolments per student from 2003 on, which are based entirely on NSN. Regardless of the method used to count students, there has been a trend of increasing enrolments per student, reflecting the growth of part-time enrolments at sub-degree level.

Average qualifications enrolled in per student – before and after the review

Year	Before review	After review
1994	1.12	1.12
1995	1.13	1.13
1996	1.13	1.13
1997	1.11	1.12
1998	1.12	1.13
1999	1.11	1.12
2000	1.14	1.14
2001	1.15	1.16
2002	1.15	1.17
2003	1.17	1.17
2004	1.19	1.19
2005	1.20	1.20

However, aggregate counts will include net effects of both false negatives and false positives, and may mask larger underlying linkage errors. The quality of the matching at unit record level is discussed further in the Section 5.

4.2 Number of students completing qualifications

As discussed in Section 3, the major revisions made to the matching related to the derivation of SN in enrolments data, and less on the code that derives SN in the completions data. One might expect therefore to see relatively less impact on the count of students completing a qualification before the review and after the review. This is shown in the table below. Since the matching utilises NSN directly from 2003 on, there will be no differences from 2003 on.

Count of students completing qualifications by level of study– before and after the review

Year	Level 1-3 Certs	Level 4 Certs	Dip-lomas	Bach-elors	Level 8 post-grad	Mast-ers	Doct-orate	Total	Level 1-3 Certs	Level 4 Certs	Dip-lomas	Bach-elors	Level 8 post-grad	Mast-ers	Doct-orate	Total
1994	-13	-8	-8	22	0	0	0	-6	-0.1%	-0.4%	-0.1%	0.2%	0.0%	0.0%	0.0%	0.0%
1995	-13	-6	5	4	8	0	0	-4	-0.1%	-0.4%	0.1%	0.0%	0.2%	0.0%	0.0%	0.0%
1996	6	-7	0	181	4	4	0	264	0.0%	-0.5%	0.0%	1.0%	0.1%	0.2%	0.0%	0.6%
1997	-1	-1	-3	45	3	0	0	-42	0.0%	-0.1%	0.0%	0.2%	0.1%	0.0%	0.0%	-0.1%
1998	-31	-1	-2	44	4	10	0	-9	-0.2%	-0.1%	0.0%	0.2%	0.1%	0.4%	0.0%	0.0%
1999	-63	-2	-3	-31	6	4	0	-164	-0.3%	-0.1%	0.0%	-0.1%	0.1%	0.1%	0.0%	-0.3%
2000	-27	-1	-1	31	5	0	0	-68	-0.1%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	-0.1%
2001	-65	0	-4	17	8	0	0	-80	-0.2%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	-0.1%
2002	-150	-4	-3	26	-8	-2	0	-268	-0.4%	0.0%	0.0%	0.1%	-0.1%	-0.1%	0.0%	-0.3%
2003	0	0	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%	0%	0%
2004	0	0	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%	0%	0%
2005	0	0	0	0	0	0	0	0	0%	0%	0%	0%	0%	0%	0%	0%

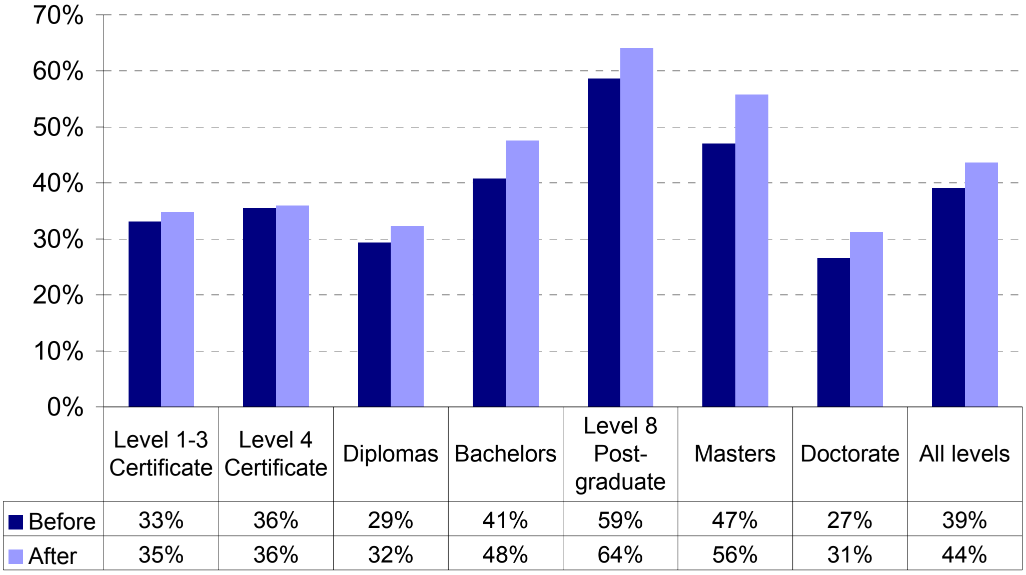
Again, as with enrolments, student counts are lower, reflecting the fact that the matching has made more linkages across institutions. Note that these counts reflect the number of unique students, not the number of qualifications. The number of qualifications completed before and after the review is unaffected.

As with enrolments, the 'total' difference differs from the sum of differences at each level. This occurs for example, when a student completing a certificate at one institution and a bachelors degree at another say, has been incorrectly assigned two different SNs under the old matching. They would have been counted as one student at certificate level and as one student at bachelors level, and as two students in the total. The new matching has now correctly assigned them the same SN. They are still counted as one student at certificate level and one student at bachelors, but now only as one student in the total, instead of two. Hence the differences at each level will generally always be lower than the total difference.

4.3 Completion rates

While the revised matching may have improved linkages across institutions within the same year, the review also aimed to improve linkages across different years, and in particular, aimed to improve retention and completion rates. The graph below shows five-year completion rates by level of study, before and after the review.

Percentage of students starting a level of study in 2001 that have completed a qualification at that level after five years – before and after the review

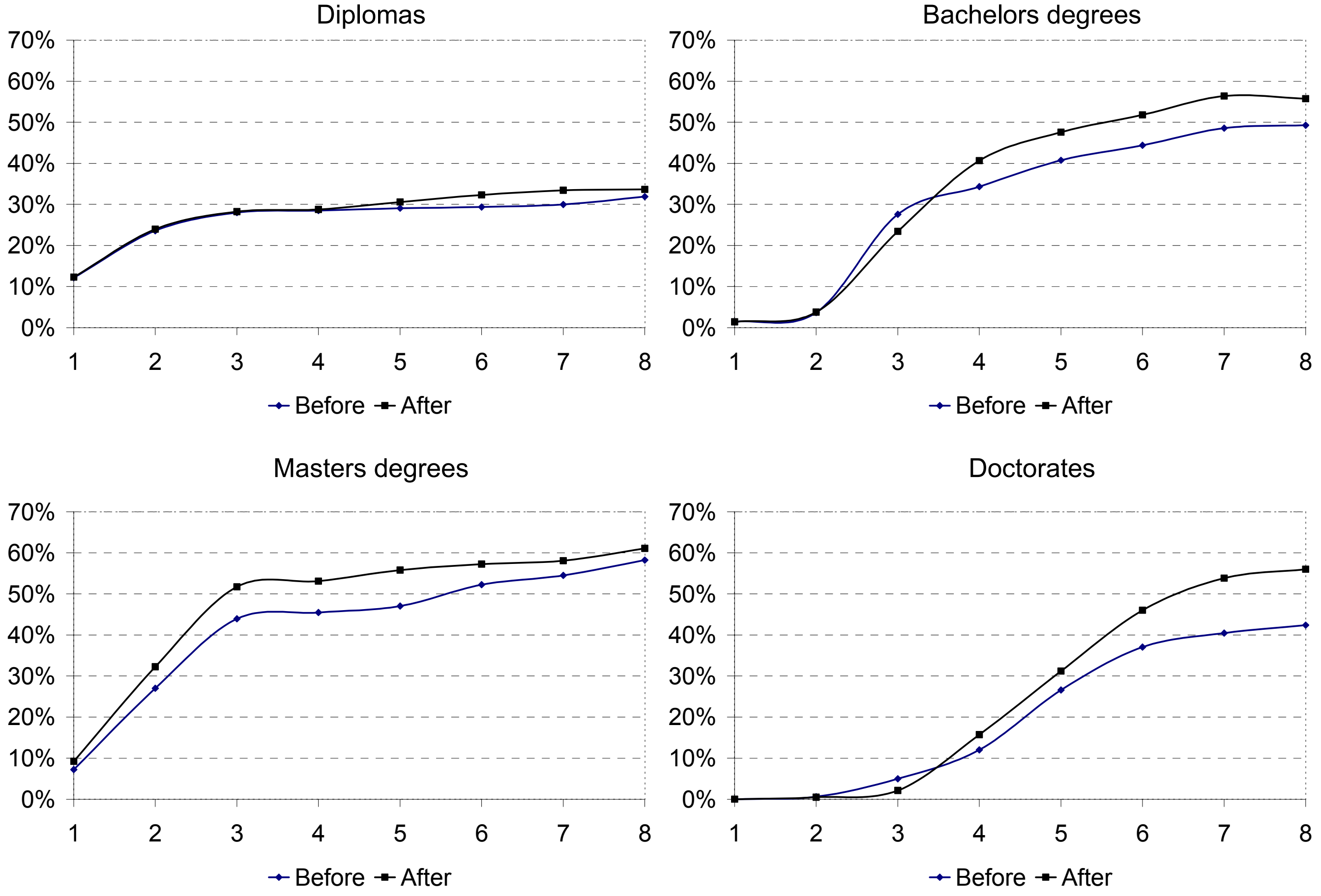


	Level 1-3 Certificate	Level 4 Certificate	Diplomas	Bachelors	Level 8 Post-graduate	Masters	Doctorate	All levels
■ Before	33%	36%	29%	41%	59%	47%	27%	39%
■ After	35%	36%	32%	48%	64%	56%	31%	44%

The total rate of completion for any level started has gone from 39 percent to 44 percent. The major change can be seen at bachelors level and above, where the rates after the review for this particular cohort are four to nine percentage points higher. This is in the direction expected, based on the expectation that the review aimed to reduce false negatives, and thereby increase linkages between students enrolling over several years in longer programmes.

Another way to view the impact is to track completion rates by year of study. The following graphs show this for selected levels of study. In these graphs, a mixture of cohorts is used for convenience. The rate of completion after one year relates to those starting at this level in 2005. The rate of completion after two years relates to those starting at this level in 2004, and so on, up to eight years which relates to those starting in 1998.

Completion rates for selected levels of study by number of years studied – before and after the review



Again, the major changes can be seen for bachelors, masters and doctorate completion rates. The new matching sees increases in eight-year completion rates of six percentage points at bachelors level from 49 percent before the review to 55 percent after the review, from 58 percent to 61 percent for Masters degrees, and notably from 42 percent to 56 percent for PhD's. The increase in the long-term rate is quite significant, and is more comparable to doctorate completion rates in other countries. The size of the change is affected by the smaller size of the cohort of students starting at this level each year (eg around 600 in 1998). At this level if the matching fails to link just six students in any of the following years, this can reduce the completion rate by one percentage point.

The following table provides a full before and after comparison of completion rates by starting year, number of years studied and level of study started.

Completion rates by level of study and number of years studied for 1998 to 2005 starters–before and after the review

	Before									After							
	Level 1-3 Certificates																
	1	2	3	4	5	6	7	8		1	2	3	4	5	6	7	8
1998	26%	30%	32%	33%	34%	35%	35%	36%		26%	30%	32%	33%	34%	35%	36%	36%
1999	27%	31%	32%	34%	35%	35%	36%			27%	31%	33%	34%	35%	36%	37%	
2000	22%	26%	28%	29%	30%	31%				22%	27%	29%	30%	31%	32%		
2001	21%	29%	31%	32%	33%					23%	30%	32%	34%	35%			
2002	20%	29%	31%	33%						21%	31%	33%	35%				
2003	22%	36%	39%							22%	36%	39%					
2004	20%	35%								20%	34%						
2005	25%									25%							

Completion rates by level of study and number of years studied for 1998 to 2005 starters – before and after the review – continued

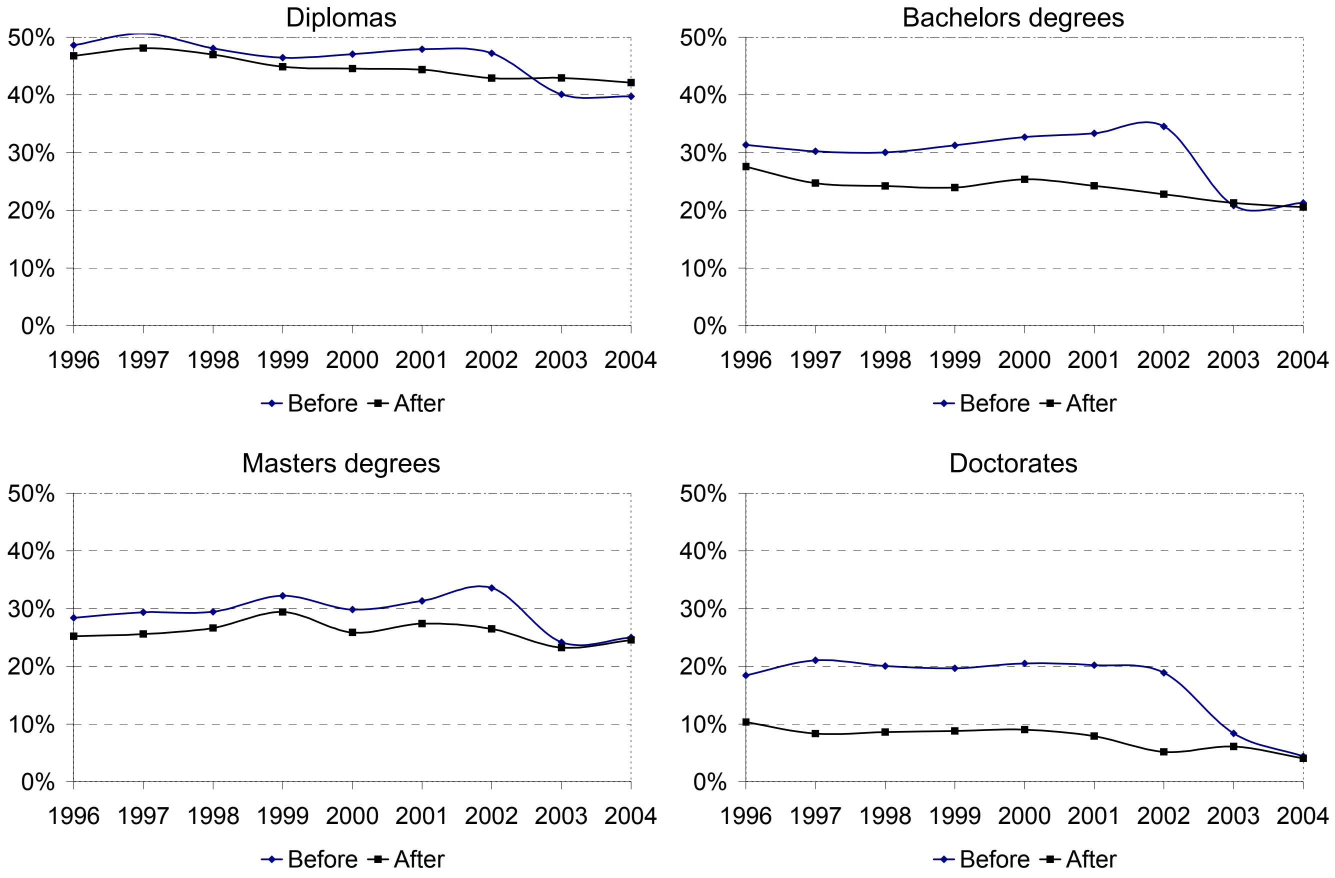
	Before								After							
	Level 4 Certificates															
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1998	14%	20%	22%	23%	23%	23%	23%	24%	20%	26%	28%	28%	29%	30%	30%	30%
1999	18%	23%	25%	26%	26%	26%	27%		22%	26%	29%	30%	30%	30%	31%	
2000	24%	31%	33%	33%	34%	34%			26%	33%	35%	35%	36%	36%		
2001	21%	32%	34%	35%	36%				22%	33%	35%	36%	36%			
2002	22%	36%	37%	38%					25%	38%	40%	41%				
2003	29%	36%	38%						30%	39%	41%					
2004	20%	31%							22%	34%						
2005	23%								24%							
	Diplomas															
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1998	11%	18%	22%	25%	26%	27%	27%	28%	10%	17%	22%	25%	26%	27%	28%	28%
1999	10%	17%	22%	25%	27%	28%	29%		9%	17%	22%	25%	27%	28%	29%	
2000	14%	23%	28%	30%	31%	32%			14%	24%	29%	31%	32%	33%		
2001	13%	22%	26%	28%	29%				14%	23%	28%	31%	32%			
2002	13%	22%	27%	29%					13%	25%	31%	34%				
2003	14%	25%	30%						12%	24%	31%					
2004	13%	24%							13%	24%						
2005	12%								12%							
	Bachelors															
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1998	3%	7%	23%	38%	44%	47%	48%	49%	2%	6%	25%	42%	49%	53%	55%	56%
1999	4%	8%	25%	38%	44%	47%	49%		2%	7%	27%	43%	51%	55%	56%	
2000	4%	7%	22%	35%	41%	44%			2%	6%	24%	39%	48%	52%		
2001	3%	5%	20%	34%	41%				2%	5%	22%	39%	48%			
2002	3%	5%	20%	34%					1%	4%	23%	41%				
2003	5%	11%	28%						1%	4%	23%					
2004	2%	4%							1%	4%						
2005	1%								1%							
	Level 8 Postgraduate															
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1998	27%	44%	53%	56%	57%	57%	58%	58%	26%	44%	53%	57%	58%	59%	59%	60%
1999	27%	45%	54%	57%	58%	59%	59%		26%	45%	54%	57%	59%	60%	60%	
2000	32%	50%	58%	61%	61%	62%			33%	52%	60%	64%	65%	66%		
2001	31%	46%	54%	57%	59%				32%	49%	59%	62%	64%			
2002	27%	43%	50%	53%					31%	49%	58%	62%				
2003	32%	51%	59%						30%	51%	61%					
2004	30%	47%							29%	49%						
2005	29%								30%							
	Masters															
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1998	9%	27%	44%	51%	54%	56%	57%	58%	8%	29%	46%	53%	56%	58%	60%	61%
1999	11%	28%	42%	48%	51%	53%	54%		11%	29%	44%	51%	55%	57%	58%	
2000	9%	25%	41%	47%	50%	52%			10%	28%	44%	51%	55%	57%		
2001	6%	20%	36%	43%	47%				6%	24%	43%	51%	56%			
2002	7%	21%	37%	44%					7%	26%	45%	53%				
2003	8%	27%	45%						6%	29%	52%					
2004	7%	27%							8%	32%						
2005	7%								9%							
	Doctorates															
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1998	1%	3%	6%	14%	23%	33%	39%	42%	0%	1%	4%	14%	28%	40%	49%	56%
1999	1%	2%	4%	11%	22%	32%	40%		1%	1%	4%	13%	28%	42%	54%	
2000	1%	3%	5%	14%	28%	37%			1%	1%	3%	14%	32%	46%		
2001	1%	3%	5%	14%	27%				0%	1%	3%	13%	31%			
2002	0%	1%	3%	12%					0%	1%	3%	16%				
2003	2%	5%	9%						0%	1%	2%					
2004	1%	1%							0%	0%						
2005	0%								0%							

4.4 Attrition rates

Another way to view the impact of the review is to track first-year attrition rates over time. The following graphs compare rates before and after the review for selected levels of study. First-year attrition measures the proportion of students starting a level in a particular year, who did not complete or re-enrol at that level in the following year.

Because only two years are needed to determine first-year attrition rates, differences between NSN-based attrition rates and SN-based attrition rates are able to be seen sooner than for completion rates, which generally require a longer period.

First-year attrition rates for selected levels of study 1996 to 2004 – before and after the review



The lower rates as based on NSN, again indicate the higher rate of false negative errors in the matching before the review. As with completion rates, the major impacts are at degree level and above, in particular, at bachelors and doctorate level, where attrition rates pre-2003 are now more in line with those based on NSN from 2003 on. Trends for all levels are now smoother, although some reduced level of discontinuity still remains at some levels in 1999 and 2000 (reflecting data collection system changes discussed earlier).

The graphs also show improvements in pre-1998 rates, which were not expected to benefit significantly from the review. Whilst information on name is not available in 1994 to 1996 data, a number of other data quality improvements made to this historical data since the original matching was run. For example, data on ethnic group and country of citizenship used different classifications in the first few years, and so were significantly less likely to match with later years. These have now all been recoded to the new classifications. Improved matching code to recognise student ID changes also impacted on several large institutions that changed student ID numbering in these early years.

Previously, rates based on starters before 1998 were not considered robust enough to be published. However, the rates for pre-1998 starters now appear more in line with those starters from 1998 on. This provides an unexpected benefit from the review in terms of being able to provide new information on long-term rates of completion. For example, it highlights the longer time to completion for PhD students, where the rate of completion increases from around 30 percent after five years to around 60 percent after nine and ten years of study.

Differences in first-year attrition rates by level of study 1996 to 2004 – before and after the review

Year	Level 1-3 Certificates			Level 4 Certificates			Diplomas			Bachelors		
	Before	After	Diff-erence	Before	After	Diff-erence	Before	After	Diff-erence	Before	After	Diff-erence
1996	52%	51%	-2%	55%	53%	-2%	49%	47%	-2%	31%	28%	-4%
1997	47%	46%	-2%	53%	49%	-4%	51%	48%	-3%	30%	25%	-6%
1998	48%	46%	-1%	53%	50%	-3%	48%	47%	-1%	30%	24%	-6%
1999	42%	40%	-2%	51%	42%	-9%	46%	45%	-2%	31%	24%	-7%
2000	51%	50%	-2%	49%	44%	-5%	47%	45%	-3%	33%	25%	-7%
2001	44%	42%	-2%	44%	39%	-6%	48%	44%	-4%	33%	24%	-9%
2002	47%	42%	-5%	43%	38%	-5%	47%	43%	-4%	35%	23%	-12%
2003	34%	33%	-1%	38%	37%	-1%	40%	43%	3%	21%	21%	0%
2004	36%	37%	0%	41%	41%	-1%	40%	42%	2%	21%	21%	-1%

Year	Level 8 Postgraduate			Masters			Doctorates			All Levels		
	Before	After	Diff-erence	Before	After	Diff-erence	Before	After	Diff-erence	Before	After	Diff-erence
1996	41%	39%	-2%	28%	25%	-3%	18%	10%	-8%	41%	39%	-2%
1997	39%	35%	-4%	29%	26%	-4%	21%	8%	-13%	38%	35%	-4%
1998	40%	37%	-3%	29%	27%	-3%	20%	9%	-11%	38%	34%	-4%
1999	39%	38%	-1%	32%	29%	-3%	20%	9%	-11%	35%	30%	-5%
2000	34%	31%	-3%	30%	26%	-4%	20%	9%	-11%	42%	38%	-4%
2001	36%	33%	-3%	31%	27%	-4%	20%	8%	-12%	38%	34%	-5%
2002	40%	33%	-7%	34%	26%	-7%	19%	5%	-14%	40%	32%	-8%
2003	31%	30%	-1%	24%	23%	-1%	8%	6%	-2%	26%	27%	1%
2004	33%	31%	-2%	25%	25%	0%	4%	4%	0%	29%	29%	1%

5 Quality

5.1 Quality of SNs assigned to enrolments

The enrolments dataset holds one record for every qualification a student is enrolled in during the year. It is at this level that each SN is assigned. One measure of quality is to estimate what proportion of these enrolment records ends up with an incorrect SN.

We estimate this by comparing one enrolment record with another. If these two have the same NSN then we might expect they should have the same SN, assuming that both NSNs are correct. If the two records don't have the same SN, the matching has failed to make the link. This is a false negative. Alternatively, if the matching has made a link and assigned two records with the same SN, but the NSNs for these records are different, then, again assuming no NSN errors, this indicates a false positive.

Where the matching has correctly determined that two records are not the same, this is a true negative, and where the matching has correctly assigned the same SN to both records this is a true positive.

The matching compares every enrolment record in a year with every other enrolment record, both in the same year, and in other years. NSN data is available for three years of enrolment data, 2003 to 2006. Hence an assessment can be made by comparing every enrolment record in 2003 to 2006 with every other enrolment record in 2003 to 2006.

After all comparisons have been made, a single status of true negative, true positive, false negative or false positive is assigned to each enrolment record's SN value. Where a record is associated with both true positives and true negatives then it is assigned as a true positive. Where a record is associated with both false negatives and false positives then it is assigned a status of false positive. Where a record is associated with both true links and false links then it is assigned a false negative or false positive status depending on the false link.

It is important to note that these rates relate to the SN assigned after all four years have been compared. The first part of the process matches one year's enrolments with another's. The second part of the process then combines the results from each year-pair match and assigns an SN on the strength of the combined information. For the purposes of presenting a single error rate for each year, the results for four comparisons are combined into a single error rate illustrated for the year 2003 in this formula - Error rate for 2003 = (errors from 2003 with 2003) + (errors from 2003 with 2004) + (errors from 2003 with 2005) + (errors from 2003 with 2006) / (number of years compared * enrolments).

This analysis indicates what errors there might have been in 2003 to 2006 data if there were no NSN present. In actual linking, NSN is used directly for these years, so assuming no NSN errors, there is, in theory, no error in SN. Because the quality and completeness of the data prior to 2003 is in general less than that from 2003 on, the error rate estimates provide a reasonable indicator of the quality of the linking for data before 2003, but may understate the prior year error rates slightly.

In all the following analysis, the assumption is made that there are no NSN errors. If SN correspondence is inconsistent with what NSN indicates, then the assumption is made that the matching is in error. This is not true all of the time, particularly for the first year of NSN in 2003, and for some institutions in particular where NSNs were assigned incorrectly. The error rates presented below therefore represent a maximum potential error rate if there were no NSN errors. See section 3.1 for more discussion on the quality of NSN.

Overall error rate

There was an overall potential error rate of 3.6 percent in the enrolments data. This comprised a 2.2 percent false positive error rate and a 1.4 percent false negative error rate.

That is, if NSN was 100 percent correct, then an estimated 22 in every 1,000 enrolments would have incorrectly linked to another enrolment that it should not have (a false positive), while 14 in 1,000 enrolments would have failed to link with at least one other enrolment it should have (a false negative).

Another measure of error is the resulting error in completion and attrition rates. These are discussed later in this section, and in absolute terms are generally lower than the error rates above. For example, the three-year completion rate for first-year students starting any level of study was 43.9 percent based on NSN, and 43.8 percent based on SN.

A review of the original matching method was done in 2004 using the first year of NSN data in 2003. The 2003 enrolments were matched across all other enrolments in 2003 (independently of NSN) and then compared with NSN. When the old error rate from this assessment is reconstructed on the same basis as that used above for the new error rate[7], there was a potential error rate of 9.7 percent in the SN matching. This included a false negative rate of 6.8 percent and a false positive of 2.9 percent.

Comparison of SN error rate before and after the review (2003 enrolments only)

Error type	Old Matching	New Matching	% Improvement
False positive rate	2.9%	2.3%	21%
False negative rate	6.8%	1.3%	81%
Total error rate	9.7%	3.6%	62%

Looking at the same 2003 to 2003 comparison after the review, the new error rate is 62 percent less than the previous error rate. The false negative rate has reduced by 81 percent, while false positives have reduced by 21 percent.

Based on this comparison alone – which is just 2003 with 2003 – the drop in the false negative rate suggests that the review has significantly improved the ability to link two students, without causing any increase in false matches.

However, this is an indicator only of the true quality. It assumes no NSN errors, which is not true, especially for the year, 2003. However, the change from old to new is a good indicator of the impact of the review of the matching code.

The following table shows the new error rate for other years:

Percentage of enrolments with incorrect SNs by year

Year	False negative rate	False positive rate	Total error rate
2003	1.3%	2.3%	3.6%
2004	1.2%	1.9%	3.1%
2005	1.3%	2.2%	3.5%
2006	1.6%	2.4%	4.0%
Over all years	1.4%	2.2%	3.6%

Typically around 55 percent of students enrolled in a year were enrolled in the previous year, while 36 percent were enrolled two years prior, and 28 percent were enrolled three years prior. Around 14 percent of students have more than one enrolment in the same year. The table below shows the percentage of enrolments potentially associated with linking errors, by how many years between the

[7] In the 2004 assessment, an estimate of NSN errors was also made, and these were removed from the SN error rate. In this study, no estimate of likely NSN errors is made. The presented SN error rates are the maximum potential if there were no NSN errors.

two enrolments being compared. There do not appear to be any major differences in quality depending on how many years apart the two enrolment records being compared are.

Percentage of enrolments with incorrect SNs by how many years apart are the enrolments being compared

Years apart	False positive rate	False negative rate	Error rate
Same year	2.1%	1.5%	3.6%
One year apart	2.2%	1.7%	3.9%
Two years apart	2.2%	1.6%	3.6%
Three years apart	2.1%	1.4%	3.5%
Over all years	2.2%	1.4%	3.6%

Linking within institutions and across institutions

Institutions (in most cases) assign unique student IDs to each student which generally do not change for a student over time. While this number is not entirely accurate for all institutions across all years, for most institutions it does provide an accurate way to link students re-enrolling in the same institution.

Because of this, the matching was significantly (eight times) more accurate linking students who were enrolled in the same institution from year to year. Resulting attrition and completion rates are also more accurate at an institution level than at a system level. These are discussed further in section 5.3 and 5.4.

The majority (over 80 percent) of students who re-enrol in the following year, do so at the same institution. For students re-enrolling for a third or fourth year, 63 percent and 53 percent respectively, they do so at the same institution.

Therefore, the availability of student ID provides a very accurate way to link a majority of students. However, it is the ability of the matching to link students enrolling across different institutions, and within institutions where student IDs have changed, that was one of the reasons for the initial development of the matching. Hence, it is of interest to examine the quality when linking both within institutions and across institutions.

Enrolment error rates within and across institutions

Student enrolled in	False positives	False negatives	Total false
Same institution	0.3%	0.2%	0.5%
Different institution	2.6%	1.4%	4.0%
Total	2.2%	1.4%	3.6%

The higher rate of false positives when matching across institutions indicates a tendency for the matching to falsely link two students who have different NSNs. However, this rate will also be affected by the level with which students enrolling in more than one institution have been assigned different NSNs. This will not affect the error rate when linking within institution, but will overestimate the error rate when linking across institutions.

Institutions such as universities, where students are more likely to enrol over more than one year, were more likely to have higher linking errors than other types of institutions, where there is a higher percentage of true negatives in their enrolments because students are less likely to re-enrol.

Enrolment error rates by type of institution (matching within institution)

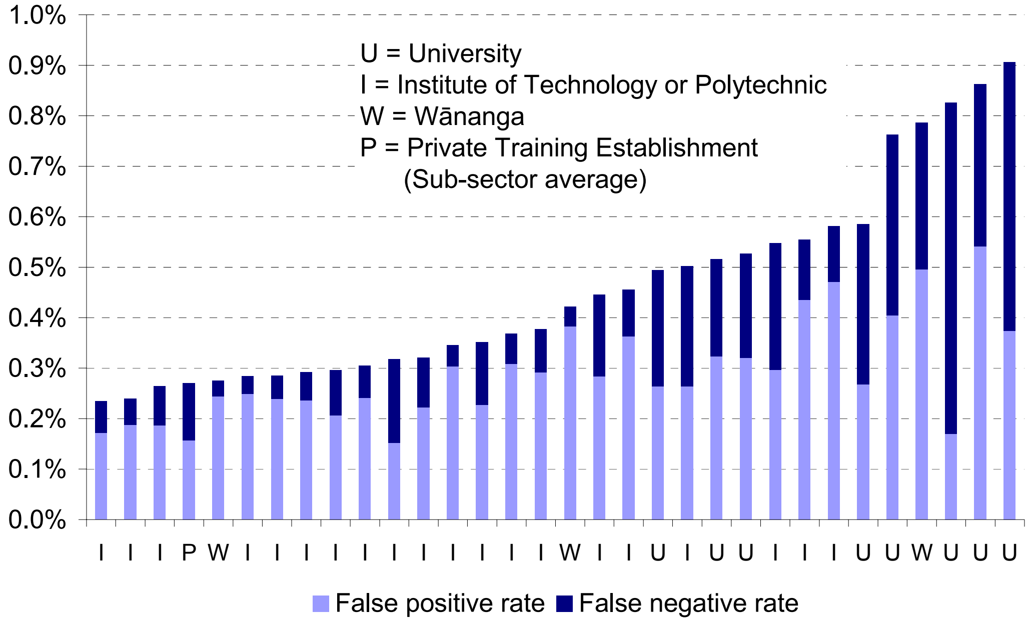
Student enrolled in	Number of institutions	Average error rate	Median error rate	Range
Universities	8	0.7%	0.7%	0.5%-0.9%
Institutes of Technology and Polytechnics	20	0.4%	0.3%	0.2%-0.6%
Wānanga	3	0.5%	0.3%	0.3%-0.8%
Private Training Establishments	301	0.3%	0.1%	0.0%-3.2%
All Institutions	332	0.5%	0.1%	0.0%-3.2%

Note: Excludes Colleges of Education.

A total of four institutions had errors rates over 2 percent. All were Private Training Establishments (PTEs). Within-institution error is important for assessing the likely quality of institution attrition and completion rates (which are only concerned with start and completion at an institution). Matching across institutions is important for assessing the quality of sub-sector or system (ie national) rates of completion of attrition. Institution error rates are therefore unaffected by the degree to which students will enrol in other institutions.

The following graph shows the distribution of SN error rates in enrolments data matching within individual institutions. As the matching works the same for each institution, the size of differences could be an indicator of possible NSN or student ID errors in some institutions.

Enrolment SN error rates by type of institution (matching within institution)



Matching across institutions was eight times more likely to be in error than matching within institution. This assumes that every student enrolled in more than one institution has correctly been assigned the same NSN. Matching within institutions takes no account of what SN or NSN value a student may have in other institutions. So while some institutions may assign the same NSN to a student from year to year, this NSN may be different to an NSN assigned to the same person in another provider. A student is more likely to have NSN errors (ie more than one NSN) when their enrolments over all institutions are considered. Both SN and NSN errors will therefore be higher when matching across institutions.

The following table shows the breakdown of error rates by sub-sector when matching over all enrolments (both across and within institutions).

Enrolment error rates by type of institution (total matching within and across institution)

Student enrolled in	Number of institutions	Average error rate	Median error rate	Range
Universities	8	3.0%	2.9%	2.3%-3.6%
Institutes of Technology and Polytechnics	20	3.6%	3.4%	2.7%-5.6%
Wānanga	3	3.9%	3.6%	2.1%-6.1%
Private Training Establishments	301	6.2%	4.4%	0.0%-33.3%
All Institutions	332	3.5%	3.5%	0.0%-33.2%

Note: Excludes Colleges of Education.

There were 17 institutions with error rates over 10 percent. These were all PTEs, with rates for six of these based on a low denominator (less than 50 enrolments). While PTEs had the highest rates, some PTEs also had very low error rates, and many had no error at all. However on average, their error rates were above other types of institution.

Differences at an aggregate level

Another way to assess the quality of assigned SNs is to compare the aggregate number of students as based on NSN, with the number of students that there would have been in the absence of NSN. This is not a true measure of the quality of the matching, as false positives and false negatives will act to cancel each other, but is useful in terms of assessing the impact of the matching on key indicators such as the count of students.

Difference between counts of enrolled students based on NSN and based on SN by level of study

Level of study	2003	2004	2005	2006	2003	2004	2005	2006
	Number				Percent			
Level 1-3 Certificates	154	227	384	29	0.1%	0.1%	0.2%	0.0%
Level 4 Certificates	-12	3	27	22	0.0%	0.0%	0.1%	0.0%
Diplomas	127	82	93	97	0.2%	0.1%	0.1%	0.1%
Bachelors	856	875	809	723	0.6%	0.6%	0.5%	0.5%
Level 8 Postgraduate	-19	-4	17	15	-0.1%	0.0%	0.1%	0.1%
Masters	11	13	11	3	0.1%	0.1%	0.1%	0.0%
Doctorates	2	2	2	2	0.0%	0.0%	0.0%	0.0%
Total students	3,650	3,690	3,726	2,688	0.8%	0.8%	0.7%	0.5%

Note: Students are counted in each level they enrolled, hence the sum of the rows may not add to the total.

In the absence of NSN, the matching would have resulted in between 2,700 and 3,700 fewer students each year. This represents an undercount of between 0.5 percent and 0.8 percent each year. Differences were largest at bachelors level and at universities (see following table) with 700-900 fewer students based on SN compared with those based on NSN. Note that the matching does not affect the total number of qualifications enrolled in, just the number of students enrolled. This indicates the higher false positive rate in the matching. That is, the matching has a higher tendency to link enrolments as the same student, whereas their NSN indicates they are different people.

Difference in enrolled students between counts based on NSN and counts based on based on SN by sub-sector

Sub-sector	2003	2004	2005	2006	2003	2004	2005	2006
	Number				Percent			
Universities	979	953	926	866	0.6%	0.6%	0.6%	0.6%
ITPs	552	573	624	348	0.3%	0.3%	0.3%	0.2%
Wānanga	-46	12	-11	7	-0.1%	0.0%	0.0%	0.0%
PTEs	46	103	92	32	0.1%	0.2%	0.1%	0.1%
Total Students	3,650	3,690	3,726	2,688	0.8%	0.8%	0.7%	0.5%

Note: Students are counted in each sub-sector they enrolled hence the sum of the rows may not add to the total. Colleges of Education have been included with universities.

The key feature of the above tables is that the sum of the differences at each level is significantly less than the difference in the total number of students. Because a student is counted at each level they are enrolled, but only once in the total, the higher false positive rate will cause connections across different levels of study, thereby causing greater differences in the total. However, it could also indicate that the same student in different institutions has been assigned different NSNs, and the matching has correctly linked them. We know there is some NSN error, and that some students are assigned more than one NSN, particularly in the first years of NSN. But the true level of error is unknown, and it is assumed that the errors in the matching would be greater. If we were to assume that there are no NSN errors, then this indicates a tendency for the reviewed matching code to over-match, that is to match where it should not, rather than under-match.

The following tables show the differences in aggregate student counts by ethnic group, age group and whether domestic or international. All but one is positive (reflecting the higher level of false positive linking). The sole negative difference results from a single institution where students with the same NSN have different student IDs and have been allocated different SNs from the matching.

Difference in enrolled students between counts based on NSN and counts based on based on SN by ethnic group, age group and domestic or international

Ethnic group	Error rate	Age group	Error rate	NZ status	Error rate
European	0.4%	Under 18	0.2%	Domestic	0.3%
Māori	-0.3%	18-19	0.6%	International	0.7%
Pasifika	0.1%	20-24	0.5%		
Asian	0.5%	25-39	0.3%		
Other	0.2%	40+	0.0%		

Of the over 330 institutions with enrolments between 2003 and 2006, 71 percent had no difference in student counts based on NSN or SN over the four years. A further 20 percent had differences of less than either three students, or 0.5 percent of students. A total of eight institutions had differences between 0.6 and 1.1 percent (including seven universities and one polytechnic). In most of these cases, there were more students based on NSN than based on SN. This reflects the higher false positive rate which tends to decrease the number of students.

In summary, 88 percent of the time there was no difference or a difference of three students or less, while in most of the remainder, the matching had identified some students within an institution as the same student even though they had different NSNs. For this to happen, institutions must have assigned different student IDs to two students with the same NSN.

5.2 Quality of SNs assigned to completions

The Ministry of Education also collects data on the number of qualifications completed each year in tertiary education organisations. In conjunction with enrolments, these are used to tell us what the rate of qualification completion is, that is, what percentage of students complete qualifications. The matching also generates an SN for each qualification completion, so that it can be linked back to the enrolments data.

The matching routine for assigning SN to qualification completions is much simpler than for enrolments. Most (but not all) qualification completion records have a corresponding student in the enrolment data, usually in the same institution and in the same year. However, on occasions the matching enrolment relates to a student in a different year or institution. For around 1 percent of completions there is no corresponding student in the enrolment data for any year.

The qualification completions data has not been used historically for funding purposes, and as such, it has not had the same level of quality validation attached to it as have enrolments data. The government has signalled that qualification completion will in future be tied to funding, and in 2007, the Tertiary Education Commission (TEC) began using qualification completion data for institution monitoring. As a result, significant efforts are underway that will improve the future quality of qualification completion data reported in the Single Data Return (SDR).

For the purposes of assessing the quality of the assigned SNs in completion data, each completion record is compared with all the enrolment records with the same NSN. If they all have the same SN as the completion record, then the matching is considered to have worked correctly. If there is more than one SN, then a false negative has occurred, that is, the matching has failed to link the student in the completion data to the student in the enrolment data.

Each completion record is then compared with enrolment records with the same SN. If there is more than one NSN then a false positive has occurred. The matching has incorrectly linked the student in the completion to the student in the enrolment data. Again, this all assumes that there are no NSN errors. As this is not true all of the time, the true error rates will therefore be lower than those presented below.

Again as with enrolments, this analysis indicates what errors there might have been in 2003 to 2006 data if there were no NSN present. In actual linking, NSN is used directly for these years, so assuming no NSN errors, there is no error in SN. Because the quality and completeness of the data prior to 2003 is in general lower than that from 2003 on, the error rate estimates below therefore provide an indicator only of the quality of the linking for data before 2003.

Overall error rate

There was an overall potential SN error rate of 7.8 percent in the completions data. This comprised a 5.9 percent false positive error rate and a 1.9 percent false negative error rate. That is, if NSN was 100 percent correct, an estimated 78 in every 1,000 qualification completions had an incorrect SN. Around 75 percent of errors related to incorrect linkages to a different student (false positives), while a quarter of errors were because the completion failed to link with the right student in the enrolments data (false negative).

The vast majority of students who complete a qualification, complete it in the same institution as they were enrolled in. Less than 2 percent of students completing a qualification have no enrolment ever at the institution they completed at. Institutions assign unique student IDs to each student which generally do not change for a student over time at that institution. For most institutions this provides an accurate way to link students who complete a qualification in the same institution as they enrolled in. Because of this, linking within an institution was noticeably more accurate than across institutions. The error rate for completions matched to enrolments in the same institution was 4.7 percent.

Completion error rates within and across institutions by year and type

Student completing in	2003	2004	2005	2006	False positives	False negatives	Total
Same institution	4.3%	4.4%	3.8%	6.1%	4.1%	0.5%	4.7%
Total	8.5%	7.4%	6.5%	9.2%	5.9%	1.9%	7.8%

However, despite the fact that most completion records should link to an enrolment at the same institution, the error rate for SN in the completions data was higher than for enrolments where the matching is more complex. In particular, the level of false positives in the completions linking is much higher than in the enrolments linking.

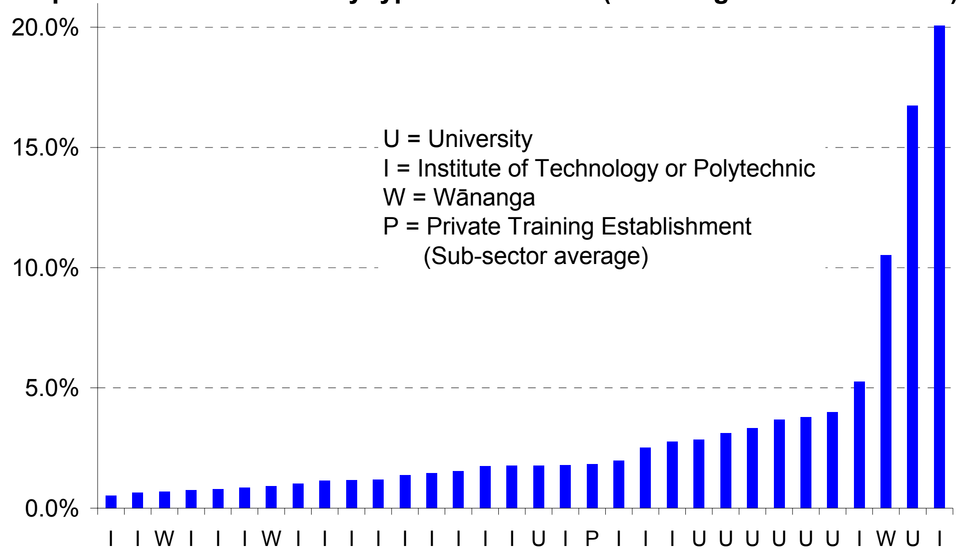
One difference between the two is that for virtually every student in the completions data we expect to see that student somewhere in the enrolments data. The level of true negative matching should be near zero. However, not all students re-enrol, so the level of true negative matching in enrolments data is much higher (around 40 percent for the years 2003-2006). Establishing that an SN in the enrolment data is a true negative is much simpler than establishing whether it is a true positive.

In the analysis so far, the false positive rate has been defined as the percentage of all records that is associated with a false link or mismatch. Another way to compare the quality of the two matching routines is to compare the percentage of positive links, rather than total records, that are false. For enrolments this is 3.9 percent, while for completions it is 4.2 percent linking between the same institution. In this sense, the level of false positives are much closer.

Another factor affecting the increased error rate in completions data is that the distribution of error rates is highly skewed by institution. While the average error rate was 7.8 percent, the median (the rate at which half of institutions are above or below) was 4.8 percent. Amongst Tertiary Education Institutions (TEIs) the average was 6.3 percent and the median was 5.2 percent. One quarter of all errors was in a single institution, while just four institutions accounted for half of all errors. If these institutions are removed the error rate reduces to 5.7 percent.

When matching within institutions, distribution of error rates remain highly skewed. In fact, just four institutions (including 2 polytechnics, one university and one wānanga) account for 66 percent of potential errors. Put another way, if these four institutions are removed the error rate reduces to 2.2 percent.

Completion SN error rates by type of institution (matching within institution)



Matching within institution relies on the institution's student ID. If the student ID on the completion file matches the student ID on the enrolment file then in most cases the SN from the enrolment file is assigned to the completions data. While the matching relies heavily on the institution number, it largely works the same for all institutions. The highly skewed nature of these results therefore indicates potentially significant quality issues with the completions data submitted for those institutions whose error rates differ significantly from the median. This in turn will impact on the quality of attrition and completion rates, which are discussed in Sections 5.3 and 5.4 below.

For institutions to have error rates that are significantly different from the median implies a mismatch between an institution's enrolment and completion file. An example of this occurs when institutions report completions for which there are no matching student (using student ID or NSN) in their enrolment data (in this year or past years), and the matching has then also incorrectly linked that completion to another student in the enrolments data. This will also occur when the student ID or NSN assigned to a student in the completions data is different to that assigned to them in the enrolments data.

Within institution error is important for assessing the likely quality of institution attrition and completion measures (which are only concerned with start and completion at an institution). Matching within and across institutions is important for assessing the quality of sub-sector or system (ie national) rates of completion or attrition. Institution error rates are therefore affected by the degree to which students transfer to other institutions.

The table below shows average institution error rates by sub-sector. In general, ITPs and two of the three wānanga had lower error rates, while universities tended to have higher error rates. Overall, PTEs had the lowest error rates, but they ranged from the lowest to the highest rates. Two thirds had an error rate under 1 percent, and 23 had error rates over 5 percent.

Completion error rates by type of institution (matching within institution)

Student enrolled in	Number of institutions	Average error rate	Median error rate	Range
Universities	8	4.9%	3.6%	1.8%-16.7%
ITPs	20	2.5%	1.4%	0.5%-20.1%
Wānanga	3	4.0%	0.9%	0.7%-10.5%
PTEs	278	1.8%	0.1%	0.0%-40.7%
All Institutions	309	4.7%	0.5%	0.0%-40.7%

Note: Excludes Colleges of Education.

In each case, the average error rate is higher than the median, indicating a skewed distribution, with a relatively smaller number of institutions with larger error rates.

Differences at an aggregate level

The table below shows the difference in the number of students completing qualifications based on SN and based on NSN by level of study for the years 2003 to 2006. Again at an aggregate level, false positives (leading to undercount) can cancel false negatives (leading to overcount), and so the impact of matching errors on aggregate totals can be masked. However, in terms of assessing the impact of the matching on key indicators such as the count of students gaining qualifications, this comparison is useful.

Difference in completing students between counts based on NSN and counts based on based on SN by level of study

Level of study	2003	2004	2005	2006	2003	2004	2005	2006
	Number				Percent			
Level 1-3 Certificates	126	533	339	515	0.3%	1.0%	0.5%	1.0%
Level 4 Certificates	668	16	36	-10	3.5%	0.1%	0.2%	-0.1%
Diplomas	18	25	29	36	0.1%	0.2%	0.2%	0.2%
Bachelors	125	117	114	239	0.5%	0.4%	0.4%	0.8%
Level 8 Postgraduate	-18	-17	-12	3	-0.3%	-0.3%	-0.2%	0.0%
Masters	7	10	28	16	0.2%	0.3%	0.7%	0.4%
Doctorate	0	0	0	0	0.0%	0.0%	0.0%	0.0%
Total students	1,101	1,115	987	1,138	1.0%	0.9%	0.7%	0.9%

Note: Students are counted in each level they enrolled hence the sum of the rows may not add to the total.

If NSN were not available, the matching would have resulted in between 1,000 and 1,100 fewer students completing qualifications each year. This represents an undercount of between 0.7 percent and 1.0 percent of the total number of NSN-based students each year. As with enrolments, differences were largest at bachelors level and at universities. Note that the matching does not affect the total number of qualifications completed, just the number of students completing them. As with enrolments, the matching is has a higher tendency to link completions to the same SN, whereas their NSN indicates they are different.

Again as with enrolments, the sum of the differences at each level is significantly less than the difference in the total number of students. Because a student is counted at each level of enrolment, but only once in the total, the higher false positive rate will cause connections across different levels of study, thereby causing greater differences in the total. However, it could also indicate that the same student in different institutions has been assigned different NSNs, and the matching has correctly linked them. We know there is some NSN error, and that some students are assigned more than one NSN, particularly in the first years of NSN. However, the true level is of NSN error is unknown, and it is assumed that the errors in the matching would be greater. If we assume that there are no NSN errors, then the results indicate a tendency for the reviewed matching to over-match, that is to match where it should not, rather than to under-match.

As mentioned earlier, the distribution of errors is very highly skewed. Ten institutions account for over 90 percent of the differences. Because the matching works the same for each institution, that fact that differences are not that evenly distributed indicates potential data quality differences with particular institutions' reporting of completions data.

5.3 Completion rates

Completion rates were calculated using SN and compared with the same rate using NSN. Two types of rates were compared, institution rates and system rates. The institution rate counts only those students completing at the same institution as the one they started. This is the concept used for Baseline Monitoring Reports produced by the TEC[8]. System rates include students who transfer and complete at a different institution and are therefore generally higher than institution rates. This is the basis of the completion rate statistics published by the Ministry of Education.

Institution completion rates

With four years of NSN data available, we are able to derive three-year completion rates. Three-year institution completion rates are determined for those who indicated they were first-year students in 2003 or 2004. Aside from telling us whether a student has completed a qualification or not, unique student numbers (SN or NSN) are needed to determine when a student started a level of study, and hence whether they belong in the starting cohort. That is, SN or NSN are used to determine the size of both the numerator and denominator in the completion rate calculation.

Because we have NSN data only back to 2003, we can only accurately determine the year a student started a particular level of study for first-year students. Using first-year students only, effectively limits available cohorts to certificate, diploma and bachelors starters.

All completion rates include students completing a qualification at a higher level than the one they started at. For example, if a student starts a diploma, and graduates with a bachelors degree, they are counted in the completion rate.

The table below shows the percentage of first-year students who have completed a qualification at the same institution as the one they started in, three years after starting. The rates represent the *average* over all institutions in each sub-sector, and average the results for both 2003 and 2004 cohorts. The average and median differences represent absolute magnitudes, ie a difference of -0.5 is treated as 0.5. Any group with less than five starters has been removed because of the distortion that resulting rates can have on averages. The resulting differences are the differences that would have occurred in the absence of NSN. In the actual data, there is no difference as the matching is constrained so that there is a one to one correspondence between SN and NSN.

[8] The definition for institution completion rate that is used by the TEC in its Baseline Monitoring Reports differs slightly from that used in this review. For detailed information on definitions and methods used by the TEC refer to the TEC's Baseline Monitoring Report Methodology papers.

Comparison of NSN- and SN-based three-year institution completion rates (first-years only)

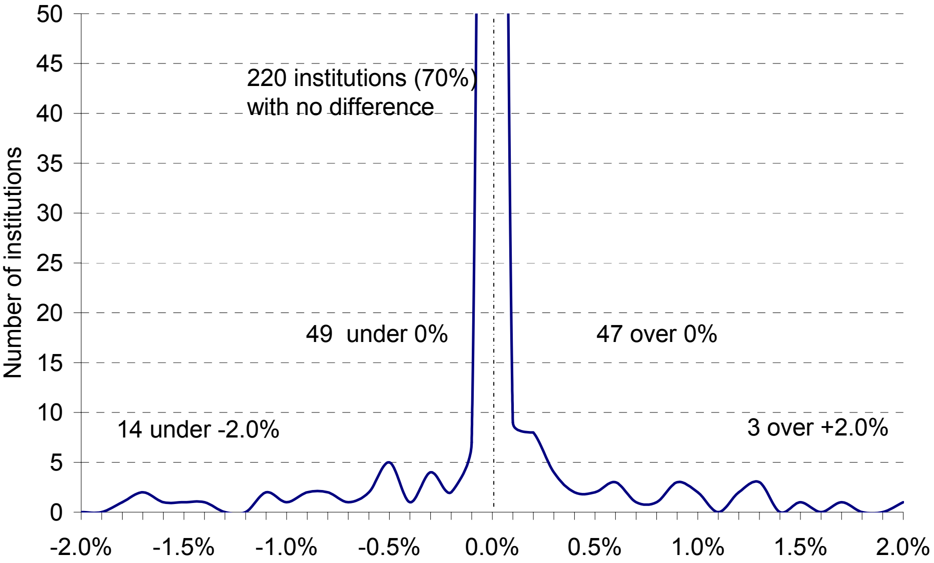
Level of study	Sub-sector	Based on NSN	Based on SN	Average difference	Median difference	Range of differences
				in percentage points		
Level 1-3 Certificates	Universities	44.9%	45.0%	0.7%	0.6%	0.0%-1.8%
	ITPs	35.2%	35.2%	0.3%	0.2%	0.0%-1.8%
	Wānanga	63.3%	63.6%	0.5%	0.5%	0.5%-0.6%
	PTEs	48.0%	47.8%	0.5%	0.0%	0.0%-8.7%
	All	45.9%	45.8%	0.5%	0.1%	0.0%-8.7%
Level 4 Certificates	Universities	47.6%	47.7%	0.6%	0.6%	0.0%-1.2%
	ITPs	27.8%	27.8%	0.3%	0.1%	0.0%-3.0%
	Wānanga	52.9%	50.7%	2.9%	2.2%	0.7%-5.2%
	PTEs	44.1%	43.0%	1.6%	0.0%	0.0%-35.3%
	All	37.7%	37.1%	1.0%	0.1%	0.0%-35.3%
Diplomas	Universities	31.5%	32.4%	1.1%	0.5%	0.0%-4.1%
	ITPs	24.9%	24.9%	0.2%	0.2%	0.0%-0.6%
	Wānanga	31.1%	30.9%	0.4%	0.4%	0.3%-0.5%
	PTEs	39.4%	39.5%	0.5%	0.0%	0.0%-3.2%
	All	33.1%	33.3%	0.5%	0.2%	0.0%-4.1%
Bachelors	Universities	22.1%	22.6%	0.6%	0.3%	0.1%-3.2%
	ITPs	25.6%	25.6%	0.1%	0.0%	0.0%-0.6%
	Wānanga	39.8%	41.3%	1.5%	1.5%	1.5%-1.5%
	PTEs	23.7%	23.7%	0.0%	0.0%	0.0%-0.0%
	All	24.9%	25.1%	0.3%	0.1%	0.0%-3.2%
Any level	Universities	28.6%	29.1%	0.6%	0.3%	0.1%-3.2%
	ITPs	32.3%	32.3%	0.2%	0.2%	0.1%-1.5%
	Wānanga	52.9%	52.7%	0.5%	0.5%	0.2%-0.8%
	PTEs	46.7%	46.5%	0.5%	0.0%	0.0%-10.7%
	All	43.9%	43.8%	0.5%	0.1%	0.0%-10.7%

Note: The rates for each specific level relate to the percentage of first-year students that have completed at that level or higher after 3 years. The rate for 'Any level' however refers to those who have completed at any level of study after 3 years. The rate is an average for the years 2003 and 2004 over all institutions with more than five students. Colleges of Education have been included with the respective university they merged with.

For example, for bachelors degrees, when the three-year completion rates are averaged over all institutions offering degrees, the three-year completion rate was 22.1 percent based on NSN, and 22.6 percent when based on SN. When the magnitude of the difference is taken, regardless of whether the SN-based rate is less than or more than the NSN-rate, the average difference for institutions was 0.6 percentage points.

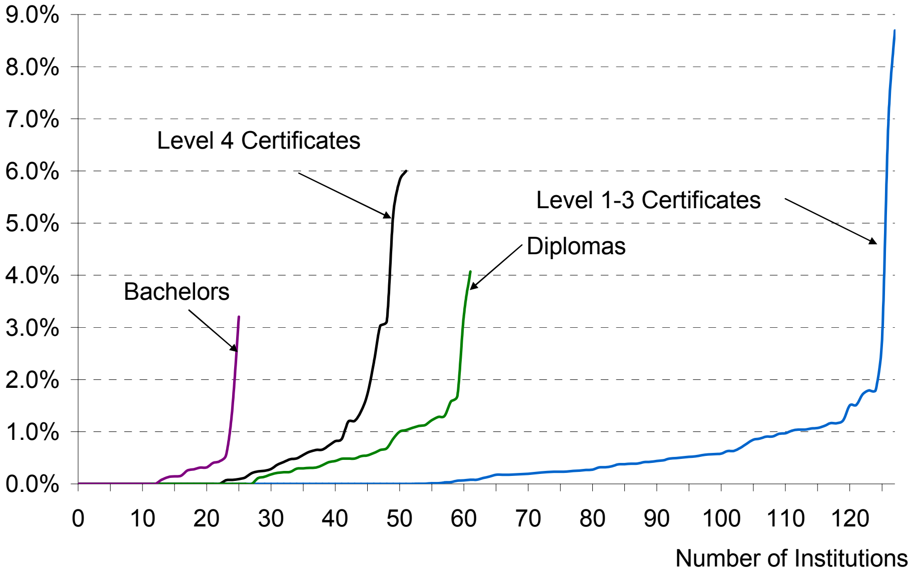
Differences over all institutions and levels of study ranged from no difference to 35.3 percentage points for a level 4 completion rate for one PTE. Of 315 institutions, about 70 percent had nil difference in rates, when based on SN or NSN. There did not appear to be a tendency for the matching to either over- or under-estimate three-year institution completion rates. About 15 percent of institutions had higher completion rates when NSN was used, and about 15 percent had lower rates. Higher rates with NSN mean that the matching failed to make connections to completions records, while lower completion rates meant that SN was incorrectly linking to future completion records. Of the 17 institutions with differences over 2 percent, most were small PTEs with less than 30 first-year students. Just one university, one polytechnic and one wānanga had differences of more than 2.0 percent.

Distribution of differences between NSN- and SN-based three-year institution completion rates (first-years only)



By considering the absolute magnitude of the difference only, we can see the highly skewed nature of the differences across institutions. The graph below shows that differences for the majority of institutions are low, while for a few they are quite high.

Distribution of differences between NSN- and SN-based three-year institution completion rates (first-years only)



The large differences for only a small number of institutions and the normal distribution of differences suggest specific data quality issues with these institutions' data returns, rather than systemic issues with the derivation of SN. With more systemic matching issues, we might expect to see a more uniform distribution of differences. Either these institutions have incorrectly assigned NSNs (and the SNs are correct), or there are incorrect student IDs or other errors in the fields used for matching, that have resulted in an incorrect SN being assigned.

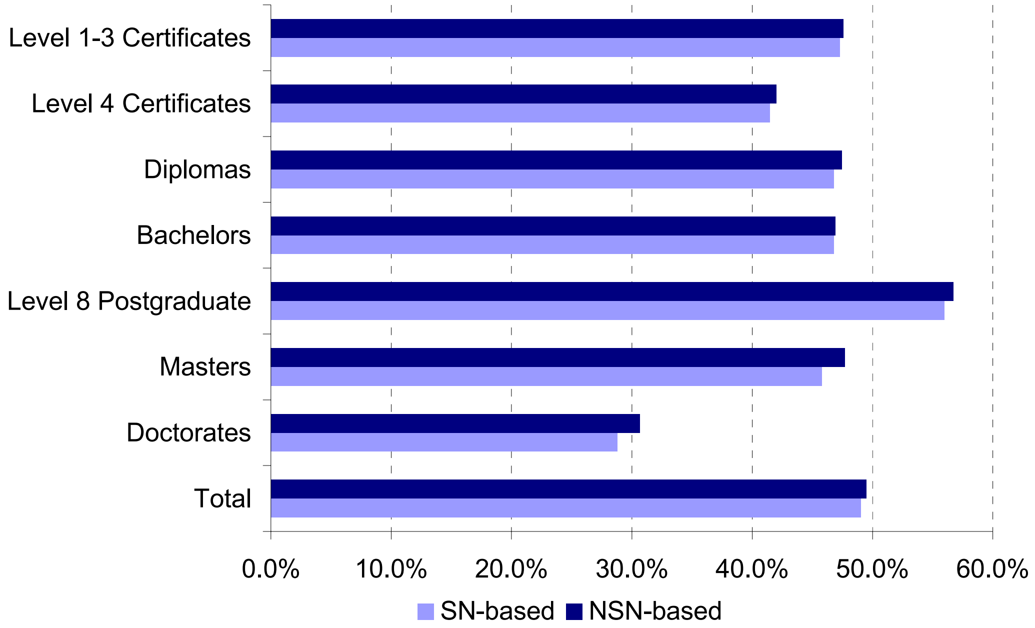
Because the aim of this report is to assess the matching, and because first-year students generally represent a minority of all students matched, SN-based rates are then also assessed using all an institution's students, regardless of how many years they studied. Here, the rate is defined as the proportion of all students enrolled (rather than just those starting), that have completed three years later. By extending the rates comparisons to all students, we can assess the quality of the matching over the full range of students and levels of study. However, in doing this, we are mixing cohorts of starters with those in their second, third or later years of study. So while the absolute value of the rate may be less meaningful, the difference between the two nevertheless provides a good indicator of the matching across the full range of students.

Comparison of NSN- and SN-based three-year institution completion rates (for all students enrolled in 2003 and 2004)

Level of study	Based on NSN	Based on SN	Average difference
Level 1-3 Certificates	47.6%	47.3%	0.6%
Level 4 Certificates	42.0%	41.5%	0.9%
Diplomas	47.5%	46.8%	0.9%
Bachelors	46.9%	46.8%	0.3%
Level 8 Postgraduate	56.7%	56.0%	0.8%
Masters	47.7%	45.8%	1.9%
Doctorates	30.7%	28.8%	1.8%
Any level	49.5%	49.0%	0.7%

Note: The rates for each specific level relate to the percentage of students (regardless of how many years enrolled) that have completed at that level or higher after three years. The rate for 'Any level' however refers to those who have completed at any level of study after three years. The rate is an average for the years 2003 and 2004 over all institutions with more than five students.

Comparison of NSN- and SN-based three-year institution completion rates (for all students enrolled in 2003 and 2004)



When considered over all students – rather than first-year students – the matching shows a slight tendency to understate three-year completions.

For bachelors-level study, rates differences ranged from no difference to 6.5 percentage points for a small PTE. However, there are only five institutions (including one university) with differences between 1.0 and 2.0 percentage points. All other institutions have differences of less than one percentage point. At masters level, there were seven institutions with large differences in rates – ranging from 2.0 percentage points to 7.3 percentage points. At PhD level, differences range from zero to 4.5 percentage points – with an average of 1.8 and a median of 1.5 percentage points. These are shown in the table below.

Difference between NSN-based and SN-based three-year institution completion rates by sub-sector and level of study (for all students enrolled in 2003 and 2004)

Level of study	Unis	ITPs	Wānanga	PTEs	All
Level 1-3 Certificates	0.6%	0.2%	0.6%	0.6%	0.3%
Level 4 Certificates	0.6%	0.3%	1.5%	1.1%	1.0%
Diplomas	0.3%	0.2%	0.4%	1.2%	0.2%
Bachelors	0.4%	0.2%	0.5%	0.5%	0.1%
Level 8 Postgraduate	1.6%	0.3%	0.0%	0.1%	0.3%
Masters	3.0%	1.0%	0.0%	0.9%	2.8%
Doctorates	1.8%				1.8%
Any level	0.6%	0.1%	0.4%	0.8%	0.1%

Note: The rates for each specific level relate to the percentage of students (regardless of how many years enrolled) that have completed at that level or higher after three years. The rate for 'Any level' however refers to those who have completed at any level of study after three years.

System completion rates

System completion rates include students who transfer and complete at a different institution. These generally will be higher than institution rates, which count transfer as non-completion. System rates based on SN are more likely to be different from corresponding NSN-based rates, because of the higher error rate associated with matching across institutions. The following table shows the difference in three-year completion rates by level of study for first-year students in 2003 and 2004.

Comparison of NSN- and SN-based three-year system completion rates (first-year students)

Level of study	Based on NSN	Based on SN	Difference
Level 1-3 Certificates	42.0%	42.0%	0.0%
Level 4 Certificates	40.1%	38.2%	1.9%
Diplomas	32.2%	32.7%	-0.5%
Bachelors	21.5%	22.3%	-0.8%
Any level	36.8%	36.9%	-0.1%

Note: The rates for each level relate to the percentage of first-year students summed over all institutions more than five starters that have completed at that level or higher after three years. The rate for 'Any level' however refers to those who have completed at any level of study after three years.

These rates sum up starters and completers over all institutions. They are not averages of institution rates as used in the institution rates above. Hence, the total system completion rate appears, counter-intuitively perhaps, lower than the average institution completion rate. This is because the average institution rate is increased by a large number of smaller institutions with higher rates, while the system rate is lowered by a number of larger institutions with lower rates.

As with institution rates, by extending the basis to all students enrolled, regardless of how many years of enrolment, we can assess the difference in three-year system completion rates based on NSN and SN. The pattern of difference for system rates follows that of institution rates, with larger differences for Masters and PhD completion rates. This is shown in the following table.

Comparison of NSN- and SN-based three-year system completion rates (all students)

Level of study	Based on NSN	Based on SN	Difference
Level 1-3 Certificates	42.0%	42.3%	-0.3%
Level 4 Certificates	41.4%	40.5%	0.9%
Diplomas	40.4%	40.6%	-0.2%
Bachelors	49.7%	49.8%	-0.1%
Level 8 Postgraduate	62.0%	61.7%	0.3%
Masters	59.1%	56.3%	2.8%
Doctorates	32.0%	30.1%	1.9%
Any level	44.4%	44.3%	0.1%

5.4 Attrition rates

The table below shows the percentage of first-year students in 2003 or 2004 that did not complete a qualification at the institution they started and did not re-enrol at that same institution in the following year[9]. The rates represent the average institution first-year attrition in each sub-sector, and averages the results for 2003 and 2004 students. Rates based on SN were compared with rates based on NSN. The resulting differences shown below represent absolute magnitudes, ie a difference of -0.5 is treated as 0.5. These are the differences that would have occurred in the absence of NSN. In the actual data there is no difference, as the matching is constrained so that there is a one to one correspondence between SN and NSN. Any group with less than five starters has been removed because of the distortion that resulting rates can have on averages.

Comparison of NSN- and SN-based first-year institution attrition rates

Level of study	Sub-sector	Based on NSN	Based on SN	Average difference	Median difference	Range of differences
				in percentage points		
Level 1-3 Certificates	Universities	34.3%	33.9%	0.7%	0.7%	0.0%-1.2%
	ITPs	47.1%	46.8%	0.3%	0.3%	0.1%-1.1%
	Wānanga	20.9%	20.6%	0.5%	0.5%	0.4%-0.6%
	PTEs	35.0%	34.9%	0.6%	0.1%	0.0%-5.3%
	All	36.6%	36.5%	0.5%	0.2%	0.0%-5.3%
Level 4 Certificates	Universities	36.2%	35.7%	0.8%	0.7%	0.5%-1.3%
	ITPs	39.9%	39.5%	0.5%	0.3%	0.0%-3.3%
	Wānanga	42.0%	42.5%	1.7%	1.7%	0.7%-2.7%
	PTEs	35.9%	35.9%	1.0%	0.3%	0.0%-14.7%
	All	37.8%	37.7%	0.8%	0.4%	0.0%-14.7%
Diplomas	Universities	39.2%	38.6%	0.9%	0.5%	0.3%-1.9%
	ITPs	46.1%	46.0%	0.4%	0.4%	0.0%-1.1%
	Wānanga	46.1%	45.9%	0.3%	0.3%	0.2%-0.4%
	PTEs	30.2%	30.3%	0.6%	0.3%	0.0%-2.0%
	All	37.3%	37.3%	0.6%	0.4%	0.0%-2.0%
Bachelors	Universities	18.4%	18.1%	0.6%	0.4%	0.1%-1.5%
	ITPs	34.8%	34.7%	0.3%	0.2%	0.0%-1.3%
	Wānanga	41.3%	41.3%	0.0%	0.0%	0.0%-0.0%
	PTEs	45.4%	45.7%	0.3%	0.0%	0.0%-0.7%
	All	30.6%	30.4%	0.4%	0.2%	0.0%-1.5%
Any level	Universities	22.3%	22.0%	0.6%	0.4%	0.2%-1.3%
	ITPs	45.8%	45.6%	0.3%	0.3%	0.1%-1.4%
	Wānanga	29.1%	29.0%	0.6%	0.5%	0.2%-1.0%
	PTEs	33.4%	33.4%	0.5%	0.1%	0.0%-5.4%
	All	34.4%	34.3%	0.5%	0.2%	0.0%-5.4%

Note: The rates for each specific level relate to the percentage of first-year students that have not completed at that level or re-enrolled at the same institution in the following year. The rate for 'Any level' however, refers to those who have not completed at any level of study or re-enrolled in the following year. Colleges of Education have been included with the respective university they merged with.

For example, for bachelor degrees, the average difference between SN-based and NSN-based institution first-year attrition rates was 0.6 percentage points, and the median was 0.2 percentage points. Differences were slightly more likely to be positive than negative, with 19 percent of institutions with error rates greater than 0, and 11 percent of institutions with error rates less than 0. A positive difference is consistent with the higher false positive rate in the completions linking.

As with completion rates, the distribution of the absolute magnitude of the differences was highly skewed. Over 70 percent of institutions had no difference in rates, 20 percent had less than one

[9] The definition for attrition that is used by the TEC in its Baseline Monitoring Reports differs slightly from that used in this review. For detailed information on definitions and methods used by the TEC refer to the TEC's Baseline Monitoring Report Methodology papers.

percentage point difference, and 10 percent had over one percentage point difference. Some of these are based on small numbers, however, there still remain a few large institutions with large differences. These institutions are generally the same as those which had large differences in completion rates. As discussed earlier, the large differences for a small number of institutions suggest specific data quality issues with these institutions' data. Either these institutions have incorrectly assigned NSNs (and the SNs are correct) or there are incorrect student IDs or other errors in the fields used for matching, that have resulted in an incorrect SN being assigned.

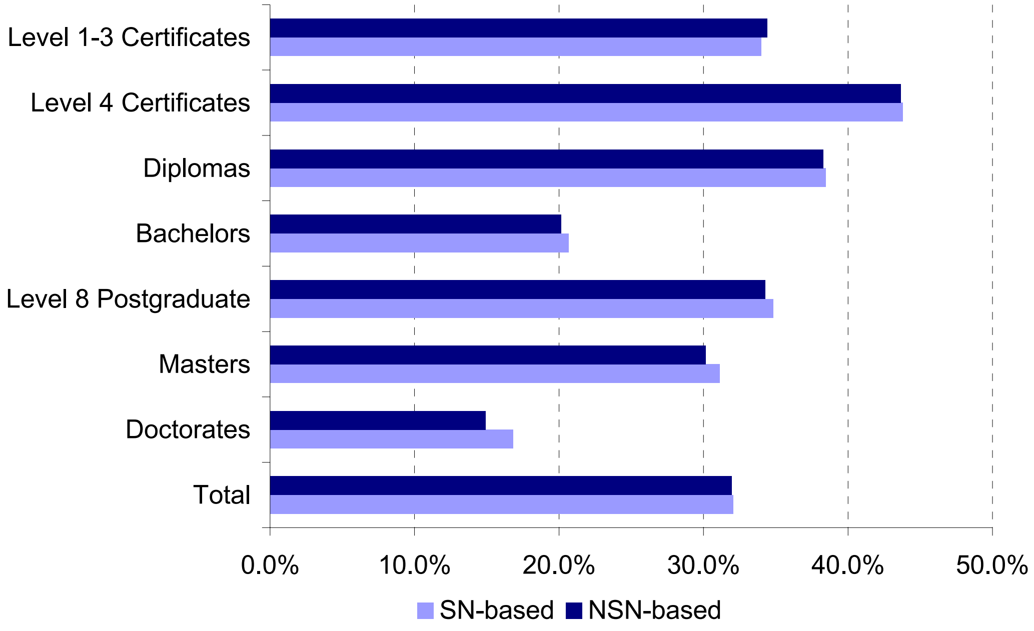
Attrition rates are now assessed using all of an institution's students, regardless of how many years they have studied. By including all students, rather than first-year students, we can assess the quality of the matching over the full range of students and levels of study. However, in doing this we are mixing cohorts of starters with those in the second, third or later years of study. So while the absolute value of the rate may be less meaningful, the difference between the two nevertheless provides a good indicator of the matching over the full range of students.

Comparison of NSN- and SN-based one-year institution attrition rates (all 2003 and 2004 students)

Level of study	Based on NSN	Based on SN	Difference
Level 1-3 Certificates	34.4%	34.0%	0.4%
Level 4 Certificates	43.7%	43.8%	-0.1%
Diplomas	38.3%	38.5%	-0.2%
Bachelors	20.1%	20.7%	-0.5%
Level 8 Postgraduate	34.2%	34.8%	-0.6%
Masters	30.1%	31.1%	-1.0%
Doctorates	14.9%	16.8%	-1.9%
Any level	31.9%	32.1%	-0.1%

As with completion rates the largest differences occur with masters and doctorates. This partly reflects the smaller base of students enrolled at these levels. Consistent with the higher NSN-based completion rates, the NSN-based attrition rates are slightly lower than the SN-based rates, reflecting the higher false positive rate associated with the completions matching.

NSN- and SN-based one-year institution attrition rates (all 2003 and 2004 students)



Bibliography

The following is a sample of research and analytical reports that use SN, and so rely on the ability of the matching to link students in historical tertiary enrolments and completions data. All of these reports can be found on the Ministry of Education's Education counts website at www.educationcounts.govt.nz

- Ministry of Education. Online Tertiary Education Statistics at http://www.educationcounts.govt.nz/statistics/tertiary_education

- Ministry of Education (2007) *Profile & Trends 2006: New Zealand's Tertiary Education Sector*. Ministry of Education, Wellington.

- Earle, D. (2007). *Te whai i ngā taumata atakura, supporting Māori achievement in bachelors degrees*. Ministry of Education, Wellington.

- Earle, D. (2007). *The System in Change, Tertiary Education Strategy, 2002/07 Monitoring Report 2005.* Ministry of Education, Wellington.

- Scott, D. (2006). *Passing courses*. Ministry of Education, Wellington.

- Scott, D., Smart, W. (2005). *What factors make a difference to getting a degree in New Zealand?* Ministry of Education, Wellington.

- Scott, D. (2005). *How long do people spend in tertiary education?* Ministry of Education, Wellington.

- Scott, D. (2004). *Pathways in tertiary education* Ministry of Education, Wellington.

- Scott, D. (2004). *Retention, completion and progression in tertiary education 2003.* Ministry of Education, Wellington.

- Scott, D. (2004). *Retention, Completion, and Progression in Tertiary Education 2003, Technical Documentation.* Ministry of Education, Wellington.

- Scott, D. (2004). *Assessment of TSPAR Matching (SNs and NSNs).* Ministry of Education, Wellington.

- Smart, W. (2007). *Persistence in doctoral research. Analysing the impact of the PBRF on the retention of doctoral students.* Ministry of Education, Wellington.

- Ussher, S. (2006). *What makes a student travel for tertiary study?* Ministry of Education, Wellington.

- Ussher, S. (2006). *From school, work or unemployment: A comparison of pathways in tertiary education* Ministry of Education, Wellington.

MINISTRY OF EDUCATION
Te Tāhuhu o te Mātauranga

newzealand.govt.nz