Using data integration

to validate official statistics

Felibel Zabala, Statistics NZ

# Introduction

Statistics New Zealand has been undertaking data integration since the beginning of the 1990’s (Statistics NZ, 2013). The earliest uses were mostly Statistics NZ driven but later data integration projects have been advanced from external interest. The push from the NZ government for the design of new statistics using integration of administrative data from different government agencies came in 1997 with the first trial being the integration of income data from Inland Revenue and beneficiary data from then the Ministry of Social Development. Since then data sources integrated expanded to include traditional survey data.

Statistics NZ’s Integrated Data Infrastructure (IDI) brings together linked datasets from a range of government agencies (including our own). The IDI is a large research database containing microdata about people and households and is continually growing. The IDI has paved the way to answer complex research questions to improve outcomes for New Zealanders.

The linking in the IDI is done by record linkage. The aim in record linkage is to link together two records from different datasets that belong to the same person or entity. Where the data cannot be directly linked on common identifiers, the linking is based on probabilistic linking. In probabilistic linking, records from one dataset are compared with records from another dataset. The aim of the linking in the IDI is to minimise creating erroneous links, that is, to maximise the precision rate. The program used in the linkage is IBM QualityStage. This program provides the software needed to perform probabilistic record linkage, according to the Fellegi and Sunter methodology. Details of the linking methodology in the IDI are available in [Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project](http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-resources/linking-methodology-statsnz-idi.aspx).

This paper describes different applications of data integration at Statistics NZ including its use for validating statistics. It will document issues identified with the use of administrative data to create and validate official statistics. It will also present lessons learnt from undertaking such activities.

# Design of new statistics

## Linked Employer-Employee Data (LEED) project

The first statistics designed from data integration were person-level statistics obtained from the LEED project Statistics NZ, 2012a). LEED is one of Statistics NZ’s first linked datasets. Information in LEED provided the backbone of the IDI prototype. LEED is created by linking a longitudinal employer and non-employer series from Statistics NZ’s Business Frame to a longitudinal series of payroll tax data from Inland Revenue (IRD). The Business Frame is a regularly maintained list of all businesses and organisations with a turnover greater than NZ$30,000 that are engaged in the production of goods and services in New Zealand. Since employees and employers have unique IRD numbers, the Inland Revenue client number is used as the key linking variable. However, probabilistic matching is used where exact matching cannot be established.

LEED is used to produce quarterly statistics that measure labour market dynamics at various levels – including industry, region, territorial authority, business size, sector, sex, and age. These statistics provide an insight into the operation of New Zealand's labour market. Some statistics it provides are filled jobs, worker flows, and total earnings. These statistics, together with other official statistics from Statistics NZ on aggregated movements on employment, help explain the causes of aggregate movements in jobs and worker flows and are therefore useful for explaining changes in the labour market. In LEED, a job is defined as a unique employer-employee pair present on an employer monthly schedule (EMS) on the 15th of the middle month of the reference quarter. Counts of filled jobs are restricted to people aged 15 years and over. The LEED production system was created in 2005 and since then official statistics have been produced quarterly and annually.

### Issues identified in integrating data sources in LEED

* The need of standardised processes which are responsive to administrative changes in the data supplied and to new administrative data available to Statistics NZ. First and foremost, standard software for automated data linkage robust to these changes should be identified. Various evaluations of record linkage software carried out within Statistics NZ led to the choice of QualityStage.
* Timing, unless receipt of data are common and regular. This will always be an issue for linked datasets and for statistics produced from these datasets. An example is where personal income needs to be provided annually in the June quarter. Personal income is an output from one of the HLFS supplements, the NZIS. If new entrants to the workforce are not picked up quickly in the IRD tax data, potential bias may result in personal income reported. Initial research work indicates this is an issue but needs more investigative work.
* A common editing issue with linked datasets is inconsistencies in the records linked. Where inconsistencies occur in records linked from two different data sources, it is important to know which of the two data sources is more reliable. Sometimes, even the order in which the datasets are linked is important in determining where an inconsistency arose. It is expected that as the number of datasets being linked together increases, the potential for efficiencies in detecting and treating inconsistencies in records increase as the number of variables increase. However, this may also increase the amount of editing required for the linked datasets. Issues to be addressed by an editing strategy for linked datasets can be summarised by its ability to: edit inconsistencies from the same unit from different sources, treat erroneous and missing variables in a record and ensure consistency in variables across a record for a time period and over time.

## Student loans project

The Student Loans project created an integrated dataset on student loans and allowance. The dataset provided new information about the recipients of student loans and allowances (Statistics NZ, 2012b). The integrated dataset assists with understanding the cost of the Student Loan Scheme to the Crown, forecasting, reporting of the asset in the Crown Accounts, costing policy changes, and assessing the socio-economic impact of, and the return on, education and more recently, outcomes.

The Student Loan Scheme was introduced in New Zealand in 1992 to provide financial assistance to people enrolled in tertiary study. Statistics NZ created a new statistical information source on student loan borrowers and allowance recipients by integrating, at the unit-record level, administrative data from the following government agencies:

* Ministry of Education data on study (e.g., field of study, programme of study, type of tertiary provider)
* Ministry of Social Development data on loan and allowances (e.g., course costs, living costs, interest accrued, student allowance and accommodation benefit, repayments)
* Inland Revenue data on loans and income (such as loan transfer details, debt outstanding, and income).

The student loans and allowance integrated dataset was created in two stages. The first stage used exact matching to merge source datasets using Inland Revenue client numbers. Education records were then added by probabilistic matching, using the student identification number, the tertiary institution number, and demographic variables. The integrated dataset is updated annually as each year's administrative data becomes available.

The first output of the integrated dataset was published in December 2002. Recent releases have focused on:

* male and female borrowing, debt, repayments, and income
* borrowing, debt, and income by age group and ethnicity
* loans and allowances
* borrowing, debt, and income of borrowers who completed their qualification compared with those who did not.

The Student Loans integration project faced similar issues with the LEED project. These two projects spearheaded the consideration of the quality of administrative data in the production of official statistics. A metadata information template was developed to initially assess the information provided by the administrative data sources. The template was designed to be independent of the intended use of the data source. The information it includes can be used to make an initial decision about whether the source has the potential to meet a particular statistical need. This type of template can also provide a list of potential quality problems which could be investigated in more detail with a particular statistical use in mind. The fields included in the template are shown in Table 1.

|  |  |
| --- | --- |
| Data Source | Name of agency |
|  | Application of data |
|  | High level summary of time period, variables, file structure |
| Population | Target and actual population |
|  | Reporting units |
|  | Coverage |
| Variables | Name |
|  | Definition |
|  | Values |
| Glossary | Definitions of terms used |
| File Structure | Conceptual |
|  | IT |
| Data collection and data entry | Application |
|  | Collection method |
|  | Frequency and timing of collection |
|  | Consistent approach (over different centres) |
|  | Question adequacy and respondent understanding |
|  | Contextual or methodological biases |
|  | Data capture, coding and editing |
|  | Updating procedures |
| Data Accuracy | Missing data |
|  | Imputation |
|  | Duplication |
|  | Rounding |
|  | Internal consistency |
| Changes over time in: | Concepts |
|  | Coverage |
|  | Data collection and data entry |
|  | Data accuracy |
| Accessibility | Privacy/security/confidentiality |
|  | Documentation available |
|  | Ease of access |
|  | Forms of dissemination |
|  | Timeliness of publications |
|  | Storage of historical information |
| Comparison with other data sources | Consistency at aggregate level |
|  | Comparison of variable concepts and values with SNZ standards |
| Summary of Statistical Integrity | Is the population well-defined? |
|  | Do reporting units map to statistical units? |
|  | Are variables well-defined? |
|  | Do data collection and entry systems have good quality controls? |
|  | Is data accuracy reasonable (bias, reliability, internal consistency)? |
|  | Is there a consistent time series? |

**Table 1.** Components of the Statistics NZ metadata template for administrative data

This template has been reviewed and an updated version is available in [Guide to reporting the quality of administrative data](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx).

# Validating income survey data

Various different projects are using data integration to examine the use of administrative data in social and population statistics. Inland Revenue data, specifically from the Employer Monthly Schedule (EMS) returns, used to create LEED, has been investigated as a possible replacement for income data in various projects. These projects have generally used comparisons between EMS and survey data, either at aggregate level or probabilistically linked. The measure of quality used has typically been how closely the administrative data is able to reproduce the results of surveys. Measures of link rates, coverage, and timeliness have also been used.

Statistics NZ uses a range of surveys to measure the income of New Zealand households and individuals. The Quarterly Employment Survey (QES) provides income measures for those in paid employment at the job level. The following surveys provide income at both the individual and the household levels:

* New Zealand Income Survey (NZIS)
* Household Economic Survey (HES) / HES (Income)
* Survey of Family, Income and Employment (SoFIE)
* Census of Population and Dwellings (Census).

All these surveys include income from various sources - wages and salaries, self-employment, government transfers (e.g., benefits and tax credits), private superannuation, and investments. Income data are collected from households; however, person-level and household-level income statistics are available and published.

Validation projects carried out to replace income aimed to determine how closely the income information obtained entirely from tax information reproduces survey results. Replacing income may fall into three categories:

* Full replacement of survey questions with tax data
* Production of income statistics using tax data
* Using tax data for benchmarking, imputation, validation, or other methods to improve the statistics obtained from income surveys.

No validation project has concluded that administrative data available in the IDI can fully replace the survey questions available in either NZIS or HES. Most projects support the use of administrative data in either the production of some income statistics or for benchmarking, imputation, validation, or other methods to improve the statistics obtained from income surveys.

### Issues identified in the validation projects

* Timeliness of administrative data. These include:
* The promptness in picking up of new entrants to the work force. A high number of new entrants not picked up quickly enough by administrative data leads to potential bias in official statistics.
* Timing issues around getting the linking accomplished in time for production
* Some key statistics are not available from existing sources of data. These include hours for each job and some income sources such as income from investments.
* Sources of potential bias have been identified. These include:
* Coverage and conceptual issues especially for some age groups such as 15-19 and 65+. A solution is to correct LEED totals to account for conceptual differences in the definition of income from wages and salaries as well as coverage differences.
* People with some income source seem more likely to match. This has a potential for bias especially if LEED data is used to replace NZIS data with non-matches being imputed.
* Using LEED will result in a break in the data series that needs to be managed
* Backwards and forward casting of the link especially for longitudinal data. A person may link in one quarter but not in another due to data quality reasons (or may link to a different record).
* Pay periods are not consistent with the survey reference period. NZIS collects weekly income while HES collects annual income. EMS reports monthly earnings. Scaling administrative will be an issue especially at unit record level.

### Recommendations from validation projects on income

* A quality framework is beneficial in assessing how closely the administrative data is able to reproduce the results of surveys.
* The key quality dimension to use is accuracy. The normalised root mean square error is a useful tool to evaluate the accuracy of statistics sourced from integrated datasets.
* The linked SoFIE-IR data should be used to evaluate the impact of non-response and the improvement of surveys through calibration.
* Tax data can be used to produce income statistics sourced from salaries and wages while various administrative data available from the IDI can be used to produce income statistics sourced from government transfers. Some calibration will still be required to correct for potential biases arising from measurement and representation errors.
* Administrative data is useful for benchmarking, imputation, validation, providing revisions to statistics or other methods to improve the statistics obtained from income surveys.
* Collect IRD number from survey respondents to improve linking. The IRD number can be used for exact matching rather than probabilistic matching.
* Investigate the use of other variables besides names and date of birth to improve on probabilistic matching especially for respondents who do not want to provide their IRD numbers.

# Improving the design of existing statistics

The Household Labour Force Survey (HLFS) was linked to the 2013 Census data to analyse non-respondents to the HLFS. The non-respondents were either refusals or non-contacts. The proportion of non-responders has been increasing over time, with response rates now around 86%. The purpose of the data integration was to gain a greater understanding of the demographic and labour force characteristics, using the linked 2013 Census data, of those who did not respond to the March 2013 quarter of the HLFS. Furthermore, the study wanted to investigate if non-response adjustment was required, over and above calibration methods already implemented in the survey, to further reduce the potential non-response bias.

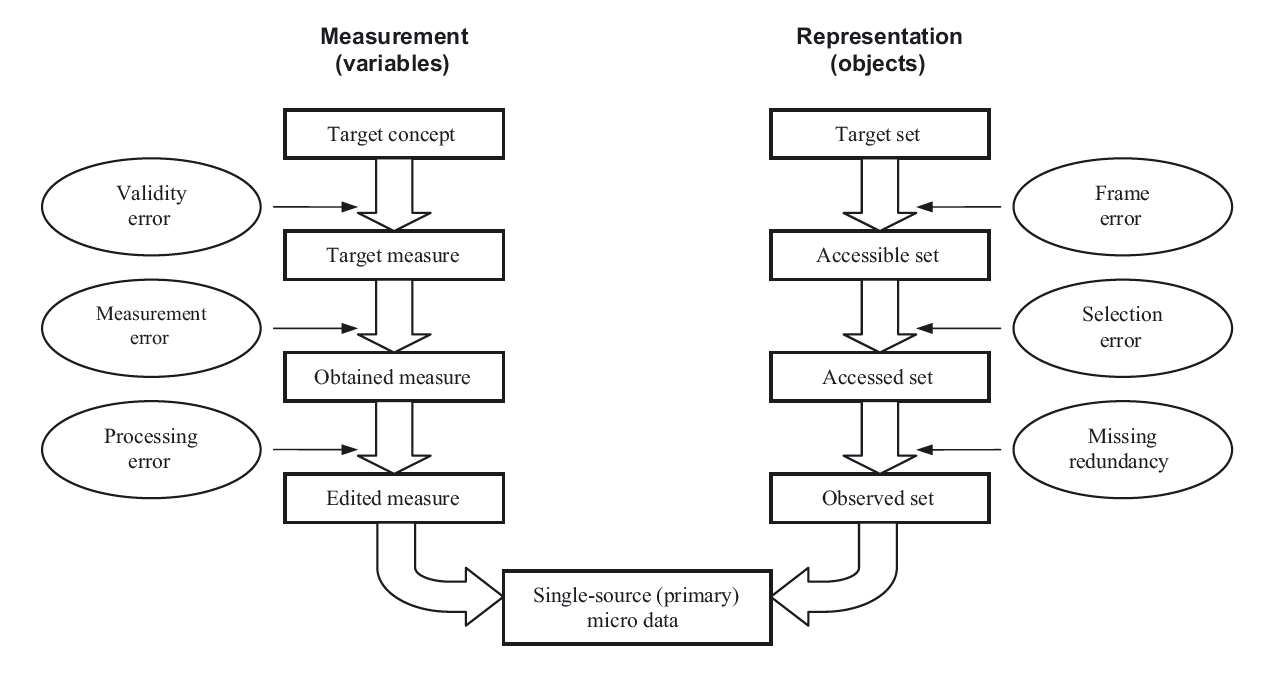
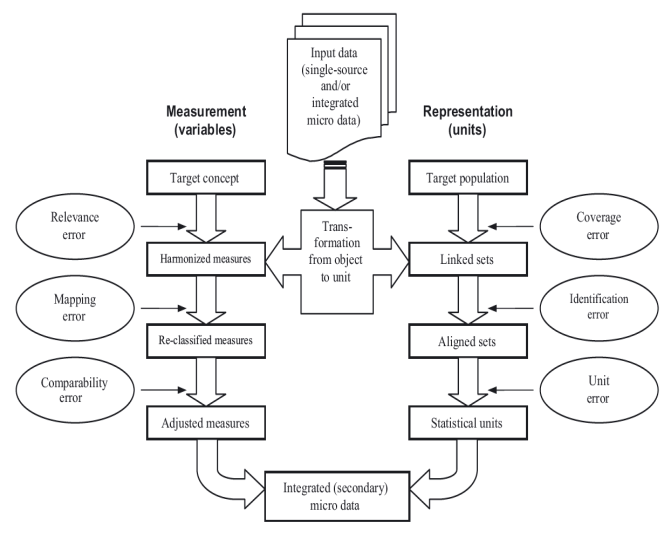
The linked dataset showed that bias from non-response exists in the estimates of labour force status from the HLFS. The unemployment rate in the responding sample is underestimated by around 0.3% to 0.5%, while the labour force participation rate is underestimated by around 0.2%. Unemployed people had both the highest non-contact rate and the highest refusal rate. These results, however, showed minimal impact on labour force estimates and relative sampling errors. While a non-response adjustment appears to correct for some of the bias introduced by non-response, the effect of this is within current sampling errors. Thus, Statistics NZ has therefore abandoned a non-response adjustment step in the weighting procedure for the HLFS, and relies on calibration to adjust for some of the non-response bias that is present.

The study revealed some interesting results with regard to data integration. Non-responding individuals have a lower match rate (98.0%, or 96.4% for households) and a higher estimated percentage of spurious matches (27.3%, or 33.5% for households) than responding individuals. These may be caused by the quality of data available from non-responding individuals. Further investigations need to be carried out for studies undertaking non-response analysis using integrated data. For the HLFS study, a sensitivity analysis showed that this was not a problem for the non-response analysis carried out.

# Determine suitability of administrative data for census-type attribute information

Data integration has also been used in Statistics New Zealand’s [Census Transformation Programme](http://www.stats.govt.nz/Census/census-transformation-nz.aspx). The programme is investigating alternative ways of running New Zealand’s future census. The programme aims to modernise the current census model as well as investigate different ways of producing small-area population, social, and economic statistics. Investigations are also looking at the feasibility of using linked administrative data to replace census questions. Quality in terms of the ability of administrative data to satisfy the information needs currently met through people filling out a census questionnaire is being assessed.

Statistics NZ’s quality frameworks were used as basis of quality assessments carried out. At the highest level, Statistics NZ’s quality framework is based on six dimensions of quality widely used internationally: Relevance, Accuracy, Timeliness, Accessibility, Coherence and Consistency, Interpretability. A more detailed quality assessment process is described in [Guide to reporting the quality of administrative data](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx). The quality framework is based on Li-Chun Zhang’s [two-phase life-cycle model for integrated statistical microdata](http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2011.00508.x/abstract) (Figure 1) which expands the total survey error paradigm to include administrative data. The framework enables understanding of the error sources from the individual data sources including those arising from the integrated datasets. In the model for an administrative census, attributes will be derived from multiple administrative sources linked together at the individual level. Zhang’s two-phase life-cycle model assists in determining the associated methodological and operational issues that may impact on quality resulting from producing statistical information from linked administrative data sources.

Phase 1 of Zhang’s framework Phase 2 of Zhang’s framework

Figure 1. Zhang’s two-phase life cycle model

Although results of the investigations were mostly descriptive, Zhang’s model was useful in assessing the differences between data from the 2013 Census with information from the administrative data sources available in Statistics NZ’s IDI. Conceptual differences in the census variables and coverage differences at aggregate levels were identified and likely reasons for the differences were determined. The model was also instrumental in determining areas to focus on when 2013 Census data linked to the IDI were compared to determine consistency between census responses provided by an individual with the values derived from the administrative data for the same person. Cross-comparisons (often in the form of two-way tables) helped understand measurement error in the census variables. Since the individual-level results were based on the linked dataset, they would not have been affected by differences in population. Although results may have been affected by linkage error between the Census and the IDI, its impact may not be significant since 95% of returned 2013 Census forms were linked to the IDI with an estimated 1% of incorrect links.

One issue identified with using multiple linked administrative data was the substantive inconsistency in values of some variables across these datasets, e.g., ethnicity and address, when compared to census data.

More details are available from the paper [Using quality frameworks to assess the potential use of administrative   
sources in the census](https://consol.eventsair.com/QuickEventWebsitePortal/itsew2016/itsew16/Speaker).

# Summary

Statistics NZ has used data integration for various purposes, including its use for validating statistics, since the beginning of the 1990’s. Issues identified with use of administrative data to validate official statistics include the following:

* The need of standardised processes which are responsive to administrative changes in the data supplied and to new administrative data available to Statistics NZ.
* Timeliness of administrative data, unless receipt of data are common and regular. This will always be an issue for linked datasets and for statistics produced from these datasets. These include:
* The promptness in picking up birthed units to an administrative data source. A high number of birthed units not picked up quickly enough by administrative data leads to potential bias in official statistics.
* Timing issues around getting the linking accomplished in time for production
* Coverage errors and measurement errors typically occur for quite different reasons, and suggest quite different approaches to improving quality. Only some of these are within the control of the administrative agencies. Several main causes of coverage error have emerged, and with them different possible solutions.
* A common editing issue with linked datasets is inconsistencies in the records linked. Where inconsistencies occur in records linked from two different data sources, it is important to know which of the two data sources is more reliable. Sometimes, even the order in which the datasets are linked is important in determining where an inconsistency arose. It is expected that as the number of datasets being linked together increases, the potential for efficiencies in detecting and treating inconsistencies in records increase as the number of variables increase. However, this may also increase the amount of editing required for the linked datasets. Issues to be addressed by an editing strategy for linked datasets can be summarised by its ability to: edit inconsistencies from the same unit from different sources, treat erroneous and missing variables in a record and ensure consistency in variables across a record for a time period and over time.
* Some key statistics are not available from existing sources of data.
* Sources of potential bias have been identified with regard to integrating datasets. These include:
* Coverage and conceptual issues may only apply for some groups of a population so care should be taken in generalising results.
* Some variables have the potential to affect the quality of linking and may be a source for potential bias in carrying out analysis on resulting datasets. Investigations on linkage rates across different subpopulations may be required.
* The use of linked datasets even for validation purposes may result in a break in the data series that needs to be managed
* Backwards and forward casting of the link especially for longitudinal data. A person may link in one quarter but not in another due to data quality reasons (or may link to a different record).
* Need for a weight to adjust for missed links
* Methods to better estimate linkage errors are required to determine models appropriate to account for these linkage errors. Linkage errors contribute to potential coverage errors in the resulting target population. Care should also be undertaken when creating statistical units from integrated datasets wherein one dataset is sourced from an administrative dataset since the unit may be defined differently in the administrative dataset.
* Administrative data may suffer from measurement errors, e.g., validity error, and these errors propagate when the administrative data is integrated with other data sources to produce a statistical output. Hence, target concepts used in an administrative dataset should be well understood before being used in the production of official statistics.

# Recommendations

* The quality assessment framework described in [Guide to reporting the quality of administrative data](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx) is beneficial in carrying out validation studies. The quality assessment involves 3 steps.

**Step 1: Initial metadata collation:** Basic information is collected about each of the source datasets used in the validation project. The information relates to the source agency, purpose of the data collection, populations, variables and timeliness of the data.

**Step 2: Phase 1 evaluation:** Errors occurring in phase 1 of the quality framework are determined and categorised for each source dataset. This involves detailed consideration of how the methods, purpose, known issues, and other aspects of the original data collection contribute to each of the specific error categories in the phase 1 flow chart in figure 1.

**Step 3: Phase 2 evaluation:** As for the previous step, errors arising in phase 2 of the quality framework are listed and examined in a similar way, taking into account the dataset(s) being integrated to produce the final output. These errors are considered with respect to the intended statistical target concepts and population. The effects of phase 1 errors on the creation of statistical units, or the particular details of the misalignment between concepts on different datasets, must be understood.

* Understanding the error sources from the input data sources, including those arising from the linked datasets, is essential in determining the associated methodological and operational issues that may limit the use of administrative data in the validation of official statistics.
* The key quality dimension to use is accuracy. The normalised root mean square error is a useful tool to evaluate the accuracy of statistics sourced from integrated datasets.

# References:

Bycroft, C. 2016. [Using quality frameworks to assess the potential use of administrative sources in the census](https://consol.eventsair.com/QuickEventWebsitePortal/itsew2016/itsew16/Speaker). Paper presented at the 2016 International Total Survey Error Workshop, Sydney, Australia.

Hall, J. and Anand-Kumar, V. 2014. [Analysis of non-respondents to the Household Labour Force Survey using linked 2013 Census data, Statistics New Zealand Working Paper No 16–02](http://www.stats.govt.nz/methods/research-papers/working-papers-original/analysis-non-respondents-hlfs-16-02.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2012a. [Linked Employer-Employee Data (LEED) project](http://www.stats.govt.nz/browse_for_stats/income-and-work/employment_and_unemployment/guide-interpreting-the-leed-data/definitions-of-measures.aspx). . Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2012b. [Student Loans project](http://www.stats.govt.nz/methods/data-integration/data-integration-projects/student-loans-project.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2013. [Data integration manual: 2nd edition](http://www.stats.govt.nz/methods/data-integration/data-integration-manual-2edn.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2014. [Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project](http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-resources/linking-methodology-statsnz-idi.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand. 2016. [Guide to reporting the quality of administrative data](http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Zabala, F. 2016[. Using administrative data to validate income in Statistics New Zealand's household surveys](https://consol.eventsair.com/QuickEventWebsitePortal/itsew2016/itsew16/Speaker). Paper presented at the 2016 International Total Survey Error Workshop, Sydney, Australia.

### Unpublished documents from Statistics New Zealand:

Anastasiadis, S. 2013. *EMS administrative data – thoughts on the barriers*.

Anastasiadis, S. 2014. *Hourly earnings from the HLFS-EMS link September 2014*.

Anastasiadis, S. 2014. *Pay period variability – can we remove by aggregating data?*

Archibald, J. 2014. *Measuring the accuracy of mixed source statistics*.

Bycroft, C. 2009. *HLFS-LEED interaction*.

Enright, J. 2011. *Using LEED administrative data to improve income survey: A feasibility study*.

Hansen, C. 2011. *The role of administrative data in New Zealand’s Household Labour Force Survey*.

Moes, W. 2010. *Comparing estimates from NZIS to LEED values.*

Plugh, R. 2014. Administrative data to replace government transfers questions in NZIS.

Reid, G. 2012. *Report for the administrative data roadmap project Deliverable 8: The replacement of respondent data with tax data in surveys that measure income*.

Young, N. 2012. *HLFS-LEED link on IDI round 1 link investigation*.