DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 477–511, http://dx.doi.org/10.1515/JOS-2017-0023

Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ

Giles Reid[1], *Felipa Zabala*[2], *and Anders Holmberg*[3]

Many national statistics offices acknowledge that making better use of existing administrative data can reduce the cost of meeting ongoing statistical needs. Stats NZ has developed a framework to help facilitate this reuse. The framework is an adapted Total Survey Error (TSE) paradigm for understanding how the strengths and limitations of different data sets flow through a statistical design to affect final output quality. Our framework includes three phases: 1) a single source assessment, 2) an integrated data set assessment, and 3) an estimation and output assessment. We developed a process and guidelines for applying this conceptual framework to practical decisions about statistical design, and used these in recent redevelopment projects. We discuss how we used the framework with data sources that have a non-statistical primary purpose, and how it has helped us spread total survey error ideas to non-methodologists.

Key words: Total survey error; multiple data sources; official statistics; survey design.

1. Introduction

Producers of official statistics are facing increasing pressures to save money while maintaining or even increasing the quality and timeliness of outputs. In many countries, response rates for traditional surveys have been falling, but more and more administrative and other non-traditional data have become available. There is an urgent need to find ways to use administrative sources to increase the efficiency and effectiveness of statistical production. Administrative data cannot solve all of our problems, and traditional survey data collection is still needed in many cases. At Statistics New Zealand (Stats NZ), we have been faced with the challenge of redesigning statistical outputs to make better use of administrative data while maintaining the data quality required by our users.

In this article, we propose a quality framework that we have developed to provide a systematic approach to meeting the Stats NZ goal of using administrative data as the first source of data, supplemented by survey data collection only when necessary. This framework is widely applicable to all kinds of input data and statistical outputs and includes considerations of estimation models and continuous improvement alongside its total survey error foundations.

[1] Statistics New Zealand, 2018 Census, PO Box 2922, Wellington 6140, New Zealand. Email: giles.reid@stats.govt.nz.
[2] Statistics New Zealand, Statistical Methods, PO Box 2922, Wellington 6140, New Zealand. Email: felipa.zabala@stats.govt.nz.
[3] Statistics Norway, Division for Methodology, Akersveien 26 Oslo, Norway. Email: Anders.holmberg@ssb.no.

In 2016, Stats NZ released its vision to "unleash the power of data to change lives, which will enable data-led innovation across society, the economy, and the environment". Its aim is "to increase the value of data to decision-makers tenfold in the next 15 years" (Statistics NZ 2016b, 1). Survey methodology provides good ways to answer questions about how to measure and improve data quality, but it requires us to have a high degree of control over the entire process from collection to output. Designing statistical outputs that use administrative data creates many new challenges because we have to give up direct control over many processes, including population definitions, collection methods, classifications, and data editing. Each administrative source has its own particular problems that must be understood both for our own design work and to assure the final users of the data that our outputs are fit for purpose. When we use administrative data instead of a traditional survey, we need new processes, such as data integration and re-coding or adjusting administrative variables, which can introduce new types of errors.

The quality assessment framework presented in this article provides a basis for understanding how these factors fit together. We expand and unify earlier conceptual work from various writers to make it more directly and easily applicable to practical statistical design in official statistics. Based on our experiences from survey redevelopment projects within Stats NZ, we also provide a sequence of practical steps which can be followed during the design process.

Our three-phase framework applies the Total Survey Error (TSE) paradigm (see, for example Groves and Lyberg 2010; Biemer 2010) to the new realm of statistical production, which involves integrating and combining data from various sources. It builds on Li-Chun Zhang's extension of this TSE thinking to administrative and integrated data (Zhang 2012). It makes use of various quality indicators and measures, such as those developed as part of the European Statistical System Network (ESSnet) and BLUE-Enterprise and Trade Statistics (BLUE-ETS) projects, alongside earlier Stats NZ quality work, like metadata templates, output quality reviews, and process reviews. (Burger et al. 2013; Daas et al. 2011; Daas et al. 2012; Statistics NZ 2016a). The framework assists in understanding how well different data sets meet their originally intended purpose (Phase 1) and what their strengths and limitations are. It provides a way of determining what effects these strengths and limitations may have on the quality of a statistical output that makes use of these "found" data sources, statistically designed (possibly sample survey based) data, or a combination of the two (Phase 2). In that sense, our framework suggests an extension of Zhang's work including a third phase with evaluations between design options for a statistical output.

Quality assessments carried out with this framework can help answer statistical design questions on how to use available data to meet user needs in an efficient way. They help to decouple the true statistical needs of our users from design decisions: our goal should be to meet these needs as best we can with the data available. Sometimes reproducing the results of a sample survey using administrative sources may not be feasible, but a new alternative output can be produced which still meets existing needs, or meets emerging needs that the old survey outputs did not.

The framework is also part of Stats NZ's continued efforts to be better equipped for a changing data environment with an increasing array of unconventional data sources. Our new strategy is to increase use and reuse of data already collected, both in the production

of traditional official statistics and for new research projects. More reuse of data also means we need to always consider that all new data we collect may have new uses in the future. To make this reuse possible, documentation is essential; the framework gives a clear guide to what should be recorded and how the documentation should be structured. The framework also helps with managing data from multiple data sources simultaneously, and improving the opportunity to use them in an integrated way.

In this article, we adopt the United Nations Economic Commission for Europe (UNECE) definition of administrative data: "data that is collected by sources external to statistical offices" (UNECE 2011b, 2). and "administrative sources are data holdings containing information which is not primarily collected for statistical purposes," (UNECE 2011b, 4). We use the term "survey" in a classical sense, which does not necessarily mean the data acquired was selected with probabilistic methods.

We used the Organization for Economic Cooperation and Development (OECD) definition of a statistical product to define *statistical outputs*: "an information dissemination product that is published or otherwise made available for public use that describes, estimates, forecasts, or analyses the characteristics of groups, customarily without identifying the persons, organisations, or individual data observations that comprise such groups. This may include general-purpose tabulations, analyses, projections, forecasts, or other statistical reports" (OECD 2007, 745) as well as data sets containing unit record data.

In Section 2 we present the quality assessment framework and discuss how it is used. We also provide a list of quality measures and indicators that can be used to measure different types of error. The versatility of the framework is illustrated by three applications: the redesign of our Quarterly Building Activity Survey in Section 3, the use of tax data to measure personal income in Section 4, and the use of linked administrative data to estimate resident population counts in Section 5. We conclude with a summary and discussion.

2. The Quality Assessment Framework

2.1. *Developing the Framework*

Stats NZ's "administrative data first" goal means that during the design phase of a proposed statistical output, we first have to confirm if an existing data source can be used to provide all or part of the required information and satisfy data needs. To ensure an administrative data approach is comparable to a classical survey design (with a tailored questionnaire, sample selection scheme, controlled data collection, data processing etc.), measures are required to assess the quality of alternative data sources and determine how they fit together to answer statistical needs. Assessments that determine whether the data sources are fit for purpose enable sound decisions on whether using them is a cost-effective alternative to directly collecting new data ourselves.

We first investigate useful quality measures for statistical outputs that use administrative data. The most important motivation is the need to understand in detail, the risks and benefits involved when we are redesigning statistical outputs to make more use of administrative data. We must be able to assure users that our new designs will produce fit-for-purpose data that will meet their needs. Without a thorough understanding

of the sources of error affecting output quality, it is very difficult to evaluate whether the savings and efficiencies from the use of administrative data will be worth the potential loss in output quality.

There are many approaches to the quality assessment of administrative data (see Daas et al. 2010, 2012; Daas et al. 2011; Wallgren and Wallgren 2014; UNECE 2011b). However, our work focuses on how the quality of statistical outputs that use administrative data can be assessed. To do this and to enable an administrative-data first production environment with easy reuse of data, we developed quality measures both for the administrative data we use as input to our statistical outputs ("input quality") and for the statistical outputs produced from administrative data ("output quality").

To assess the input quality of the administrative data entering a national statistics office, our framework includes qualitative as well as quantitative indicators based on the quality concepts given by Daas et al. 2010. These indicators have also been included in Stats NZ's meta-information template for evaluating new data sets (an online version of the template can be found in Statistics NZ 2016a). As for indicators of the output quality, our framework is influenced by the work by Burger et al. (2013). They investigated the use of administrative data to avoid unnecessary reporting burden on businesses and provided quality indicators for statistical outputs that use mixed sources of data. Other agencies have adapted or developed frameworks for measuring quality. Examples include Australian Bureau of Statistics Data Quality Framework (Australian Bureau of Statistics 2009) and Statistics Canada's Quality Guidelines (Statistics Canada 2009). These are of limited practical use in determining what the quality of our outputs will actually be since quality indicators are not explicitly defined. The United Kingdom's Office for National Statistics' Guidelines for Measuring Statistical Quality (Office for National Statistics 2013) provides quality indicators useful in the assessment of output quality.

Li-Chun Zhang's two-phase life-cycle model for integrated statistical microdata (Zhang 2012) helpfully expands the TSE paradigm in a way that makes it applicable to mixed-source statistical outputs. We adopted this model for the first two phases of our framework because its systematic list of the ways in which error arises in statistical outputs, is applicable to designs using traditional survey methods, administrative data, and mixtures of the two. This enables us to compare the various sources of error affecting rival statistical designs aiming to produce the same statistical output with different mixtures of input data. The two phases cover the processes used to create a final unit record data file. In Phase 3 of our framework, the errors that arise from the estimation process are considered, alongside the evaluation and correction of errors. Our framework also gives a useful vocabulary for this error and statistical design comparison, which can be explained to non-methodologists with limited familiarity of administrative data, TSE, or both. It also provides a structure to organise the practical knowledge from processing which analysts have about the sorts of errors that affect their statistical output.

One major attraction of the framework is that it explicitly distinguishes "input quality" and "output quality". Input quality, or how well a single data source meets its original purpose, is particularly relevant to Stats NZ's aim of reusing data and matches well with our existing meta-information template for evaluating new data sets (Statistics NZ 2016a). The sources of error under Phase 1 of Zhang's model are a result of the initial data

collection and processing, and will flow through into any use of the data in the production of a statistical output. Phase 2 errors relate to using these source data sets to produce a *particular* statistical output. They depend on the desired outputs and the design under consideration. When a new statistical output is being designed, previous Phase 1 evaluations of the data sources under consideration can be reused. Some additional Phase 1 evaluation work may still be required but the previous Phase 1 evaluation still saves a lot of information gathering and initial investigations.

To practically apply the concepts in the framework with an overall aim of identifying and understanding all sources of error that affect a statistical output, an organised list of the sources of error, and at least a rough idea of their relative magnitude, is essential. Rigorous measurement is often difficult, but is necessary for design, process monitoring, and reporting to users.

2.2. *Components of the Framework*

The three phases of the quality assessment framework are separated so we can understand the effects of data processing on the quality of the statistical output.

2.2.1. Phase 1

Figure 1 shows a flow chart illustrating Phase 1 of the quality assessment framework from Zhang (2012). The flow chart is similar to those in works such as Groves and Lyberg (2010, Figure 3). The main difference is that Zhang (2012) uses more generic terms that apply to both survey and administrative sources. The most important aspect of this diagram is the flow (shown by arrows) between the rectangular boxes from the initial target concept and target set to the final data stored. At each stage errors can arise (represented by the ovals). Throughout the Phase 1 assessment process, it is the target concept and target set (intended by the organisation that created the data) that we must assess against. Using someone else's data means we cannot control any of their decisions

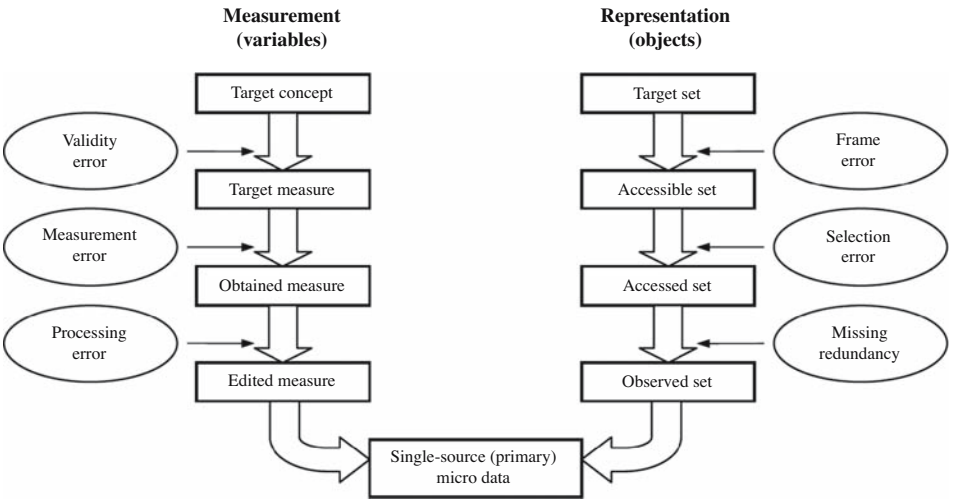

Fig. 1. Phase 1 of the quality assessment framework (Zhang 2012)

on measurements and populations. We need to understand their design decisions so we can determine what to do to turn their data into the information we want. Definitions of terms in Phase 1 of the quality assessment framework are given in Appendix A.

Although the framework applies to both administrative and traditional survey data, different types of errors tend to dominate. Our test cases (Sections 3–5) show that administrative collections, particularly for business data, usually have very good alignment between the target concept and the measure used to capture it. For instance, the value of sales taxes paid by a retail business in a given calendar month is objective and well defined, so validity errors are small compared with conceptually complex individual survey questions about ethnicity or well-being.

The distinction between frame error and selection error can be confusing, especially when the administrative data have been designed with restrictions already in mind. An example of these errors is in the recording of transaction events. Suppose that a retail chain wants to produce statistics on the transactions across all its stores, but the system they use can only record purchases that use electronic cards. Cash transactions could be said to be "inaccessible" since they will never be in the database. On the other hand, if a store manager forgets to run the reporting tool for a week, the transactions missing from the data set due to that mistake will be selection errors: they were accessible, but were not accessed.

Phase 1 of our framework provides some measures for each of the identified error components of a data source. Examples of quality measures for measurement error include the item imputation rate of a variable and the lag time between the reference period and the time of receipt of the data source. Quality measures for frame error include undercoverage and overcoverage. In instances where a metric assessment is not possible, the framework will assist in identifying processes where potential errors may arise so these can be addressed during the design of the output statistic. More complex measures are also possible: Bakker (2012) used a structural equation model to assess bias arising from measurement errors from various data sources, and Scholtus and Bakker (2013) used a simulation study to test the robustness of the model to additional components of measurement error as well as selection errors.

See Appendix A for a list of quality measures and indicators for Phase 1. Note that we focus on administrative data use and the new potential for errors it raises, so our examples are centred on administrative data. Many of the same or similar measures are also relevant to survey data, or can be made so with small modifications.

2.2.2. Phase 2

Phase 2 of the quality assessment framework is illustrated in Figure 2. Phase 2 focuses on errors that arise when data sets from several sources are integrated to produce an output that meets a certain statistical purpose. Phase 2 also includes errors from an output produced mainly from a single administrative data set.

In this phase the reference point is the statistical (target) population we would ideally have access to and the statistical (target) concepts of the units we want to measure in the target population. In practice, it takes some care to precisely define the true targets. In an established survey design, for instance, sometimes there is not a clear distinction between the sampling frame developed for practical purposes and the true target population. Some of the errors that arise during Phase 1 can also propagate through to the final output, and

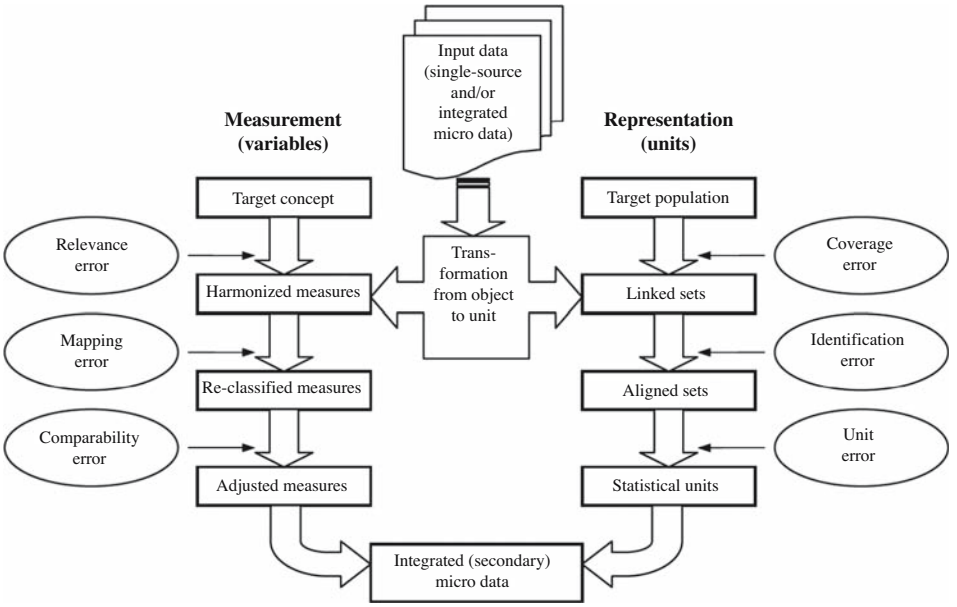

Fig. 2. Phase 2 of the quality assessment framework (Zhang 2012)

the flows in the figures are not necessarily directly related to specific or sequential steps in a statistical production process. See Appendix B for a definition of terms in Phase 2 of the quality assessment framework.

If we again look at sales tax data from our tax agency and consider Phase 2; the sales taxes paid by a business in a given month might not actually correspond to the true sales in that month, which is generally what the statistical output is more concerned with. Depending on the details of the tax system, the sales taxes may be paid in the month after the actual sale of the goods, or there may be sales taxes paid on items at the time they are brought into the store rather than at the time they are actually sold. These mismatches give rise to errors, specifically mapping errors, in the "measurement (variables)" column (see Figure 2).

For the "representation (units)", other difficulties may be encountered. If a particular branch of a retail store franchise changes ownership, New Zealand tax reporting rules often result in an entirely new tax unit being created in the administrative source. From the point of view of the tax agency, as long as the tax owing is paid correctly, this does not result in any error in the units. From the point of view of a business survey, however, such changes result in the old unit being dropped from the survey (because it is marked as ceased in the tax data), while the new unit may not be selected because the magnitude of its tax activity is too small to qualify it for selection. In reality, the store continued in the same way, but our rules for creating and selecting survey units using the administrative data have introduced errors for "representation (units)".

Data integration is also an important source of error in Phase 2. Stats NZ's Integrated Data Infrastructure (IDI), discussed in more detail in Sections 4 and 5, combines information from several government agencies, and so create a central list of individuals who interact with the government. We found many cases where an individual has multiple

records with the same agency. In some cases these duplicates are flagged and linked by the agency, but if they cannot be detected and removed, we will be creating too many individual records. This in turn leads to problems when linking other data sets, because our record-matching process effectively has to choose between two different duplicate records when integrating to another data set, and the result will be rather unpredictable.

Phase 2 of the framework provides measures for each of the identified error sources of an integrated data set. These are listed in detail in Appendix B.

For some data sets there may be no linking, just processing and conversion from a raw input data set into an output. In these cases, measures such as link rates may not be very useful, but concepts such as coverage of the target population, and conversion of administrative objects to statistical units, can still be valuable.

2.2.3. Phase 3

The end point of Phase 2 is a unit record file containing a set of units and a set of variable values for each of these units. Typically, this unit record file is not itself the final output desired: this file is used to derive estimates, such as the unemployment rate, or the population for a range of geographic regions. We included Phase 3 in the framework to account for the processes and errors that can arise in the creation of these final outputs.

In our framework, Phase 3 includes the work done to evaluate or estimate the quality of the final outputs, taking into account all error sources. It also concerns the inaccuracies introduced by estimation methods that attempt to correct for sources of error that arise in the first two phases.

In a traditional survey context, the estimation process can include a variety of techniques, from simple sums and averages to complex model-based methods that use auxiliary data to calibrate or correct for selection or nonresponse biases. Other processes, like seasonal adjustment, may also be carried out to further correct or adjust final estimates. Seasonal adjustment could be thought of as a correction for relevance errors: for example our desired output could be to measure the underlying growth rate of, say, an industry sector, but our raw results only measure the combination of seasonal and true growth. Using seasonal adjustment we can estimate the size of the seasonal movements and remove them, but this process itself is subject to error.

It is difficult to create a generic set of steps for this phase, but the aim is to consider the estimation methods and the corrections that can be applied to deal with various sources of error. It should also include an evaluation of the estimated level of error remaining in the final estimates. Traditionally, the key indicator published by statistical agencies is the sampling error, but as we saw in Phases 1 and 2, there are many more non-sampling errors that, ideally, we would try to estimate. Ultimately, if this error estimation can be done for competing designs that are candidates to estimate the same underlying quantity or concept, then we can use these error estimates as the foundation for a cost/quality trade-off.

In the planning (or design) stage we can use comparative production costs and our best "guesstimates" of the total error in the desired outputs to determine whether an overall statistical design is well-motivated, compared with some other configuration and use of input data sources from Phase 2 (or just Phase 1 if no integration is considered as in many traditional surveys). Ideally, these estimations and evaluations would include optimising a multivariate TSE measure subject to cost restrictions, but this is unrealistic because of

complexities and possible shortcomings in assessing errors in Phases 1 and 2, including the use of different indicators with different scales. Instead, we advocate a practical approach that analyses different options in Phase 3 by appropriately weighting and comparing individual error components, and on that basis reach a decision on design. Although this approach will include a significant amount of judgements, we argue that this is a more thorough, methodological, and systematic way of achieving better-quality outputs than making the statistical-design decision based on a single (first) choice of data set. The approach enforces the practice of setting and thinking of competing objectives and comparing design options. This increases the chances of getting a good outcome that considers the cumulative effect of errors. It is in a phase that compares outputs (estimates) where evaluations that affect final choices on statistical design can be made.

Laitila and Holmberg (2010) give an example of how a Phase 3 comparison can be made. They suggest estimating the total error of an estimator from one data source by deriving lower and upper boundaries for a Total Mean Square Error (TMSE) measure. Let $\tilde{Y}_1(r,m)$ and $\tilde{Y}_2(r,m)$ denote an estimator of a parameter Y under representative (r) and measurement errors (m) from two different Phase 2 data set alternatives. By decomposing TMSE of the estimators with respect to the error sources and comparing them, there is guidance about which one is best. The derivation of $TMSE(\tilde{Y}_i(r,m))$ can be done in different ways (for example Biemer and Lyberg 2003; Biemer 2010; Laitila and Holmberg 2010; Smith 2011). Each of these approaches involves different assumptions, so the best choice depends on the particular case under consideration, the error sources, and the indicators available from Phase 1 and Phase 2. The derivation is an important and non-trivial step. A full consideration of the choice of a total mean square error measure would be too complex to include in this article, but one particular challenge is how to address the cases when randomisation theory does not easily apply to some data sources.

A very important aspect of these comparisons is the recognition that errors can be accounted for and potentially corrected within our estimation process. If we know from an independent survey (for example an audit sample) that a certain administrative data set has systematic undercoverage of our target population, then including this as a correction factor will bring our estimates closer to the true value. Ideally we would repair or eliminate errors at source or during the production of the unit record file, but this may be impossible. For instance, we can quite easily measure the rate of erroneous links in integrated data sets using a clerical sample, but searching through the entire linked file removing all incorrect links is impractical. Instead, we can use the error rate as an input to an estimation model that aims to produce corrected final estimates.

One possibility for estimation in these scenarios has come from work done at Stats NZ to estimate the size of the New Zealand resident population. In the past, we relied on data from the Census of Population and Dwellings, collected in a full-coverage survey of the country, but Bryant and Graham (2015) describe a Bayesian approach for population estimation from administrative data under coverage errors. By expanding that estimation approach to also include other types of errors identified by the framework, and comparing the uncertainties arising from different combinations of data, we have a tool to assist in making a better design choice. The error decomposition and the knowledge from the indicators in Phase 1 and Phase 2 can be used as inputs to the model and contribute to the uncertainty of estimates. Substantial further work is required to develop this idea, more

generally. There are other alternatives, but we believe that these ideas are a promising solution to dealing with errors in administrative sources that cannot necessarily be identified and repaired at a unit record level.

2.3. *Applying the Framework in the Design of a Statistical Output*

Our quality assessment framework is useful for designing a statistical output that considers either the use of a single data source or an integration of several data sources. Because the error categories and concepts in the framework are often quite abstract, it can take some time and effort for analysts to come to grips with them when they are first introduced. When the framework was developed, we carried out practical tests on various outputs together with subject matter analysts who are very experienced in the practicalities of working with their data and processing systems. Based on these tests, we settled on a rough sequence of tasks for applying our framework.

The evaluation process we developed has four steps.

- **Initial metadata collation:** Basic information is collected about each of the source data sets that contributes to the final output. The information relates to the source agency, purpose of the data collection, populations, reporting units, variables, timeliness of the data, and so on.
- **Phase 1 evaluation:** Errors occurring in Phase 1 of the quality framework are determined and categorised for each source data set. 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.
- **Phase 2 evaluation:** As in the Phase 1 evaluation, errors arising in Phase 2 of the quality framework are listed and examined in a similar way, taking into account the data set(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 data sets, must be understood.
- **Phase 3 evaluation:** The previously identified sources of error are evaluated and further investigations are done into how they might be measured, controlled, or reduced. This may include developing and applying tailored quality measures and indicators. It also includes determining which sources of error should be minimised or which data source minimises a specific source of error so that the final statistical output is optimised. The error measurements may eventually feed into an estimation model that attempts to correct known data problems as much as possible.

Once this four step process is completed, the final outputs will include a list of the sources of error that affect both the input sources and the final statistical output, and corresponding measures to be used to assess the size or effect of each of these errors, where possible.

An important principle we agreed on during our tests was that the framework should be used in a flexible way. For some major design projects we might need to examine every detail of every type of error that might arise. In other cases, the goal might be to produce a basic report that explains the data under investigation in general terms and highlights its

main features and potential flaws. The effort spent on an evaluation should depend on the requirements, and the process should never be a routine box-ticking. A more detailed guide to the implementation of our quality assessment framework is available in Statistics NZ (2016a).

2.4. *Case Studies*

The following sections describe three projects that we used to test and develop our framework in practice.

The first, the Building Activity Survey redevelopment, was the first full redesign project carried out at Stats NZ where we tried to apply the process of mapping out the sources of error and systematically measuring or correcting for them. It is a relatively simple survey so was a good test case for administrative data replacement in business surveys, and balancing the cost savings made against any quality risks introduced.

The second case study relates to the measurement of personal income in household surveys, and the potential for using linked personal tax records to replace population census collection of this variable. We have included it because it is a good demonstration of the way our framework can capture, not only the issues with administrative data sets, but also categorise and understand the errors that arise in the traditional collection of this variable.

Our final case study is more of a work in progress, and examines the problem of population estimation from (imperfectly) linked administrative sources. It is important because it shows the value of the Phase 3 thinking that we have introduced in our framework and how new estimation models can take advantage of our systematic approach to the evaluation of error. It is also a good example of how we separate out the causes and effects of the various types of error that arise in a complex way when linking many data sets.

3. Case Study 1: Redesign of the Building Activity Survey

This case study is an example of a redevelopment project in which the aim was to reduce the amount of direct surveying through the use of administrative data. The process we followed is applicable to surveys in which the desired response variable can be approximated by using a statistical model based on a closely related administrative variable (or variables). A more complete statistical discussion of the changes made to the survey is available in Statistics NZ (2015a, 2015b).

3.1. *Introduction to the Building Activity Survey*

In the past, Building Activity Survey estimates were based on a stratified sample survey. The frame for this survey was of approved construction jobs from local government administrative data on building consents. It used a postal survey to gather information on the value of construction work completed each quarter. The redesign project aimed to replace our building activity sample survey with modelled values derived from the building consents administrative data and the relationship between building consents variables and building activity variables in past data. The redesign aimed to greatly reduce the number of survey forms posted out while maintaining or improving quality. The processing and analysis for the new survey also had to be built on a new software system

because the existing one used legacy tools and software that were very difficult to maintain. This meant that many of the software tools used for coding, editing, and estimation were being updated and improved.

To guide decisions on how much reduction in survey data would be possible without putting data quality at risk, we mapped out the sources of error affecting the old and new designs using our quality framework. The framework was applied in joint collaboration with experienced subject matter analysts. They helped us to understand the issues or problems they encountered in their existing design, including any issues that did not appear to easily fit one category of error or another. The outcome of these discussions was a detailed, organised list of the known sources of error, which was used to understand the impact of the new design.

3.2. *The Three Phases Applied to the Building Activity Survey*

The final outputs of the Building Activity Survey are quarterly tables of the dollar value of work put in place in construction jobs, broken down by several variables, including the type of building and the geographic region. Both old and new designs use building consents data, the source of the building type and other variables, and a survey to collect the value of construction work done in the previous quarter. The building consents data are the selection frame for the survey in both cases, but the new design only surveys large construction jobs.

The findings of the steps described in Subsection 2.3 are summarised below.

Initial metadata collation

Table 1. Summary of the initial metadata collation for the Building Activity Survey.

Information object	Building consents	Building Activity Survey data (before redesign)
Source agency	Local government authorities	Stats NZ
Purpose of data collection	Track new construction work and provide an early indicator of building activity planned throughout New Zealand.	Provide an estimate of the value and volume of work put in place on construction jobs in New Zealand.
Target set	All building consents issued by local authorities in New Zealand with a value of NZD 5,000 or greater.	All construction jobs in New Zealand active during the reference quarter.
Main variables collected	Consent date, consent value, building type, geographic location.	Dollar value of work put in place during the reference quarter.
Mode of collection	Administrative lists requested from each local authority on a monthly basis.	Quarterly (panel) sample survey using building consents as the sampling frame.
Time span of data	1998–present in the current form, historical data from 1965.	1998–present in the current form, historical data from 1965.

Phase 1 evaluation

Phase 1 in this example relates to the building consents data that provide the number and dollar value of construction jobs granted formal approval by local administrative authorities in New Zealand (which we publish as a separate economic indicator series), and the survey data collected by Stats NZ about the construction work actually carried out in each quarter.

Some of the errors arising in this stage are:

Table 2. Examples of the Phase 1 errors which arise in Building Consents and Building Activity Survey Data.

Error type	Building consents	Building Activity Survey data
Validity error	The target concept is the amount recorded on the consent so there is no validity error.	Work done on a job is a well-defined concept easy for respondents to understand, so the question is very closely aligned with our target concept, minimising validity errors.
Measurement error	Values are often rounded down by applicants because there is a financial incentive (lower fees) to have a lower consent value.	Respondents can make mistakes or provide round numbers.
Processing error	The main errors that occur in processing are related to coding: in some cases it is extremely difficult to determine the correct building type based on the description given on the consent.	Processing errors at this point are minimal because the variable – work put in place – is scanned from the survey form. There may be some errors in capturing the information from the form (for example messy handwriting).
Frame error	Cases of consents being given the wrong consent date, and thus not being included in the data extraction for a given month provided to us by the consenting authority.	Some construction work does happen on unconsented jobs, especially small ones.
Selection error	Every consent in the frame is included in the data by definition.	Actual sample drawn from the consents can be incorrect when building consents data contains errors. This results in a building job being placed into the wrong sample stratum. Sampling errors also arise from the random sample drawn in the lower value strata in the old design.
Missing/redundancy error	We do not get missing records on the consents because the consent itself is the unit of interest – any consent issued is available in the data.	Unit and item nonresponse are difficult to distinguish on the Building Activity Survey because (aside from simple confirmations of contact details and so forth) only one statistically important variable is collected on the questionnaire. There is about 10–15% nonresponse to the survey.

Phase 2 evaluation

Phase 2 of the framework applies to the combined unit record data, which is created using the combination of building consents data and survey responses. In both the old and new Building Activity Survey designs, errors arising from data integration are minimal because survey responses are very easily matched to the consent they relate to.

The most important error sources in Phase 2 arise from the corrections for nonresponse and erroneous respondent values, and from the modelling of building work done for small construction jobs below the cut-off and hence deliberately not sampled.

Errors due to editing and item imputation fit clearly into the category of comparability errors (Zhang 2012). The question of where to place the errors arising due to modelling of building work done is more complex. These errors could be considered to be similar to inaccuracies in item imputation, because technically and methodologically the solutions are very similar. Conceptually, however, the two sources of error are quite different. Imputation errors are the result of trying to correct for nonresponse for a variable already being collected using the final harmonised measures. Modelling is a conscious decision not to collect the data in this form and to instead use statistical techniques to convert administrative data into the harmonised measure. Thinking of modelling in this way, it is more closely aligned to mapping error, which arises when "turning primary input-source measures into harmonized measures" (Zhang 2012, 51).

Phase 3 evaluation

Other types of modelling and estimation do not fit this description quite as well, though, so to talk of "modelling error" generically is difficult. This is in part the motivation for introducing a Phase 3 to the framework, which includes modelling and estimation that takes the unit record data as an input and applies adjustments, models, or other techniques to derive final outputs. For the Building Activity Survey, the application of an estimator (Horvitz-Thompson) to the old sample design would be a Phase 3 activity and sampling errors arise as errors at this point.

Estimates for the new design, such as 'total value of work done in the quarter in all of New Zealand', are simpler: a basic sum is taken of the work done for all jobs, since every job has a modelled, surveyed, or imputed value for this variable. Another one of the errors arising in this phase, though, would be from the seasonal adjustment process.

3.3. Examples of Measures Developed and Used

The changes and quality impacts of the redesign fell into two main categories:

1. Changes to existing processes that are needed in the new design.
2. New methodology that would fundamentally change the way estimates were derived.

In the first category, the most important change was in the coding of building type. Building consent forms include an open-ended text box for applicants to describe the construction job they intend to carry out. Under the old system, all building consents were manually coded by a member of the processing team, which required large amounts of

effort and was quite a tedious job. The manual process was generally assumed to have very few errors, but had limited formal evaluation of the quality.

This manual process was replaced by automatic coding using a series of rules that looked for certain words and phrases in combination. To determine whether this new solution was of sufficient quality, the project team developed a set of criteria that focussed on the outcomes of the coding process and the impact on the final estimates. These criteria were used to check the new coding method against the original manual coding for the past ten years of data. They included:

1. Checks on the proportions of building consents coded into high-level categories (residential, non-residential, non-building construction): the criterion was that on a monthly basis the proportion of consents (by dollar value and count) coded to each category should fall within the lower and upper quartile of historical proportions.
2. Proportions by count and value at lower levels of classification, also using the upper and lower quartiles of historical coding as an acceptable range.
3. Specific building types or key words which were required to be coded in a certain way, such as "prison" and "relocated".

These criteria were developed along with the expert analysts, and an iterative process of refining the rules, checking for errors, and determining fixes was carried out until the quality standards were met. Further analysis included examining differences between the time series created using the old and new methodology, such as comparing seasonal adjustment diagnostics to determine whether the seasonal patterns and trends were significantly altered by the changes. In several cases, we found discrepancies due to problems with the codes originally assigned.

For the second category of changes, which included changes in the editing, imputation, sampling, and estimation methodology, we needed to set some criteria on the allowable differences between the old and new methodologies. In the old design we had traditional sampling error estimates, and comparing the old and new time series gave a measure of the accuracy of the new design.

One major challenge was in estimating both the estimation error in the model and the risk that changes in the construction sector might result in our model parameters being outdated and inaccurate. We addressed this challenge with two measures. First, we used bootstrap estimation to produce an estimate of modelling error. This estimate allowed us to fine-tune how many units would still need to be sampled to maintain a similar level of variance as the old design. Second, we ran simulations using the widest plausible range of the modelling parameters to understand the effects on the final time series. These results could then be used to make statements such as "the parameters would have to change by *x*% before the final estimates would fall outside the sampling error range in the old design". By comparing the historic changes over time in these parameters with the impacts of those changes, we could quantify the risks of our methodological changes.

We assessed potential imputation methods in a similar way. We used simulations to develop and test several methods and understand whether the changes would be significant compared with the old sampling errors and the new modelling errors. Having a clearly defined acceptable range of error for comparison was very useful, because it showed us that the choice of a simple imputation method would be more than accurate enough. This

saved us from creating a much more complex and slow solution that would have been less suited to the tools we had available in the processing system.

An important secondary benefit was that many of the measures were suitable to be published as quality indicators for users. We presently publish measures alongside each monthly and quarterly release (see, for example, the June 2015 release of Quarterly Business Activity Survey, Statistics NZ 2015d), which include:

- estimated modelling error,
- proportion by value that is modelled (rather than surveyed),
- imputation rates and proportions.

3.4. *Broader Outcomes of the Application of the Framework*

The workshops and discussions we conducted to understand sources of error and apply the quality framework to the building activity survey redevelopment, also had great benefits for the wider team. Methodologists and subject matter analysts understood clearly where the most critical and important errors might arise, and where more work was needed to control or measure potential new errors introduced, for example errors in the model. Comparing the old and new designs also helped us see existing monitoring or measures that were not effective or valuable and could be removed or replaced, and points where carrying out fixes or edits earlier in the process could reduce work. From a methodological point of view, we had a very detailed picture of the quality effects of different design decisions to guide investigations.

This work helped us to understand trade-offs and make better decisions about the design, and also to prove the value of the framework and demonstrate that we had quality risks under control. The study also had other beneficial side effects.

First, the detailed and comprehensive list of the sources of error affecting the new design compared with the old meant we could alleviate the concerns of users who relied on the existing survey. We clearly described and explained the problems of the old methodology and convinced users that although some time series might be changing, most of the change was due to fixing problems in the old design rather than introducing new errors. The way the framework forces the true statistical target to be clearly stated without reference to our existing measurement of it was very valuable in these discussions.

Second, analysts involved in the discussions understood the "TSE" mindset we brought and took a larger view of the proposed changes. At times, analysts who focus on certain parts of the process are very concerned with maximising the quality of the particular step they are responsible for. While this is not necessarily a bad thing, giving them the opportunity to follow the whole process while explaining the effect of their work on the final statistical quality helped us work together to determine where their effort might have the greatest impact.

4. Case Study 2: Evaluating Administrative Data for Personal Income

This section discusses an application of our framework to income data derived from combining the 2013 New Zealand Census of Population and Dwellings and Stats NZ's Integrated Data Infrastructure (IDI). As with the previous example (redeveloping the

Business Activity Survey), this project is an example of evaluating the potential of administrative data to replace specific survey questions. This may save costs and lower respondent burden. This example also helps to illuminate the challenges that arise from imperfect linkage and coverage of administrative sources and how the limitations of an administrative source can be weighed against the limitations of survey data. It is a good example of how comparing administrative and survey data can shed light on the limitations of both sources, as long as the limitations of each are clearly understood.

The IDI is a collection of linked data sets supplied by various government agencies (including Stats NZ). A key component of the IDI, called the 'spine', is a main data source to which all other person level data sets for research link. The target population for the spine is anybody who has ever resided in New Zealand. At present, the spine is a single list of individuals created by a union of tax, birth, and long-term visa records, to which all other data sets, such as income data from administrative sources, can be linked. For further details about the structure of the IDI, see Black (2016).

Stats NZ has linked 2013 Census records to the spine of the IDI as part of a Census Transformation Programme. The aim of this work is to evaluate the potential for administrative data sources to supplement or replace some census data in the future. The evaluations so far have been relatively quick and exploratory, and used a simplified version of the quality framework, primarily focusing on the coverage of administrative sources and the accuracy of the administrative variables assessed by comparison with census.

One of the most promising studies was a comparison of personal income data collected in the census with personal income from Inland Revenue (New Zealand's tax agency). Our framework can be used to understand the differences between census records and administrative data records on personal income. The results of the investigations can also help Stats NZ understand how to improve measurement of personal income in future censuses.

Initial metadata collation

The census personal income information is collected by two questions. The first asks which sources of income a person has received in the previous year, such as wages and salaries, investment income, or government benefits. The second asks for total gross income from all the sources in the previous year, with the respondent asked to tick the income band they fall into. The bands are roughly in NZD 5,000 increments (NZD 5,000–10,000; NZD 10,000–15,000, etc).

Information on personal income is available as administrative data from Inland Revenue. The data we have access to in the IDI includes tax returns for the self-employed and records from businesses that deduct tax directly from employees' regular pay (Pay As You Earn or PAYE tax), withholding payments (usually relating to contractor's pay), and registers of the main government payments, such as government pensions and unemployment benefits. Each earner in New Zealand has an individual tax number to which their various earnings and tax payments throughout the year are attached. Generally, anybody earning a wage or salary has the amount earned recorded in the tax system, and many government payments are also included. Investment income and superannuation or pension funds other than the main government pension, are not included.

Phase 1 evaluation

In this example the relevant sources of error are on the measurement side of the Phase 1 diagram (Figure 1). We briefly discuss the census income measurement, then move to tax data. Census is a single source, so we only need to consider the Phase 1 diagram.

The concept of personal income is clearly defined in the census in a technical sense: gross annual income from all sources. This is the target concept in the framework. The questions used to operationally collect this information are very well-aligned to this concept, although they make some compromises. In particular, the banded totals mean that the results are "blurred" compared with the exact, true amount.

Including bands rather than a specific dollar value makes it easier for the respondent to answer. However, many measurement errors are still possible (and observed) in the census responses. Item nonresponse is a problem, with only about 83 per cent of working-age respondents having a valid response. Other common measurement errors include:

- confusing gross with net income,
- recall errors when someone does not remember receiving income from a particular source,
- approximations made by respondents, such as rating up their latest pay cheque to an annual figure when they also received bonus payments or pay increases; or roughly mentally rounding their income and pushing themselves into a different band,
- proxy responses where a household member responds on behalf of another and does not precisely know how much money their housemate earns,
- mistakes when summing all sources, or when rating up the net pay cheque (for example the amount actually on a bank statement) to a gross amount.

Deliberate over- or under reporting of income is also a possible source of measurement error. In our investigations limited evidence of this occurs, in part because high incomes are covered by only a small number of wide income-band checkbox options. A general trend towards underestimation at all income levels seems to be stronger than any effect from deliberately overstating incomes at low levels.

Some potential measurement errors, such as respondents making factors of 100 errors when cents are or are not included, are reduced by using income-band tick boxes. The income bands also encourage more response, since people might know their income very confidently to within a few thousand dollars but not a precise amount. This is a good example of trading off different errors against each other: the bands result in some uncertainty, but also make it easier for people to respond and hopefully reduce measurement errors.

Processing errors are minor compared with other types of error because the tick-box responses are easy to code. In the 2013 Census, few important edits were made on responses, so processing errors are small contributors to the total error.

To assess the administrative income data, we made use of both Phase 1 and Phase 2 of the framework because income data comes from different sources and is not collected to measure personal income. The general process would be to understand the precise purpose of the administrative collection and determine what can go wrong within the administrative agency with respect to that purpose. For income, the variables we are concerned with are also the most crucial for administrative purposes. For instance, pension

payments are recorded so that the government can accurately track those entitled to receive payments, and company payroll tax is audited to ensure the correct tax amount is paid to the government.

If we want to understand all the sources of error fully, we need to look at all the particular administrative processes and constraints in different agencies. For example, do systematic errors (such as under reporting or processing mistakes) occur in pension data but not in personal tax returns? Earlier studies by Stats NZ suggest these errors are small and that administrative measures are very good measures of the administrative concepts (such as amount of pension paid, amount paid to an employee during a tax period). A significant practical issue is that processing for some sources like tax data takes considerable time, which means there can be a delay of several months (or more) until full records for a given date are available.

Phase 2 evaluation

Phase 2 of the framework focuses on errors that arise when data sets from several sources are integrated to produce an output that meets a certain statistical purpose.

Data integration is done using unique identifiers (tax numbers) from administrative data sets. In order to use administrative data to impute (or completely replace) the current census income question, we need to link the administrative data belonging to each individual to the right census respondent. In our prototype linkage we were able to link about 94 per cent of people to the IDI spine, with a false positive rate of about 0.7 per cent. Low-quality linking information (primarily names and dates of birth) is the main reason for not linking to the spine, but there are also several sources of undercoverage in the administrative data which mean that some people who filled in the census are not included in the administrative data at all.

One source of undercoverage is from individuals working in the "underground" market. Roemer (2002) integrated administrative data on workers earnings with earnings data from the United States Census Bureau's March Current Population Survey (CPS) and showed missing earnings from the administrative data. Earnings missing from the administrative data are exhibited across all wage sizes but are prominent across certain occupations.

In addition to linkage error, other coverage errors result from the mismatch between the tax population and the New Zealand census night population. People can be filing tax returns from overseas in some cases, causing overcoverage, although using tax data for only those census people who we link to the administrative data will help alleviate this problem.

On the other hand, people who receive income only from investments or untaxed sources may not appear in the tax data, causing undercoverage. The same error could be better described as a relevance error in some cases, such as if a person is present in the data but has no income recorded in the tax data. Unlike the census, the tax income measure does not include all sources of income.

Phase 3 evaluation

Given the high link rates, the crucial question about using administrative data over census income data is whether the conceptual mismatch between administrative data and the standardised statistical definitions results in more error than the problems caused by measurement error in the census. Note that errors in administrative data are considered to

be relevance errors in Phase 2 of the framework, while census errors are Phase 1 measurement errors that have flowed through to the final data.

These comparisons require an appreciation that the census (or any other existing source) is subject to its own errors, and that a difference between administrative data and an existing survey is not in itself proof of error in the administrative data. The comparison (as far as possible) must be between the statistical ideal and the different data sources we have. At times we consider census data to be a 'gold standard' whose results must be exactly reproduced by administrative data. In many cases the comparison with administrative data can help us understand the limitations of the gold standard. Some findings have already resulted in suggestions for improving the census questionnaire, where we can empirically show that many respondents are making similar mistakes.

For example, the missing sources of income in administrative data will cause a systematic underestimate of total income. Is this underestimate greater than that resulting from imperfect recall by census respondents? Using the linked census–IDI data, we compared the figures from the two sources. We found that even with some income sources missing from administrative data, census income responses were typically lower. This is a good argument for administrative data, along with issues of nonresponse and the lack of precision from the banded responses in the census which prevents analysis of income distributions in more detail.

It is useful to compare this investigation to that from the *Canberra Group Handbook on Household Income Statistics* (UNECE 2011a). The handbook contains detailed and careful descriptions of errors known to arise in measuring income. It allows us to clearly define our target concepts and populations so that we have a sound basis to compare against both census and administrative data. Generally, the sources of error mentioned in the handbook are similar to what we described above. Our framework puts these errors into a TSE and statistical design context in a systematic way, helping to make the evaluation of errors more practical and allowing for comparisons of the relative influence of different error sources.

5. Case Study 3: Population Estimation in New Zealand

The aim of Stats NZ's population estimates is to produce an accurate count of the number of people who usually live in New Zealand at a certain reference date. Our published population estimates are based on a variety of sources, including the five yearly Census of Population and Dwellings and some administrative data.

It is possible to use the IDI data directly (independently of the census) to produce estimates of the size of the New Zealand population. As part of our Census Transformation initiative, assessments and studies have been carried out to assess how accurately the population can be estimated from administrative sources (Gibb et al. 2016). The goal of this work is not to replace existing estimates (at least not yet), but to understand the limitations of the administrative data so that progress can be made towards improving our own processes in combining and using available administrative data. Another major goal is to identify which sources are more reliable, and whether there are any significant issues with the administrative data which could be fixed by source agencies.

This case study is included here to demonstrate how the three phase framework can be used to understand the complex interplay between coverage and linking errors. It is also a

demonstration of the start of what we see as a very promising path for continuous improvement of estimates derived from complex combinations of administrative data where there are many known and significant sources of error.

Initial metadata collation – definition of population

The New Zealand official Estimated Resident Population (ERP), defined as the "estimate of all people who usually live in New Zealand at a given date", was about 4.8 million people at the start of 2017 (Statistics NZ 2016d).

The target population of the IDI spine lists any person who has ever lived in New Zealand, and currently contains about nine million people (Black 2016). In order to generate a list of people who usually live in New Zealand that can be compared with the official ERP, we use a set of rules to restrict the spine to only those people who reside here as of a certain date. The resulting list is called the IDI-ERP. It is derived by selecting only those people in the spine who have shown recent activity in one of the administrative data sets linked to the spine. For example, those who have filed a tax return or have interacted with the health system during the previous twelve months, or who were born less than five years ago, are included in the IDI-ERP. The rules also take other information into account, such as death registrations and data about people who have travelled overseas and not returned.

Phase 1 evaluation

In this example, Phase 1 of the framework applies to each of the source data sets integrated in the IDI. For the purposes of population estimation, many of the administrative variables are not important, but some have measurement errors that affect estimates. First, measurement errors in linking variables such as names and dates of birth result in links not being made in the IDI processing. Errors in other major demographic variables (sex, ethnicity, and address) do not affect overall population counts but cause inaccuracies in subpopulation estimates.

New Zealand Customs data is a good example of the complex effects of measurement errors. Passengers crossing New Zealand borders complete arrival/departure cards that are collected by Customs officers. To identify that someone has left the country and later returned, their departure and arrival cards must be linked. Errors in scanning or recording names on these cards, or respondents writing incorrect or changed details (such as different spellings of a name transliterated from another language) can flow through to population estimation. In many cases, the records can be linked using passport numbers, but people may travel on different passports or renew their passports resulting in a different number that is not necessarily recorded in the administrative data.

Another crucial measurement error arises when we create subnational population estimates using administrative address information to assign people to different locations. Here many problems arise, such as out-of-date addresses, missing or poor-quality addresses that cannot be accurately geocoded to a certain location, and conflicts between different administrative sources that must be resolved. Some errors might be ignored by the agency: for instance, if the tax department wants someone's address, but enters the address of their accountant instead, this could be considered a validity error (depending on

what the tax agency's true purpose for collection is). Unless it results in difficulties with getting the right amount of tax paid by the person, they are unlikely to correct it. Similarly, if an agency's usual contact with an individual is by cellphone or email, they might never check if the address in the person's file is a valid one.

Phase 2 evaluation

The most obvious and significant sources of error are in the representation side of Phase 2. The linked sets are created by identifying records across multiple data sources that we believe belong to the same individual. Identification errors and unit errors are not an issue in this case because we are not creating new statistical units, but using the linked list directly as our list of units. The only error of concern is coverage error, but this can arise in many ways. In some cases, different sources of error can cause similar net effects on population counts and yet require different treatment.

A simple coverage error is when the available data does not include a person from the target population. The visa data we have access to starts from 1997, so if a couple moved to New Zealand in 1990 and only the husband has ever paid tax, the wife might not be included in the spine at all. Overcoverage is also possible because in some situations overseas residents could be paying tax to the New Zealand tax agency, and could look like an active resident under the IDI-ERP rules. Both the IDI-ERP time-band rules for activity and the time lag in updates of the spine, create issues in classifying a person as a 'usual resident' which causes coverage errors.

Linkage errors are also a major source of error. False negative links (for example when the link between someone's birth and tax records is not made due to a name change) effectively cause duplicates in the population. Some of these duplicates will be removed by the activity rules, but there are many complex possibilities. For some reason, if a person's (active) health record is linked to their birth record, but their (also active) tax record is not, two separate and active records for one person will exist.

False positive links can have different effects. If someone who has moved overseas is erroneously linked to an accurate health record of someone with a similar name, we may get overcoverage. But if a person is falsely linked to a departing immigration record, they may be removed from the population, causing undercoverage. Depending on how the rules for inclusion and exclusion are defined and which one takes precedence, linking errors between particular data sets will result in different effects.

Phase 3 evaluation – the estimation phase

The population estimation problem highlights that the end point of Phase 2 is the final integrated microdata, rather than the final estimates derived from this data. No matter how much effort we spend on improving our processes and data, our final integrated data set will have significant amounts of undercoverage and overcoverage. Therefore, we need to devise an estimation procedure that can correct these errors. Within Stats NZ's Census Transformation project, Bryant and Graham (2015) described one attempt to construct such a model using multiple administrative data sets. However, a conclusion of this work was the need for an independent sample survey to assist with coverage estimation.

Conceptually, the problem can be described by considering a large population, the union of the total coverage of the various administrative populations with the target resident population. The problem is to construct a model that describes which individuals on the administrative data make their way into the final data set, and which target population individuals are not represented in any of the administrative data sets. The processes that lead to somebody not being present in a given data set are part of the model, as are missing or erroneous variables (for example errors in ethnicity measurement). Parameters such as coverage rates can then be estimated and used to create a final estimate of the total population correcting for undercoverage, overcoverage, and other errors.

A complete model taking all error sources into account is still a work in progress, but this approach has a clear synergy with the error framework described in this article. If we have measures for some of the errors in the administrative data, these can be used to improve this model. Conversely, if a particular source of error (for example overcoverage in a particular administrative data set) is poorly understood, the model can give us some insight into how much uncertainty this causes in the final estimates. We can then make decisions about where to target our efforts; either by helping an agency improve their data, studying the coverage in more detail, or running coverage surveys to target measures for improving our overall estimates of the population size.

5.1. *Phase 3 and Continuous Improvement of Population Estimation*

The process for producing Stats NZ's official population estimates following a Census provides a good example of the usefulness of the Phase 3 concept. We can consider the final unit record census data after all editing, imputation, and other processing (the so-called "clean unit record file") to be the outcome of Phases 1 and 2 in the census, where error arising from combining administrative data and survey data have been incorporated in Phase 2. Most output tables produced for New Zealand's 2013 Census were based on tabulating the relevant variables from this clean unit record file.

However, in deriving the base estimated resident population counts, results of the Post-Enumeration Survey (PES) were used to correct and adjust for the estimated undercount in the Census. These final results do not come directly from data integration between the PES and Census unit record data, although data integration is a part of the process. Instead, the PES allows for coverage rates to be estimated, and these rates, as well as the raw counts from the Census data, are used as part of an estimation method that aims to produce more accurate counts of the population than the raw data alone. These estimates are updated in the period between population censuses using administrative sources such as birth, death, and immigration records, which are again incorporated into an overall estimation model.

Work continues at Stats NZ to improve population estimation and understand the sources of uncertainty in population estimates and projections. See for example Bryant et al. (2016), and Statistics NZ (2016c). Evaluations of errors in individual administrative data sets, the IDI linking process, census data, and coverage surveys can all be captured in a systematic way using our framework and this adds to our understanding of the quality of our final estimates. The models developed so far can be expanded to include new sources of error as we improve our understanding of the input data and linking processes.

6. Summary and Discussion

The quality assessment framework discussed in this article facilitates the reuse of both existing data and previous quality assessments. This was successfully demonstrated in the three case studies. The framework supports Stats NZ's goal to use administrative data first. The basic idea behind the framework is that with a clear understanding of both the limitations of all source data sets, and of the way errors propagate through our statistical production processes we can obtain a complete picture of the quality of the final output. Measuring an error is the first step to correcting it. We need to separate what the collecting agency has done from our own processes and what users intend to do with the data.

Phase 1 of the framework focuses on how well a data set meets its original, intended purpose. This information is valuable for anyone who wishes to investigate whether the data can meet any other needs. The framework can provide a common language for talking about data quality issues, and be a valuable decision-making resource for the organisation. This also applies for users outside Stats NZ, when data is reused and shared for research purposes. The framework and documentation is a pedagogical instrument to help explain a data source so that researchers and other users can determine how useful or suitable it might be for their own purposes. Besides helping users, applying the framework also raises internal awareness at Stats NZ of quality and sources of errors.

Phase 2 of the framework deals with the problems that can arise when integrating data sets from different sources during processes like transforming the original variables to match statistical needs and identifying and creating statistical units from integrated data sets. The reference point in the quality assessment in Phase 2 is the statistical population we would ideally have access to, and the statistical concepts we want to measure about the units in the target population. The measurement side in Phase 2 is concerned with how variables from each source data set are reconciled. This may differ in various ways from the target concepts. The representation side is about creating a set of statistical units from the objects in the original data sets.

Phase 3 of the framework focuses on estimation, design, and evaluation. The aim is to determine the data source(s) that can minimise the cumulative effect of errors on output statistics produced from integrated or combined data in Phase 2. If there are no integrated data sets but two or more alternative data sources (thus making Phase 2 redundant), then assessments from Phase 1 can be used to determine the best statistical design. The Phase 3 investigation can also provide a list of quality risks that need to be mitigated or checked over time to ensure the consistency of the resulting statistics. For statistics that the organisation can influence, this gives valuable input into which/how production/data generation processes can be improved.

The framework provides a list of measures and indicators that can be used to quantify key aspects of data quality. The measures can be used during the design phase of a survey to determine if survey needs have been met, during statistical production to monitor the process, and for dissemination to explain the quality of a statistical output to users. They can also be used to provide feedback on the improvement of the input data sets, including suppliers of administrative data. The measures do not cover every possible situation, but give a starting point and ideas for more detailed or technically complex measures that

could be developed for specific outputs. The framework also helps us prioritise further work so that investigations can be focused on the most crucial quality issues.

From our experience, the generic or standardised lists of measures can be very useful for initial input quality evaluation and for output reporting. Stats NZ also publishes output quality reports based on a standard list of required information (for example see Statistics NZ 2015c.)

When we make technical design decisions, we often have to develop more customised measures depending on the details of the design, population, and variables. Some of these measures are important for understanding output quality, such as measuring the uncertainty in modelled estimates instead of sample survey sampling errors. In many cases, specialised measures are needed to understand particular sources of error. We advocate flexibility in the measures and indicators used, and recognise that in some cases no satisfactory way exists to measure the effect of a given source of error.

An interesting future area of work will be to develop estimation models that can work in a positive feedback loop with our error assessments. The Bayesian estimation framework (Bryant and Graham 2015) may be one way to do this. We would like to be able to use our three phase quality framework to identify sources of error that could be built into the estimation model. The model would then give us a way to isolate the effects of each error on the final estimate, so that we can focus further improvements on the areas which have the largest impact, whether that is advocating for input data set improvements or processing improvements.

In trial applications like the Building Activity Survey, we found that our quality framework was a useful tool for teaching analysts about quality and TSE concepts. Analysts responsible for statistical production may be extremely knowledgeable about the types of error that occur in their data without having a methodologist's understanding of end-to-end effects of design and data quality. The framework allows their extensive practical knowledge to be translated into standardised and structured metadata, which other people can use to investigate data reuse. It also helps the analysts think about the connection between the initial user needs that are met by their output and the effects their decisions have on data quality.

To get a full picture of the quality of statistical outputs that reuse data not originally intended for official statistics, we also need to measure the improvements in processing costs, respondent burden, and other aspects of statistical production. Issues such as public attitudes towards data integration and the risk of relying on outside data suppliers also need to be considered by decision makers. We intend our framework to be an expanding information bank as Stats NZ gains access to more administrative data. A shared understanding of what data is useful for what purposes, captured with the help of our framework, will increase the pace at which both Stats NZ and data users can get the most value from new data sources and outputs.

Appendix A

Here are definitions of terms and quality indicators and measures useful to measure Phase 1 of the quality assessment framework.

Representation Side

Target set is the set of all objects the data producer would ideally have data on. This includes, for example, people, businesses, events, and transactions.

Accessible set is the set of objects from which measurements can be taken in theory.

Accessed set is the set of objects for which measurements are obtained in practice. For example, the electoral roll doesn't include people who fail to enroll despite being legally entitled, or whose forms get lost in the mail.

Observed set is the set of objects that end up in the final, verified data set after all processing by the source agency.

Frame error is the difference between the ideal target set of objects and the accessible set. These errors refer to objects that are inaccessible even in principle. In a survey context the accessible set is the sampling frame. For an administrative source, objects may be inaccessible for a variety of reasons.

Table A1. Quality Indicators for Frame Errors

Quality indicator / measure	Definition
Lag in updating population changes	Delays in registration.
Undercoverage	When units in the target population are not on the accessible set.
Overcoverage	When units in the accessible set are not in the target population.
Authenticity	Percentage of records in the administrative data with an incorrect identifier key, including records with multiple identification keys.

Selection errors arise when objects in the accessible set do not appear in the accessed set. For example, if a store manager forgets to run the reporting tool for a week then the transactions missing from the data set due to that mistake will be selection errors: they were accessible, but were not accessed.

Table A2. Quality Indicators for Selection Errors

Quality indicator / measure	Definition
Adherence to reporting period	Proportion of units that provide data for a different period than the required reporting period for the administrative data set. This may be due to lags, delay, or non-compliance with reporting period.
Dynamics of births and deaths	Changes in birth and death rates of units in the data over time.
Readability	Proportion of records that can be accessed using existing software for reading data.
Inconsistent objects/units	Proportion of units that are (and cannot be made) internally inconsistent. Examples are objects involved in non-logical relations with other (aggregates of) objects in the data source.

Missing/redundancy errors arise from the misalignment between the accessed set and the observed set. For example, errors where an agency mistakenly rejects or duplicates objects due to their own processing could mean that objects are missing from the data set even though correct data was received about them. This category of error exists so that such errors are kept distinct from reporting-type errors.

Table A3. Quality Indicators for Missing/redundancy Errors

Quality indicator/measure	Definition
Unit nonresponse rate	Fraction of units missing in the data source.
% of duplicate records	Proportion of duplicate records present in the data.
% of units that have to be adjusted to create statistical units	Proportion of units that have to be adjusted to create statistical units. For example, the proportion of data at enterprise group level, which needs to be split to provide reporting unit data.

Measurement Side

Target concept is 'the ideal information that is sought about an object'. The target concept is usually connected to the underlying purpose of the collection and may be quite abstract. Examples could include household income, political views, advertising effectiveness, or population counts.

Target measure is the operational measurement used in practice by a source agency to capture information. A target measure can include elements such as variable definitions, classifications, a questionnaire, or rules and instructions for people filing out forms.

Obtained measures are the values initially received for specific variables against objects in the data set.

Edited measure refers to the final values that are recorded in an administrative or survey data set, after any processing, validation, and other checks.

Validity error refers to misalignment between the ideal target information and the operational 'target measure' used to collect it. The error arising from the translation from an abstract target concept or 'the ideal information sought from the administrative data set about an object' to a concrete target measure that can actually be observed in practice, and does not include issues such as misunderstanding a term used on a form.

Table A4. Quality Indicators for Validity Errors

Quality indicator / measure	Definition
% of items that deviate from target concept definition	Fraction of items from the administrative data that deviate from the target concepts. In this context, 'items' are variables or fields entered on the final unit record data set.
% of items that deviate from Stats NZ/international standards or definitions	Proportion of items from the administrative data that deviate from Stats NZ / international standards or definitions.
% of inconsistent records	Proportion of units (or records) from the administrative data that violate logical, legal, accounting, or structural relationships between variables in a record.
% of items affected by respondent comprehension of questions asked in collection process	Proportion of items from the administrative data affected by the quality of questions in the data collection process.

Measurement errors occur when the obtained measure (the value actually recorded in the data set) differs from the measurement intended. These could include people misremembering details or interpreting the questions differently from how they were designed. In more automated administrative systems, such as electronic transaction records, measurement errors could include computer system problems that corrupt some values or introduce ambiguity.

Table A5. Quality Indicators for Measurement Errors

Quality indicator / measure	Definition
Item nonresponse	Fraction of missing values for a variable.
Percentage of records from proxies	Proportion of units from the administrative data whose data were provided by proxies.
Lagged time between reference period and receipt of data	Lapsed time between the end of the reference period and the time of receipt of the data source.
% of units in administrative data which fail checks	The proportion of units that fail one or more edits.

Processing errors arise from editing and other processing carried out by the source agency to correct or change the initial values received (the obtained measures).

This kind of processing is usually intended to improve the quality of the data with respect to the target concept, but it is important to understand how much improvement the processing makes, as well as any limitations introduced by the processing.

Table A6. Quality Indicators for Processing Errors

Quality indicator / measure	Definition
% of transcription errors	The proportion of units of a variable coded or recorded incorrectly.
Modification rate	The rate of editing changes done on a variable. Editing changes refer to changes to non-missing, values being changed to other non-missing values, which in most cases will be the result of editing.
Item imputation rate	Fraction of the values of a variable modified by editing and imputation by the administrative data provider.

Appendix B

Here are definitions of terms and the quality indicators and measures that apply to the error sources from Phase 2 of the quality assessment framework.

Representation Side

Target population is the ideal set of statistical units a final data set should cover.

The **linked sets** include all the basic objects from across all source data sets that are matched together to make base units. These units will not necessarily be the final statistical units of the output.

Aligned sets are the groups of base units which have been determined (after linking and other processing) to belong to each composite unit in a final output data set. For instance, we might create household units based on dwelling units and person units. In this case, the aligned sets could be represented by a table that contains all these relationships (for example Household 1 consists of dwelling A and persons X, Y, Z, Household 2 consists of dwelling B and person W, etc.).

Statistical units are the entities for which information is sought and for which statistics are ultimately compiled. These units can, in turn, be divided into observation units and analytical units (OECD, 2007).

Coverage errors are the differences between the units actually included in the linked data sets in practice (linked set) and the full set of units included in the (ideal) target population. Coverage errors can arise in several ways. For example, the data sets themselves may not cover the whole target population, or linking errors may mean some members of the linked sets are not identified. This error may also be caused by measurement errors. For example, if the date of birth variable on an administrative data set is not of good quality and we are filtering on age to select our population, we could end up with undercoverage even though the units are not missing from the source data.

Table B1. Quality Indicators for Coverage Errors

Quality indicator / Measure	Definition
Undercoverage	The proportion of units in the target population that are missing from the final data sets.
Overcoverage	Overcoverage occurs when units that are not in the target population are present in the final linked data.
Percentage link rate	The fraction of objects in each data set that can be connected with units in other data sets.
Proportion of duplicated records in the linked data	The fraction of units duplicated in the linked data.
False positive and false negative rates	False positives are record pairs deemed to be links but are actually true non-matches. False negatives are true matches that remain unlinked.
Delay in reporting	The time difference between the period each data set relates to and when you receive the final data set.

Identification error refers to the misalignment between the linked set and the aligned set. This type of error also includes situations where the target statistical units cannot be adequately represented using combinations of base units. For example, if we wanted to measure the economic activity of all manufacturing businesses by industry, we would ideally have separate statistical units to capture different types of manufacturing done by a single company. However, in practice we might have to define statistical units via legal entities. Changes in company or legal structures might result in statistical units being absorbed into others, despite no real-world change in economic activity occurring.

Table B2. Quality Indicators for Identification Errors

Quality indicator / measure	Definition
Proportion of units with conflicting information	Proportion of linked units that contain conflicts that need to be resolved during the production process.
Proportion of units with mixed or predominance-based classifications	When assigning objects from the input data sets to composite units, a single classification may have to be assigned to the composite unit based on the properties of the base objects that make it up. If the underlying units fit under one classification code, this decision will be simple. If they don't, the decision may be based on predominance, importance, or some other decision rule. However the decision is made, the units will not completely capture the properties of the real-world object they represent. A simple indicator of the quality of the final classification is the proportion of units for which such a decision must be made.
Rates of unit change from period to period	For many statistical outputs, the target population changes relatively slowly, so significant changes in the units in the input data sets may indicate quality problems with the data, linking, or other aspects of the process. This indicator is a simple measure of the rate of change of the population.

Unit errors are introduced when the final statistical units are created for the output data set. For instance, to create household units from the aligned sets of dwellings and people we must simultaneously decide which dwellings should have a household created, and which people should go into which household unit. Because statistical units may not correspond to any of the units in the source data, a variety of errors can arise at this stage.

Table B3. Quality Indicators for Unit Error

Quality indicator / measure	Definition
Proportion of units that may belong to more than one composite unit	The fraction of units that don't have a single clear composite unit to which they can be assigned without doubt. This could be units that cannot be assigned to any composite unit for some reason, or units equally likely to belong to two different composite units.

Measurement Side

Target concept is 'the ideal information that is sought about the statistical units'. The target concept is usually connected to the underlying purpose of the collection and may be quite abstract. Examples could include household income, political views, advertising effectiveness, or population counts.

The **harmonised measures** are the operational measures decided upon in the design of the statistical output to capture the target concepts. They include elements such as questions, classifications, and variable definitions. A common example would be a survey question aligned with a standard classification.

Re-classified measures are the values of the harmonised measures.

Adjusted measures refer to the final values in an integrated microdata, after any processing, validation, and other checks.

Relevance errors are errors at a conceptual level that arise from the fact that the concrete harmonised measure usually fails to precisely capture the abstract statistical target concept. For example, if we want to find out about personal income but decide that in practice we will only measure taxable income, this creates a relevance error since non-taxable income is part of our target concept but not our harmonised measure.

Table B4. Quality Indicators for Relevance Errors

Quality indicator / measure	Definition
Percentage of items that deviate from Stats NZ / international standards or definitions	Proportion of items in the final data set that deviate from Stats NZ / international standards or definitions.

Mapping errors arise from the transformation of variables on the input data sets into output variables that have been defined (the harmonized measures). These could include transformations like:

1. Reclassification from a non-standard classification, or coding a free text field.
2. Derivation of a numerical variable from a source data set, such as removing gross sales tax from a transaction value.
3. Modelling of a target variable using a combination of several variables on a source data set and some model parameters.

In each of these cases the value of the output variable may differ from the true value, and these differences are mapping errors.

Table B5. Quality Indicators for Mapping Errors

Quality indicator / Measure	Definition
Proportion of items that require reclassification or mapping.	Fraction of the variables on the input data set that requires transformation into relevant output variables.
Proportion of units that cannot be clearly classified or mapped.	Fraction of units of which values of its target variables cannot be clearly determined using classification rules.
Inconsistency of variable definitions in linked data	Differences in variable definitions across linked data sets.
Indicators and measures of modelling error	If the output design involves modelling a target variable using one or more of the original data set variables, this introduces errors. These errors can be measured, but the method to do this depends on the chosen model. Many indicators can be applied to statistical modelling. A few examples are goodness-of-fit tests (for example R-squared), confidence intervals of model parameters, or for Bayesian models credible intervals).

Comparability error arises from editing and other treatment methods applied to values obtained from reclassified measures – to correct for missing values, inconsistencies, or invalid values.

Table B6. Quality Indicators for Comparability Errors

Quality indicator / Measure	Definition
Proportion of units failing edit checks	The fraction of units, of the total units checked, failing one or more edits.
Proportion of units with imputed values	The proportion of units that have been imputed.

18. References

Australian Bureau of Statistics. 2009. *The Australian Bureau of Statistics Data Quality Framework*. Canberra: Australian Bureau of Statistics. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0. (accessed June 2013).

Bakker, B. 2012. "Estimating the Validity of Administrative Variables." *Statistica Neerlandica* 66: 8–17. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00504.x.

Biemer, P. 2010. "Total Survey Error: Design, Implementation and Evaluation." *Public Opinion Quarterly* 74: 817–848. Doi: http://poq.oxfordjournals.org/content/74/5/817.full.pdf + html10.1093/poq/nfq058.

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley.

Black, A. 2016. *The IDI prototype spine's creation and coverage*. Available at: http://www.stats.govt.nz/methods/research-papers/working-papers-original/idi-prototype-spine. (accessed August 2016).

Bryant, J., K. Dunstan, P. Graham, N. Matheson-Dunning, E. Shrosbree, and R. Speirs. 2016. *Measuring Uncertainty in the 2013-Base Estimated Resident Population* (Stats NZ Working Paper No 16-04). Available at: http://www.stats.govt.nz/methods/research-papers/working-papers-original/measure-uncertainty-2013-erp.aspx (accessed March 2017).

Bryant, J. and P. Graham. 2015. "A Bayesian Approach to Population Estimation with Administrative Data." *Journal of Official Statistics* 31: 475–487. Doi: http://dx.doi.org/10.1515/JOS-2015-0028.

Burger, J., J. Davies, D. Lewis, A. van Delden, P. Daas, and J.-M. Frost. 2013. *Deliverable 6.5/2011: Final List of Quality Indicators and Associated Guidance*, Report for Work Package 6 of the ESSnet on the Use of Administrative and Accounts Data for Business Statistics. Luxembourg: Eurostat. Available at: https://ec.europa.eu/eurostat/cros/system/files/SGA%202011_Deliverable_6.5.pdf_en. (accessed August 2016).

Daas, P.J.H., S.J.L. Ossen, and M. Tennekes. 2010. "Determination of Administrative Data Quality: Recent Results and New Developments." In Proceedings of the Q2010 European Conference on Quality in Official Statistics, May 4–6, 2010. Available at: http://www.pietdaas.nl/beta/pubs/pubs/Q2010_Session34_presentation.pdf. (accessed June 2012).

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, A. Bernardi, F. Cerroni, T. Laitila, A. Wallgren, and B. Wallgren. 2011. *Deliverable 4.1: List of Quality Groups and Indicators Identified for Administrative Data Sources*, Report for Work Package 4 of the European Commission 7th Framework program BLUE-ETS. Brussels: European Commission. Available at: http://www.blue-ets.istat.it/index.php?id=7. (accessed December 2015).

Daas, P., S. Ossen, and M. Tennekes. 2012. *Deliverable 4.3: Quality Report Card for Administrative Data Sources Including Guidelines and Prototype of an Automated Version*, Report for Work Package 4 of the European Commission 7th Framework program BLUE-ETS. Brussels: European Commission. Available at: http://www.blue-ets.istat.it/index.php?id=7. (accessed December 2015).

Gibb, S., C. Bycroft, and N. Matheson-Dunning. 2016. *Identifying the New Zealand Resident Population in the Integrated Data Infrastructure (IDI)*. Available at: http://www.stats.govt.nz/methods/research-papers/topss/identifying-nz-resident-pop-in-idi.aspx. (accessed August 2016).

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74: 849–879. Doi: http://poq.oxfordjournals.org/content/74/5/849.full.pdf+html10.1093/poq/nfq065.

Laitila, T. and A. Holmberg. 2010. "Comparison of Sample and Register Survey Estimators via MSE Decomposition." In Proceedings of the Q2010 European Conference on Quality in Official Statistics, May 4–6, 2010. Available at: http://q2010.stat.fi/sessions/special-session-34/. (accessed December 2015).

Office for National Statistics. 2013. London: Office for National Statistics. *Guidelines for Measuring Statistical Quality*. Newport: Office for National Statistics. Available at: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html. (accessed February 2017).

Organization for Economic Cooperation and Development. 2007. *OECD Glossary of Statistical Terms*. Paris: OECD. Available at: https://stats.oecd.org/glossary/index.htm. (accessed August 2016).

Roemer, M. 2002. *Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation*. Maryland: U.S. Census Bureau. (Technical paper No. TP-2002-22). Available at: https://www2.census.gov/ces/tp/tp-2002-22.pdf. (accessed September 2016).

Scholtus, S. and B.F.M. Bakker. 2013. *Estimating the Validity of Administrative and Survey Variables Through Structural Equation Modelling: A Simulation Study on Robustness*. The Hague / Heerlen: Statistics Netherlands. (1572-0314, no - 201302).

Smith, T.W. 2011. "Refining the Total Survey Error Perspective." *International Journal of Public Opinion Quarterly* 23: 464–484. Doi: http://ijpor.oxfordjournals.org/content/23/4/464.short/ 10.1093/ijpor/edq052.

Statistics Canada. 2009. *Statistics Canada Quality Guidelines*. Ontario: Statistics Canada. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf. (accessed June 2013).

Statistics NZ. 2015a. *Implementing Classification and Other Changes to Building Consent Statistics*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/Construction/building-consent-changes-2015.aspx. (accessed January 2016).

Statistics NZ. 2015b. *Methodology and Classification Changes to Value of Building Work Put in Place Statistics*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/Construction/methodology-classification-changes-value-building-work.aspx (accessed January 2016).

Statistics NZ. 2015c. *Retail Trade Survey: September 2015 Quarter, Data Quality Section*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/RetailTrade/RetailTradeSurvey_HOTPSep15qtr/Data%20Quality.aspx (accessed January 2016).

Statistics NZ. 2015d. *Value of Building Work Put in Place: June 2015 Quarter*. Available at: http://www.stats.govt.nz/browse_for_stats/industry_sectors/Construction/ValueOfBuildingWork_HOTPJun15qtr/Data%20Quality.aspx. (accessed August 2016).

Statistics NZ. 2016a. *Guide to Reporting on Administrative Data Quality*. Available at: http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx (accessed at August 2016).

Statistics NZ. 2016b. *Our Strategic Direction*. Available at: http://www.stats.govt.nz/about_us/who-we-are/our-strategic-direction.aspx. (accessed August 2016).

Statistics NZ. 2016c. *How Accurate are Population Estimates and Projections? An Evaluation of Statistics New Zealand Population Estimates and Projections, 1996–2013*. Available at: http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/how-accurate-pop-estimates-projns-1996-2013.aspx (accessed March 2017).

Statistics NZ. 2016d. *Standard for population terms*. Available at: http://www.stats.govt.nz/browse_for_stats/population/standard-pop-terms.aspx (accessed May 2017).

United Nations Economic Commission for Europe. 2011a. *Canberra Group Handbook on Household. Income Statistics*. New York and Geneva: United Nations. Available at: http://www.unece.org/index.php?id=28894 (accessed December 2015).

United Nations Economic Commission for Europe. 2011b. *Using Administrative and Secondary Sources for Official Statistics – A Handbook of Principles and Practices*. New York and Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf. (accessed December 2015).

Wallgren, A. and B. Wallgren. 2014. *Register-Based Statistics:Sstatistical Methods for Administrative Data*, 2nd ed. Chichester: Wiley.

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Received January 2016
Revised March 2017
Accepted March 2017