Appendix 3: Quality indicators for phase 2 errors

Guide to reporting on administrative data quality is the source document for appendix 3.

Appendix 3 is a companion to the phase 1 quality indicators document (appendix 2). It lists 19 quality indicators and measures that apply to the error sources from phase 2 of the error framework. Generally you'll need to assess the errors in phase 1 before the sources of error in phase 2 can be properly understood.

Phase 2 categorises the difficulties that arise from taking variables and objects from input datasets and using them to measure the statistical target concept and population we're interested in. In phase 2 we consider what we're trying to do with the data, and determine how well the input datasets match up with what we'd ideally be measuring.

As with the phase 1 list of indicators, not all the phase 2 indicators will be useful for a particular output. Only calculate and use the indicators that will help you understand the data's quality.

We've organised the indicators in this document into six groups, based on the error types in the phase 2 framework.

Representation (units)

The **representation** side in phase 2 is concerned with identifying, linking, and creating statistical units using the input datasets. In a perfect situation, the final dataset would contain one unit for each member of the target population, and these units would be connected to units or objects in each input dataset – so all relevant variables could be included next to the appropriate units. The indicators listed below provide ways to understand and measure the extent to which this perfect situation is satisfied in practice.

Note: for some outputs linking and data integration is a routine process. For instance, Statistics NZ's business survey designs often incorporate admin datasets that have to be linked together, but this linking is done using very reliable identifiers (eg IRD numbers). For other admin datasets there may be no linking, just processing and converting from the raw input dataset variables 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 admin objects to statistical units, can still be valuable.

Error types for representation

The representation side of phase 2 deals with creating a list of statistical units to include in the output data, based on the objects of the input data. Individual datasets may be based on transactions or events, which must be connected together and placed into newly-created statistical units that relate to customers, households, or other entities of interest in the statistical target population.

To apply this phase 2 framework, we must define the target statistical population precisely. This population is the ideal set of statistical units that the dataset should cover. We must also understand the nature of the objects in each input dataset and their relationship to the statistical units. Do we

expect each statistical unit to have a single corresponding unit in each input dataset, or do many transaction records need to be combined for a single statistical unit?

Coverage error

To evaluate coverage error we need to determine what the linked sets are. The linked sets include all the basic objects from all input datasets that are matched together to make base units. Note: these may not be the final statistical units.

Example: we link a range of datasets, such as tax returns, survey responses, and immigration records, at the person level, and use the linked person-level dataset to produce the final target statistical units, such as households or families using the linked dataset at a later stage.

The coverage of the final linked sets depends primarily on two factors: the proportion of units in the target population with a corresponding entry in the input datasets, and the quality of the linking that connects the datasets. Each input dataset might be intended to cover a certain part of the target population. In these cases, we evaluate each dataset for a certain part of the reference dataset; the combination of all datasets could also be evaluated as a whole.

Low coverage is not necessarily a sign that a dataset should not be used. For example, if the dataset's undercoverage is well understood and consistent then it may be possible to correct quite accurately for it, and produce high-quality counts data. Use coverage rates and other similar measures to help towards the goal of improving quality rather than as strict criteria for accepting or rejecting datasets.

Measures for coverage

Undercoverage and overcoverage

Description
Undercoverage refers to the proportion of units in the target population that are missing from the datasets available. It can apply to both individual datasets and to final datasets. Overcoverage occurs when units that aren't in the target population are present in individual datasets or the final linked data. It is important to distinguish problems of under- and overcoverage, since a dataset that suffers from both may appear to have the right **number** of units, even though it doesn't represent the target population well. Example: if we use a list of a country's taxpayers to measure the permanent resident population, we would miss permanent residents who don't pay tax (undercoverage), and include non-permanent residents who do pay tax (overcoverage) – reporting the raw count of people compared with a census count makes the dataset look better for this purpose than it really is. Where a reliable population count or list is available (eg census or a business frame) use this as the **reference dataset.** When this isn't available, compare each individual dataset and the largest or most comprehensive dataset available. Sometimes a linked dataset is created by linking several datasets to a large 'base' dataset; in this case record the coverage relative to the base dataset. For datasets that aren't being linked, and for which there is no comparison dataset, record any known information about the types of units that might be missing. In the coverage measures, we use the word 'unit'. In some cases (eg event databases) it may be more appropriate to consider the coverage for the 'objects' present in the phase 1 datasets.

1. Undercoverage

How to calculate

$$\frac{\textit{Number of expected units not in dataset}}{\textit{Number of units in reference dataset}} \textbf{ x } 100\%$$

In this formula, **expected units** refers to units that the dataset under investigation would ideally contain. For example, to measure the undercoverage of the electoral roll population, we could compare it with the census population count for people over 18. It's often more informative to report a dataset's known coverage limitations, and measure undercoverage relative to the sub-population we expect each dataset to cover where we know these limitations.

Apply this simple measure of undercoverage at a total level or for sub-populations (regions, area units, industries, or other breakdowns of interest). Having measures of the coverage of each individual dataset before processing may be useful, as well as an overall measure – based on the final linked records in the output statistical dataset, including all linking and processing carried out while creating the final list of statistical units.

Ideally, the **reference dataset** would include the entire target population, but usually it's only possible to compare the populations of the different input datasets with each other. This can still provide useful information about which datasets include more (or fewer) units than others, across the subgroups of interest.

2. Overcoverage

How to calculate

$$\frac{\textit{Number of unexpected units in dataset}}{\textit{Number of units in reference dataset}} \textbf{ x } 100\%$$

In this formula, **unexpected units** refers to units that the dataset under investigation should ideally not include. This formula measures how many extra units are in the dataset that we don't want in our final dataset. This is an overall measure that includes duplicates, erroneous records, units outside the scope of the target population, and so on.

If a dataset is known to include groups that are not part of the target population, we can aim to remove them during processing and linking. It might be more informative to report on overcoverage rates after such processing is completed, while noting the units we've excluded.

Apply this measure at a total level or for sub-populations (regions, area units, industries, or other breakdowns of interest). Calculate it for each individual dataset, and at an overall level based on the final linked records produced by combining various datasets.

As with undercoverage, ideally the **reference dataset** would include the entire target population, but usually it's only possible to compare the populations of the different input datasets with each other, or with the best dataset available. This still provides useful information about which datasets include more (or fewer) units than others, across the subgroups of interest.

3. Proportion of units linked from each dataset to a base dataset, or percentage link rates between pairs of datasets

Description
The link rate is a very important measure of the quality of the linked sets (sets of objects from different datasets that were linked together). This measures the proportion of objects in each dataset that can be connected with units in the other datasets. A low link rate may indicate that different datasets cover different parts of the population, or that the linking process is not identifying all the connections that exist between the objects in each dataset.

How to calculate
$$\frac{\textbf{\textit{Number of units linked to base dataset}}}{\textbf{\textit{Number of units in dataset that is being linked}}} \textbf{ x 100}\%$$ Where a unique identifier is available on both datasets we'd expect a link rate very close to 100 percent. If probabilistic matching is being used, the link rate will be lower. There's no strict rule about an acceptable link rate, but use this measure to assess which datasets are causing more problems at the linking stage.

4. Proportion of duplicated records in the linked data

Description
Some datasets may contain erroneous duplicate records that disrupt the linking process and/or the final dataset. Knowing how many units in each dataset are duplicated (or how many are detected as being duplicates) is a useful indicator of the underlying dataset's quality: a very good, well-checked and maintained dataset should have very few duplicates, but one produced for other reasons may not have had the same care taken. If we detect duplicates, they must be resolved in some way. Mistakes in this process can result in errors in the final data, such as when duplicates differ in some variables and we need to choose one set of values.

How to calculate
$$\frac{\textbf{\textit{Number of units found to have duplicates}}}{\textbf{\textit{Number of units in dataset}}} \textbf{ x 100}\%$$ In practice we can only compute this measure based on the number of duplicates we detect. More undetected duplicates are likely in large datasets so the number may not represent the true duplicate rate, but it is a useful indicator.

5. Precision and recall in linking

Description
Precision and recall are measures of the quality of data linking. **Precision** measures how well the links are made, and is a measure of the 'goodness' of links made. **Recall** measures how well links are found, and is a measure of how many true links were captured correctly. These measures are based on false positive and false negative rates. False positives are record pairs deemed to be links but which are actually true non-matches. False negatives are true matches that remain unlinked. In practice, false negatives are much harder to identify so it may not be very efficient to use the false negative rate as a regular indicator.

How to calculate
Precision: $$precision = 1 - false\ positive\ rate$$ $$= 1 - \frac{Number\ of\ false\ positives\ found}{Number\ of\ linked\ records\ checked}\ \mathbf{x}\ 100\%$$ **Recall:** $$recall = 1 - false\ negative\ rate$$ There is no easy way to automatically measure linkage error rates, so false positive rates have to be estimated by manually checking samples of linked records. In large datasets, analysis of false positives can be time-consuming work – it's useful to group the linked data before selecting a sample. False negative rates are typically even more difficult to measure. See data integration manual **f**or more detail and information on assessing linkage quality.

6. Macro-level comparisons of the distribution of linked objects with reference distributions

Description
By comparing the distribution of units across key variables in the linked datasets with those in a reference dataset, we determine whether the linked set is missing important parts of the target population, or at least whether it represents the target population well. Note: distinguishing the effects that coverage and variable definitions have on variable distributions may be tricky. See indicator 16 for more about this.

How to calculate
There is no fixed formula for this measure. It requires reference data to compare each input dataset (or the final linked set) with. Depending on the reference data available we can make different comparisons, such as: • total values of key variables compared with reference totals • mean values of key variables compared with reference values • if datasets are being linked to a base set, compare the detailed distributions of variables (eg age, income, or sales) in each set to be linked with the distribution in the base set – using histograms or other graphical methods. We compare totals, counts, or means at different levels of aggregation if the reference dataset allows this. An example is comparing population counts from admin datasets at a regional level with census reference counts.

7. Delay in reporting

Description
There is always some delay between creating a dataset and publishing statistics derived from it. Each individual dataset that goes into the final statistical output may be delayed by a different length of time, which can affect the datasets' coverage relative to how the real world is when the output is published. Aim to measure and record these delays and, where possible, their effect on population coverage.

How to calculate
As a basic indicator, record the time difference between the period each dataset relates to and when you receive the final dataset. For example, if March quarter data is received in September this represents a six-month delay. For datasets that are available in an early but incomplete version, report the coverage at different delay times. For example, a dataset that is 100 percent complete after 12 months might be available after six months, but with only 75 percent coverage of the final version. Report this in the same way as for coverage (above).

8. Linking methodology used

Description
This is a qualitative indicator to record the type of linking used to connect records on different datasets. Include details of which datasets were linked with a unique identifier available on both sets, and which were linked using probabilistic methods. Record the details of which variables were used to match datasets, and the methods used to resolve difficult cases (eg manual inspection, random assignment, some kind of scoring). These details are valuable to help understand where quality problems with the linked datasets may be important.

How to calculate
No calculations are required for this indicator. Diagrams and descriptions of the structure of the datasets used and the process for linking them together may be used.

Identification error

The process of determining how all the linked sets relate to each other, so that statistical units can be created, is called 'alignment'. Depending on the nature of the units linked together, we may want to create 'composite units', which are made up of one or more 'base units'. Identification errors include the mistakes made during this process.

Different datasets may conflict and we need to decide how to resolve the conflict. For example, if we have person-level data linked by a common identifier across several datasets and we want to form groups of people living at the same address. If the different datasets contain different addresses for the same person, we may make identification errors when deciding on a single address for the person.

Identification errors also include situations where we can't adequately represent the target statistical units by using combinations of base units. A problem arises when we construct statistical units from legal entities such as businesses. To measure the economic activity of all manufacturing businesses, we would ideally have a separate unit for each manufacturing company reporting only manufacturing activity. But in practice a legal company might perform other activities, even if it is predominantly a manufacturer. Changes in company or legal structures can result in statistical units being absorbed into others – ideally we'd measure the units separately. Such issues also result in identification errors.

Measures for identification error

9. Proportion of units with conflicting information

Description
If the objects linked together from two different datasets contain conflicting information, this can indicate the objects are not truly the same. This indicator measures how many linked units contain conflicts we need to resolve during the production process.

How to calculate
$$\frac{\textit{Number of units with conflicting information}}{\textit{Number of units checked}} \text{ x } 100\%$$
Methods for determining if units have conflicting information vary. The simplest measure is to record how many fixes or decisions we made (either manually or automatically) during processing. For example, in linking person-level data from multiple datasets each dataset may have its own address, sex, or date of birth fields. Even a very high probability match may have a disagreement – because of different definitions, reporting periods, or other mistakes.
Note: the focus of this indicator is not on the accuracy of the variables, but the consistency of the units created by linking different datasets. It measures how reliably the units represent the underlying target population units.

10. Proportion of units with mixed or predominance-based classifications

Description
When assigning objects from the input datasets to composite units, we may assign a single classification to the composite unit, based on the properties of the base objects that make it up. If the underlying units all fit under one classification code this is a simple decision, but if they don't, the decision may be based on predominance, importance, or another decision rule. Whichever rule we use, these 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 we need to decide for.

How to calculate
$$\frac{\textit{Number of units which fit more than one classification}}{\textit{Total number of units}} \text{ x } 100\%$$
Measuring this indicator requires knowledge of how the classifications were assigned. If a variety of different rules are applied, record how many units are assigned using each decision rule.

11. Rates of unit change from period to period

Description
For many statistical outputs, the target population changes relatively slowly, so any significant changes in the units in input datasets may indicate quality problems with the data, linking, or other aspects of the process. This indicator measures the rate of change in the population.

How to calculate
Birth rate: $$\frac{\textbf{Number of unit births in output period}}{\textbf{Total number of units}} \textbf{ x 100\%}$$ Death rate: $$\frac{\textbf{Number of unit deaths in output}}{\textbf{Total number of units}} \textbf{ x 100\%}$$ Either the birth or death rates may be of interest, or use an overall measure of the number of units changing their population status. It is difficult to give guidance for acceptable rates of population change because this depends strongly on the output's design and the target population. Calculate an expected rate based on prior knowledge to use this indicator to check for unexpected population change.

Unit error

The final statistical units in the output dataset may be created from scratch, with no direct correspondence to any units in the input datasets. From a set of people linked by address, we could create dwelling units that include everyone living at each unique address. But we might create dwelling units that don't exist, or not create a unit for a given dwelling. Because the list of dwelling units is generated from the information in the other datasets, we can distinguish these errors from linking errors. We simultaneously determine which addresses should actually be given a dwelling unit and which people should be connected to each dwelling unit.

Measures for unit error

12. Proportion of units that may belong to more than one composite unit

Description
Creating the list of statistical units in a final output dataset may require connecting together 'base' units (the lowest-level units created during the linking process, or the units on the input datasets). For example, to create a list of household units we might link several input datasets to produce a list of base person units, then link all the person units with the same address to form composite household units. This indicator records how often a base unit (eg a person) doesn't have a single clear composite unit to which it can be assigned without doubt. This could be units that can't be assigned to any composite unit for some reason, or units equally likely to belong to two different composite units.

How to calculate
$$\frac{\textbf{Number of units that cannot be assigned to a composite unit}}{\textbf{Total number of units}} \textbf{ x 100\%}$$ Depending on how assignment is done, calculate this indicator by simply counting unassigned units. If the assignment process forces difficult units into a composite unit, then modify the measure to count units that can't be exactly matched with a composite unit, or units assigned with different levels of certainty.

The indicators for coverage of the base units may also be applied to the final statistical units if they are different from the base units. For example, at the linking stage we could compare person counts with census person counts, then when household units are generated, we could compare household counts with census household counts. If reference data is available, we can also examine the distributions of variables.

Measurement (variables)

Measurement errors in phase 2 result mostly from a mismatch (eg in concept, definitions, classifications) between the variables on the original input datasets and the target concept that our final output is aiming for. In an ideal situation, variables are collected using classifications and questions that match what we'd use to collect the same information from a specialised survey.

This part of the error framework is intended to assess the issues that arise when we use variables from the input datasets in ways they weren't originally designed for.

Error types for measurement

To understand measurement errors the ideal information we want from each statistical unit must be defined clearly. This is the target concept. In phase 2, we compare the original target concepts for relevant variables from the phase 1 assessment with the target statistical concepts. The definitions, classifications, decision rules, and other processes used to produce the final values of each variable, in the final dataset, also need to be well understood. This gives a full picture of how well the final output variables match the intended purpose of the statistical dataset.

Relevance error

Harmonised measures are the practical measures we use to measure the target concept in the final data. Relevance errors are the conceptual mismatches resulting from which practical measures we chose, and the rules chosen to align differing variables into a common measure.

In practice, measuring relevance error is very similar to measuring validity error in phase 1. The major difference is that we assess the statistical target concept and the measures and classifications chosen to measure it. Very similar measures are often used to measure both types of error. Note: relevance errors are a conceptual mismatch between the statistical target concept and the measures decided on for the statistical output. We address the question of how well the input datasets match with the harmonised measures in mapping errors (below).

Measures for relevance error

13. Percentage of items that deviate from Statistics NZ/international standards or definitions

Description
This indicator provides information on the proportion of items in the final dataset that deviate from Statistics NZ/international standards or definitions. In this context 'items' are variables or fields to be filled on the final unit record dataset – using some combination of input dataset variables. This indicator presumes that Statistics NZ/international definitions are the 'gold standard' for measuring a given target concept. If the output is designed (possibly as a compromise) to use a non-standard classification or measure, it's a sign that quality issues may arise, particularly when comparing with other outputs. Metadata about variable definitions is very important when output variables differ from standards. For example, personal income variables on different datasets can include/exclude some income sources – be clear what exactly is included to avoid misinterpretation.

How to calculate
$$\frac{\textit{No. of items that deviates from agreed standards or definitions}}{\textit{Total no. of items}} \text{ x } 100\%$$ Note: When assessing the usability of the variables for a statistical output, we can weight this indicator for whether or not the variables are key to the statistical output.

Mapping error

The variable values on the input datasets must be converted into values of the harmonised measures – the reclassified measures. This process produces mapping errors. An example is converting a free-text occupation field on an employee register into a standard job classification. Here, values of the raw input dataset variable may be given incorrect standard classifications due to ambiguous titles. Mapping error also covers more complex problems, such as uncertainties in modelling a target concept from the values of variables in the input datasets.

Measures for mapping error

14. Proportion of items that require reclassification or mapping

Description
Where an input dataset already contains information that precisely aligns with the harmonised measures decided on to capture the target concept, we would expect the corresponding variables on the output dataset to be of higher quality than those derived from a non-standard variable. This indicator is a way to measure how much work is required to transform the variables on the input dataset into relevant output variables.

How to calculate
$$\frac{\textit{Number of variables or measures captured directly on source datasets}}{\textit{Total number of variables/measures}} \text{ x } 100\%$$ This measure may be modified to take account of key output variables only; for instance, if the final dataset contains extra variables that aren't of high importance to the published output. Also record qualitative information about this measure, such as a list of the output variables that details variables with an exact correspondence to one on an input dataset, which ones are derived in a simple way, and which ones require complex models or assumptions.

15. Proportion of units that can't be clearly classified or mapped

Description
This indicator shows how well the rules for determining values or classifications for the target variables are working. If we've units that don't entirely align with the chosen decision rules or units, which need a default or fallback value, this can indicate a lack of accuracy in the final data. This applies to units where the classification is uncertain; for example, where the automatic coding system (or an editor) chooses between two equally likely standard classifications – for a unit with an unclear value for a free text variable.

How to calculate

$$\frac{\textit{Number of units with unclear values}}{\textit{Total number of units}} \text{ x } 100\%$$

Measuring this indicator depends on the variable and the rules being used. If the variable is categorical then use the proportion of units to be assigned a classification based on random selection or default values. For a numerical variable, this indicator could measure the proportion of units whose original values don't fit the intended rules (eg units that have negative modelled values for a variable that should always be positive).

16. Distribution of variables in linked data

Description

Analysis can be done to compare the values of variables in different datasets at various levels of aggregation. This measure isn't easy to give general rules for but we can use statistical data inspection methods to compare measures of variables present in at least two datasets. Histograms and scatter plots are graphical methods to use. Numerical measures such as means, medians, standard deviations, and skewness also help compare distributions of values.

This indicator is similar to number 5 in the representation indicators. It's not always easy to distinguish problems caused by differences in variable definitions or reporting from those caused by datasets' different coverage. On the measurement side, we'd compare related variable values for the same unit or a set of units – to identify differences between the variables. On the representation side the focus is on the overall distribution of values from all units, to determine whether target units are missing.

Example: use this indicator to compare a sales variable from an admin dataset with the equivalent variable from a survey dataset – to understand how well the admin variable matches our 'gold standard' survey variable. The representation side indicator might look at the overall distribution of the sales variable in the admin dataset to see if it matches the distribution of the survey sales variable. The aim is to identify if any parts of the survey population aren't present in the admin data, or vice versa.

How to calculate

There's no fixed formula for calculating this measure. It's useful when you've more than one dataset with a variable related to the target concept; for example, different measures of income. Examples of possible types of analysis follow.

- Calculate basic distribution parameters (eg maximum, minimum, mean, standard deviation) for variables present in more than one dataset, and compare these.
- Plot a scatter plot of the values of two similar variables from two datasets. For example, if each contains an age variable, this plot will show any problem with the values in one dataset or a systematic difference between the datasets.

Assuming the populations covered in the datasets are the same, finding differing distributions of variables intended to measure similar concepts is a valuable measure of the effect of differences between the measures.

17. Indicators and measures of modelling error

Description
If the output design involves modelling a target variable using one or more of the original dataset variables, this introduces errors. These errors can be measured, but the method to do this depends on the chosen model.

How to calculate
Many indicators can be applied to statistical modelling. A few examples are goodness-of-fit tests (eg R-squared), confidence intervals of model parameters, modelled values, totals or other quantities of interest (bootstrapping could be used, or for Bayesian models use credible intervals). The most appropriate method depends on the nature of the model. Ideally, choose a measure that relates to the uncertainty in the final values of interest resulting from the model.

Comparability error

The final stage in producing values of the output variables for each unit is to create adjusted measures. These are the final values after editing, imputation, and other changes. In phase 2 there may be discrepancies between two datasets even if the original data is correct from the point of view of the data producer. These discrepancies could be due to reporting dates or other differences in procedures or rules. In phase 2, we must resolve any conflicts or disagreements to create a single final value – mistakes made in sorting out these issues are comparability errors.

Measures for comparability error

18. Proportion of units failing edit checks

Description
This standard indicator can be used for all kinds of output. Use edit rules to detect inconsistencies or likely errors in unit values – the proportion of units failing one or more edits is a good indicator of any issues such as reporting or processing inconsistencies (between or within datasets). In phase 2, comparability errors include final editing and imputation done after the mapping and classification stage. Depending on the output's design, many errors in the input data may be fixed during the mapping stage, so the divide between these two error types is not always completely clear. Generally the term 'comparability error' applies where the values of two variables for the same unit are not consistent with each other, or where missing data or obvious errors in the data need to be repaired. Example: we derive quarterly income for a person from employee tax records, but find the person is also classified as unemployed in data on an unemployment register. For the final output, we'd have to decide whether to count the person as employed (they received income from an employer) or unemployed (they are registered for unemployment). Another common example is where we classify and map all the units we can, then use these derived values to impute for missing values in the remaining units.

How to calculate
$$\frac{\textit{Number of units failing edit checks}}{\textit{Total number of units checked}} \text{ x } 100\%$$
This rate is a standard measure for monitoring statistical survey processing, but can be used whenever edits are used. Apply this indicator to each edit check, or a class of edit checks that are likely to pick up problems with an aspect of the data. We can separate edit checks that look at values derived from one dataset from those picking up inconsistencies between linked datasets.

19. Proportion of units with imputed values

Description
This is a standard indicator for surveys that use imputation. Knowing the proportion of units imputed with each imputation method, and the proportion of the total values of interest derived from imputation, is valuable for assessing accuracy of the final figures.

How to calculate
$$\frac{\textit{Number of units with imputed values}}{\textit{Total number of units}} \text{ x } 100\%$$
or
$$\frac{\textit{Total value of imputed units}}{\textit{Overall total of the variable}} \text{ x } 100\%$$
Either the proportion of units or the proportion of value may be useful indicators, since they relate to different things. If a large number of very small units requires imputation, this indicates a different kind of issue from a few very large units needing imputation.

Published by Statistics New Zealand
February 2016
www.stats.govt.nz