

Quality Framework for Data Integration

Quality Assessment Framework (WP5 Validating Official Statistics, Coordinator: Felipa Zabala, New Zealand)

- The quality assessment framework, including the quality indicators, described in [Guide to reporting the quality of administrative data](#) is beneficial in carrying out validation studies. The quality framework is based on Li-Chun Zhang's [two-phase life-cycle model for integrated statistical microdata](#) (Figure 1) which expands the total survey error paradigm to include administrative data. The framework enables understanding of the error sources from the individual data sources including those arising from the integrated datasets. Zhang's two-phase life-cycle model assists in determining the associated methodological and operational issues that may impact on quality resulting from producing statistical information from linked administrative data sources. Phase 1 assesses the quality of an input data source that is intended to be used in the

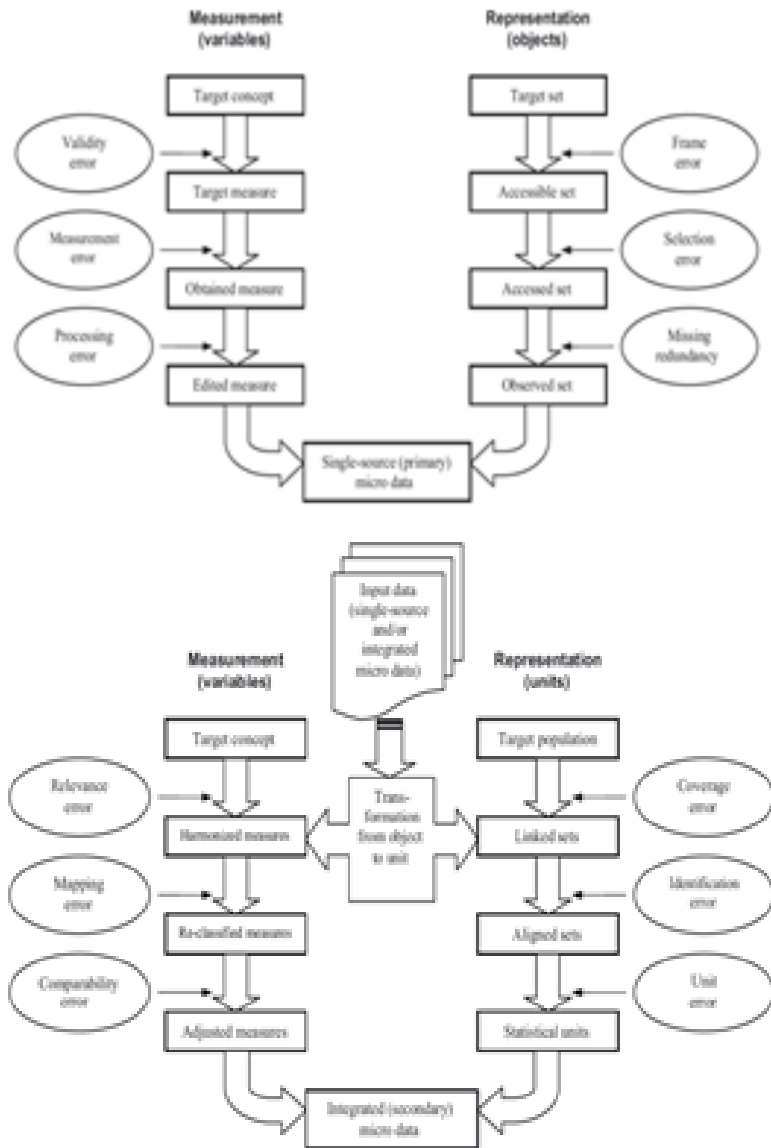


Figure 4 Zhang's two-phase life cycle model

production of a statistical product. An NSO needs to understand the design decisions undertaken by the producers of the source to determine methods to turn the data into the statistical information required by the NSO. Quality of the input data source is assessed against the purpose for which it was collected. For a survey dataset, this purpose is defined for a statistical target concept and target population. For an external data source, the entries or 'objects' in the dataset might be people or businesses, but they could also be transaction records, or other events of relevance to the collecting agency. At this stage, evaluation is entirely with reference to the dataset itself, and does not depend on what an NSO intends to do with the data. Quality issues in the input data source will flow through into any use of the data in the production of a statistical product.

Phase 2 categorises the difficulties arising from taking variables and objects from source datasets and using them to measure the statistical target concept and population an NSO are interested in. In this phase, the NSO consider what they want to do with the data, and determine how well the source datasets match what they would ideally be measuring.

The quality assessment involves 3 steps.

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

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

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

- Understanding the error sources from the input data sources, including those arising from the linked datasets, is essential in determining the associated methodological and operational issues that may limit the use of administrative data in the validation of official statistics.
- The key quality dimension to use is accuracy. The normalised root mean square error is a useful tool to evaluate the accuracy of statistics sourced from integrated datasets. Laitila and Holmberg (2010) suggest an approach to estimating the total error of an estimator from integrated dataset by deriving lower and upper boundaries for a total mean square error measure.
- The [Guide to reporting the quality of administrative data](#) provides a metadata information template that encourages thinking about the key aspects of quality in an organised way. It is also a convenient way to record a standard set of information to compare different datasets. You do not necessarily need all the meta information. The basic information required are: name of data source agency, purpose of data collection, time period covered by the data, the population (target and actual) population of the dataset, the reporting units, a short description of key variables and the timing/delay information and method of collection.