



How fit is your data?

25 June 2009

Narrisa Gilbert

narrisa.gilbert@abs.gov.au

Melissa Gare

m.gare@abs.gov.au

ABS Data Quality Framework!

- History of ABS DQF
- Brief discussion of 7 dimensions
- Examples of how to use it

- \$10,000 to buy a car

History of ABS DQF

- Adopted from two frameworks – Statistics Canada and Eurostat
- Originally six dimensions – Relevance, Timeliness, Accuracy, Coherence, Interpretability, Accessibility
- Additional dimension (2007): Institutional Environment

Dimension 1: Institutional Environment

- The institutional and organisational factors that have a significant influence on the effectiveness and credibility of the agency producing statistics
- Questions to assess:
 - Which organisation(s) have supplied the data? What sort of organisation is this?
 - Is statistical confidentiality guaranteed, and if so, under what authority?
 - Impartiality / Independence
 - Advertised release dates in advance

Dimension 2: Relevance

- How well the statistical release meets the needs of users in terms of the concepts(s) measured, and the populations(s) represented.
- Questions to assess:
 - About whom, or what, were the data collected?
 - How useful are these data at small levels of geography?
 - Does this data source provide all the relevant items or variables of interest? Does the population presented by the data match the data need?

Dimension 3: Timeliness

- The delay between the reference period (to which the data pertains) and the date on which the data becomes available; and
- the delay between advertised date and the date at which the data become available (i.e., the actual release date).
- Questions to assess:
 - Are there likely to be subsequent surveys or data collection issues for this topic?
 - What is the gap between the reference period, the time when the data were collected, and the time when the data became available?

Dimension 4: Accuracy

- Refers to the degree to which the data correctly describe the phenomena they were designed to measure.
- Questions to assess:
 - What controls or checks are in place to address data entry error, recognition error or editing error?
 - Are there particular questions which are hard to understand and to which respondents may provide an inaccurate response?
 - Have the data been adjusted in any way to account for non-response?

Dimension 5: Coherence

- Refers to the internal consistency of a statistical collection, product or release, as well as its comparability with other sources of information within a broad analytical framework and over time.
- Questions to assess:
 - To what extent can a user meaningfully compare several data items within this collection?
 - Have these data been confronted with other data sources, and are the messages consistent from all data sources?

Dimension 6: Interpretability

- Refers to the availability of information to help provide insight into the data
- Questions to assess:
 - Are terms used in the data release which are confusing or ambiguous for a user?
 - Is there information available to help the user gauge the potential magnitude of error in the data?
 - To what extent can a user of the release or dataset find supporting information about the data to enable improved interpretation?

Dimension 7: Accessibility

- Refers to the ease of access to data by users
- Questions to assess:
 - How easily can a user obtain this information? Is it publicly available?
 - What range of products are available, and what are their costs?
 - Are the data available in suitable formats?

Uses of the ABS DQF

- A data user:
 - Defining a data need
 - Assessing a dataset in the context of a data need (includes idea of assessing a quality statement)
- A data producer:
 - Assess the quality of the data produced (Creation of a quality statement)
 - Design a statistical collection or product which is fit for purpose (not covered today)

Example: A User's Perspective

- Data Need: Number of children at school in Australia.
- ABS DQF: What questions would YOU as a **user** want to ask to get clarification for your data need and to also determine what data source best fits your need?

User: Institutional Environment

- Who (organisation) collects this information?
- Is it a private, public, not for profit etc. organisation?
- What authority/legislation was the data collected? E.g. is it mandatory?

User: Relevance

- What age group(s) am I interested in?
- Am I interested in school years rather than age?
- What level of geography do I require? E.g. Australia, State level, Local level?
- What type of schools am I interested in? E.g. public, private, primary, secondary, university
- Am I interested in a sub-population E.g. Indigenous, Regional Data

User: Timeliness

- What is the year/time frame am I after?
- When is the data collected?
- When do I need the data by?

User: Accuracy

- What issues are associated with the particular method of collection?
- Is all the data collected in the same way? E.g. is there more than one form? Is it a mix of Survey and Administrative record?
- Has imputation been used on this data for missing values/non-answers?
- How was the data collected? E.g. Personal interview, report taken.

User: Coherence

- Is this data comparable over time with itself?
E.g. can I get a 'time series'?
- Are other data sources available for comparison with this source?
- Has there been a change to any definitions?
- Has there been a change to laws that may influence the starting school age / school retention age?
- What about State comparison?
- What about international comparison?

User: Interpretability

- What standard classifications have been used (if any)?
- Is there supporting information to explain concepts, methodologies?

User: Accessibility

- What format(s) are the data in?
- What isn't available? E.g. level of detail not obtainable?
- How much is it going to cost me?

Clarification of need

- The above list of questions will help a User clarify what data they are after.
- The above list of questions will also help a User determine where there are data gaps i.e. data not available to meet all specifications/needs

Uses of the ABS DQF

- A data user:
 - Defining a data need
 - Assessing a data set in the context of a data need (includes idea of assessing a quality statement)
- A data producer:
 - Assess the quality of the data produced (Creation of a quality statement)
 - Design a statistical collection or product which is fit for purpose (not covered today)

Example: Producer perspective

- What are the most important things to tell a User about your data?
- Use the ABS Data Quality Framework with the User in mind.

Example: Apparent retention rate

- What is good about this Quality Statement?
- What could be improved?

Data Quality Statement

- **Indicator:**
 - Apparent retention rate of Indigenous students from Year 7/8 to Year 10
- **Measure- Annual proxy:**
 - Numerator: Number of full-time Indigenous persons in Year 10 in the reference year (2008)
 - Denominator: Number of full-time Indigenous persons in the base year (Year 7 in NSW, Vic, Tas and ACT in 2005 and Year 8 in Qld, WA, SA and NT in 2006)

Data Source/s

- The National Schools Statistics Collection (NSSC) provides annual counts for the numerator and denominator with disaggregation by Indigenous status. For information on the NSSC scope and coverage, see [NSSC Explanatory Notes](#).

Institutional Environment

- Data on government and non-government schools are collected by the ABS through the non-finance National Schools Statistics Collection (NSSC), which was established through the work of the Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA).
- For information on the institutional environment of the ABS, including the legislative obligations of the ABS, which cover this collection, please see [ABS Institutional Environment](#).

Relevance

- The NSSC collects information on enrolment for all years of schooling. Disaggregation by State and Territory and by Indigenous status is available. Socioeconomic status information is not currently available.
- This indicator measures the proportion of Indigenous students who commenced secondary school, who have remained in secondary school to undertake Year 10, that is, an Apparent Retention Rate (ARR). It is not a measure of the proportion of Indigenous students who actually completed Year 10.

Timeliness

- The NSSC is conducted annually in August. The results from NSSC 2008 were released in March 2009.

Accuracy

- As a census, the NSSC has a high response rate. The time lapse between actual movements of students, and receipt and entry of data about such movements, results in a small percentage of duplication of student records. A small percentage of students may have left school but have not yet had their records altered at the time of the census to reflect this change.
- Care should be taken in the interpretation of ARR as the method of calculation does not take into account a range of factors such as repeating students, migration, inter-sector transfers and enrolment policies. The ARR measures change over a period of time (three years). Given the long analysis period, student transitions, such as migration or re-entry to the school system, may have an effect on the accuracy of the calculation.
- Explanatory notes are available regarding the accuracy of the NSSC methodology and ARRs, see: [Explanatory Notes](#).

Coherence

- The ARR is based on those who are undertaking study at the Year 10 level as at August in the reference year and they may not go on to complete Year 10. The NSSC data items used to construct the ARRs are consistent and comparable over time, and support assessment of annual change.

Interpretability

- Information is available for the NSSC to aid interpretation of the data. See National Schools Statistics Collection on the ABS website.

Accessibility

- See National Schools Statistics Collection for standard products available. Data are also available on request. The annual proxy measure is available on the ABS website as a standard product from the NSSC.