A Quality Framework for Longitudinal Studies

Peter Lynn, ULSC

DRAFT

Last update 26-9-2001

A Quality Framework for Longitudinal Studies

Contents

1. Introduction

This document sets out a framework for assessment of the quality of UK academic longitudinal studies. It is a conceptual framework, mapping in a structured way the concepts subsumed by the notion of quality. Most of the concepts would apply to any type of survey, but academic longitudinal studies, particularly those with study horizons of decades rather than months and years, have characteristics that raise the importance of some components of quality relative to others and influence the way that we think about others. Some of those particular characteristics are:

- Relatively high levels of sample attrition, especially amongst certain sub-groups of analytic interest. Non-response bias is an important issue for any survey, but cumulative non-response over many waves of a study can be particularly serious.

- Multiple and changing definitions of the study population. The study population changes over time and between analyses. Issues relate to the treatment of deaths, births, emigrants, immigrants and so on.

- The impact of item non-response. Many longitudinal analyses rely on measures collected on a number of occasions. The proportion of sample members providing a valid measure on every occasion can be very considerably lower than the proportion providing a measure on any one occasion, thus dramatically reducing the available sample size.

- Changing relevance of data items and variables. Questions thought to be particularly important at the start of a study may become rather less relevant years later, and *vice versa*. Long-term longitudinal studies can therefore encounter conflicts between relevance and consistency.

- Changes in technology and in the research team. For studies taking place over a number of decades, changes in the available technology for survey processing, data management and release of outputs can be considerable. This has obvious implications for compatibility and consistency. Similarly, "ownership" of the study is likely to change hands, not just between persons, but between organisations and geographical locations, with similar implications.

Our framework takes as its starting point some more general approaches to quality adopted by others. These approaches have then been adapted to the specific context as appropriate.

Over the past decade, the conceptualisation of survey quality by major survey and statistical organisations has been converging. Many organisations now have quality guidelines/ standards/ procedures/ documents that are organised under six (or seven) dimensions. The definitions and labelling of the dimensions vary slightly, but there is considerable similarity between organisations. The Statistics Canada Quality Guidelines (Statistics Canada, 1998 – see also Brackstone, 1999) identify six key dimensions of quality:

- relevance
- accuracy

- timeliness
- accessibility
- interpretability
- coherence

With minor adjustments, these six dimensions have been adopted and employed by many other organisations including, in the UK, the Office for National Statistics (Holt and Jones, 1999).

Some organisations (e.g. Eurostat, 1999) have added a seventh dimension covering costs and burden, while others, like Stats Canada and Stats Sweden, prefer to view these as a separate concept entirely, against which survey quality should be assessed. Either way, costs (and concepts such as value for money) are clearly an important component of any assessment of the quality of surveys that are carried out with finite budgetary resources.

The framework that we propose here for UK academic longitudinal studies draws upon the Stats Canada conceptualisation and is therefore set out under the six headings listed above. Under each heading, we list, define and describe the major components of survey quality relevant to academic longitudinal studies. The framework is not therefore completely general, but is tailored to the specific demands and characteristics of academic longitudinal studies. Nevertheless, many components would apply to other surveys too. We also list the major components of cost, not because cost is an inherent component of quality but rather because one needs the ability to assess quality relative to cost.

We believe that a conceptual framework provides the crucial foundation necessary as the basis upon which to develop quality profiles and quality standards. Without this, profiles and standards run the risk of being atheoretical “shopping lists”. This framework is therefore designed to be the natural reference point for discussions of quality issues in longitudinal studies.

2. Relevance

This dimension of quality refers to the extent to which the data meet the needs of users. We would note first that "users" should be interpreted in the broadest sense, to include those who might use the data in the future, those who might use the data were they more relevant to them, and those affected by the use of the data. In the case of the ESRC studies, this is potentially the entire UK population, both present and future, given that policy decisions of various sorts could be influenced by analyses of the data. The key components of relevance might be summarised thus:

2.1 Were the right questions asked?

This is obviously difficult to define in the case of multi-purpose, multi-user studies. Appropriate user consultation, expert assessment and peer review are likely to be the most productive ways of assessment. The issue might conveniently be divided into two parts – an assessment of the topics or modules included in each wave, and an assessment of the individual items within each topic. Such assessments should encompass considerations such as consistency over waves in concepts, definitions and methods and consistency between related studies (e.g. across the different birth cohort studies).

One aspect of this is the concept of "fading relevance". Topics and questions that seemed highly relevant in the early stages of a study may become rather less relevant years later. Quality measures should encompass assessment of the extent to which an appropriate balance has been struck between retaining questions over waves (relevance might fade, but consistency might be high) and replacing questions with more relevant ones (higher relevance, lower consistency).

2.2 Were they asked of the right people/ units?

This is a question of study design and, particularly, sample design. It is important to recognise that surveys sample in time and space, so this component goes beyond simple definitions of sample units, reporting units and analysis units. It encompasses issues like the frequency/timing of survey waves as well as sample structure (birth cohort *vs* cross-sectional sample, etc).

2.3 Were they asked in the right way?

This essentially concerns the validity of the questions/ data collection methods. Opportunities to make direct empirical assessments of validity are rare, so this component of quality can often only be addressed indirectly, via assessment of the question testing methods used, studies of internal validity etc. There are many aspects of validity, but a key one for longitudinal studies is the balance between cross-sectional relevance (e.g. using the best current measure) and longitudinal consistency (using the same measure). This manifests itself particularly in decisions regarding the up-dating of classifications (of occupations, industries, ethnic groups, educational qualifications, etc).

3. Accuracy

This quality dimension refers to the (estimated) magnitude of the difference between survey estimates and corresponding population parameters. This can usefully be characterised by the concept of "total survey error" (Groves, 1989). For many years accuracy was commonly viewed as the key characteristic of a survey. Only recently has the concept of quality been broadened to encompass other dimensions. Many estimates will be made from the data arising from multi-purpose, multi-user surveys and these will differ in form (e.g. point estimates of levels, differences, changes; interval estimates; model coefficients, etc) as well as in substance. It is impossible to assess the accuracy of every such estimate. Rather, "typical" analyses or "key" estimates are often employed as the basis for assessment of accuracy.

The accuracy of an estimate is traditionally decomposed into two components – bias (the result of systematic error) and variance (the result of random error). There are many sources of bias and variance; indeed, virtually every stage of the survey process can introduce error of some sort. Accuracy can be assessed via separate estimation of the contribution from each potential source. Error sources can usefully be classified as *errors of non-observation* and *observational errors* (Groves, 1989). The former include errors due to frame coverage, sampling and non-response. The latter include errors due to instrument design, interviewers, respondents, and post-interview processing of survey data.

3.1 Coverage error

Incomplete coverage of a sampling frame or sampling method (*under*-coverage) will introduce bias if the omitted units differ systematically from the included units. (*Over*-coverage can also cause bias if ineligible units are inadvertently included in a survey sample, but this is rarely important as eligibility can usually be established at the data collection stage.) Coverage bias can only be directly assessed by an external evaluation of the sampling frame. For example, the two main sampling frames for UK general population surveys, the Electoral Registers and Postcode Address File, have been evaluated by matching to each other and/or Census records (Dodd 1987; Foster 1993, 1994; Lynn and Taylor, 1995). In the absence of a direct assessment, an informal assessment should be made by establishing the likely reasons and mechanisms for omission from the frame, estimating the extent of omissions, and using auxiliary information to make general inferences about the likely nature and direction of any consequent bias. In addition to an initial assessment of the sample design, in the context of long-running longitudinal studies, the nature of coverage error may need to be reassessed periodically as the definition of the target population changes.

3.2 Sampling error

When sample data is used to make population inferences, estimates may differ from the true population values due to the play of chance in the random sampling mechanism. This is the result of random sampling *variance*. The magnitude of sampling variance will differ between estimates and will be influenced by a number of aspects of sample design (primarily sample size, sample stratification, sample clustering and distribution of selection probabilities). There exists a well-developed theory of sampling variance and the variance of estimates can be estimated from the survey data (and sample design data) alone using standard software.

Sampling can also introduce *bias* if units selected with unequal selection probabilities are not appropriately weighted at the analysis stage. This can happen, for example, due to unidentifiable duplicates in the frame or due to the sampling method allowing some units more than one chance of selection. Rarely can such bias directly be quantified (if it could be, then the appropriate correction could be made and the bias removed) so assessment involves informal indirect methods.

It should be noted that longitudinal studies often aim to represent a number of different populations (e.g. multiple cross-sectional populations and multiple longitudinal populations). The nature of sampling error could be different for each of these populations and should therefore be assessed separately. Sampling error should also be assessed for any key population subgroups of analytic interest.

3.3 Non-response error

Survey non-response can introduce *bias* to survey estimates if non-responding units differ systematically from responding units. Non-response bias can be an important component of accuracy and many surveys rightly devote considerable resources to the assessment of non-response bias and the development of adjustment mechanisms. There is a range of methods available for the estimation of non-response bias. These include direct estimation in terms of auxiliary data available from the frame or from previous waves, by data linkage, or by interviewer observation; indirect estimation by comparing survey data with external data sources; and a range of modelling approaches, for example based on survey process information or theoretical models of survey participation.

These methods are often used not only to assess non-response bias but also to develop corrective weighting. Accuracy of estimation will therefore depend not only on the non-response bias at the data collection stage but also on the success of subsequent weighting in reducing that bias. An assessment of the impact of non-response error on accuracy of estimation should therefore aim to estimate the residual bias remaining after the corrective weighting (if any) has been applied.

As with many other sources of error, non-response bias can act differently on different subgroups of the sample, particularly subgroups defined by their participation pattern over a number of waves. The effect of non-response is cumulative and the relative importance of non-response bias therefore tends to be greater for surveys with larger numbers of waves or waves spread over a larger number of years.

In addition to output quality measures (e.g. direct estimates of bias), process quality measures can be pertinent to non-response error. These could include indicators of the steps taken to remain in contact with sample members between waves, updating contact information and so on.

Non-response will also tend to increase the *variance* of estimates as it reduces the available sample size. The magnitude of this affect can be easily estimated using standard sampling theory.

3.4 Instrument errors

The way that questionnaires and other data collection instruments are designed can affect the resultant data and hence the accuracy of those data. Numerous methodological studies have explored the nature of such instrument effects (e.g. Schuman and Presser 1996, Krosnick and Fabrigar 2001). The effects can either be systematic or random and thus result in either bias or variance. The most powerful way to quantify such effects is by experimental manipulation of the design. If this is not possible on the survey in question (which, typically, it will not be), assessment must be made by analogy to external experimental and theoretical evidence. Typically, instrument errors are specific to individual questionnaire items (or even sub-items). It is therefore necessary to assess each item separately. It is rarely possible to form an overall judgement about an instrument as a whole, except as a summary of a set of item judgements. Furthermore, effects can differ across estimates based on the same items. For example, two items A and B might each introduce systematic errors that would lead to a bias in estimates of mean levels of A or B. However, if the bias is constant across the two items, estimates of B/A, say, could remain unbiased. Instrument errors are therefore best considered with respect to specific estimates.

One particular source of instrument errors on longitudinal studies is inconsistency in wording, definitions and other aspects of data collection methods across waves. This can bias estimates of change over time. Another important issue relates to the use of dependent interviewing techniques for “updating” information collected at previous waves: the nature of errors may be different depending on whether or not dependent interviewing is employed and which technique is employed.

3.5 Respondent errors

There are a range of well-established reasons why survey respondents do not always provide data that are a true reflection of the information being sought. These include lack of knowledge (never knew the answer), memory errors (once knew the answer, but can no longer recall it accurately or at all), partial knowledge (knows part of the answer, but not the full detail), and deliberate mis-reporting (knows the answer but is unwilling to report it accurately or at all – for example due to social desirability effects or confidentiality concerns). These reasons for inaccurate reporting manifest themselves as various phenomena well known to survey researchers including under-reporting, “telescoping”, “heaping” or “bunching”, etc. These effects can, again, introduce bias and/or variance to survey estimates.

Internal analyses can often confirm the presence of respondent errors but can rarely confirm their absence. External validation can be used to quantify the effects of respondent errors but is only occasionally possible and is typically costly. Respondent reliability (an important component of respondent error, but not the only component) can be assessed by simple test-retest measures. Longitudinal studies sometimes provide natural opportunities for such dual measurement and/or for direct assessment of memory effects. Indirect assessment methods can include evaluation of the appropriateness of the recall periods about which respondents are asked. There is considerable evidence that optimal recall periods differ greatly between topics. The recall period is often related to the interval between waves, so the choice of interval can also affect respondent errors.

3.6 Interviewer errors

In the case of data collection by interview, interviewers can affect the survey data. This can be a result of conscious differences in the way questions or answers are interpreted or recorded, but it can also be a result of subtle differences in tone of voice, accent, speed at which the question is read and so on (Collins, 1980). Such effects will introduce bias if they are systematic across interviewers (for example, a tendency to fail to probe fully enough at a particular question will result in a net under-reporting). They will introduce variance even if interviewers can be assumed to vary around some "true" mean (O'Muircheataigh and Campanelli, 1999). The extent of interviewer variance can be estimated by a simple extension of sampling theory, but only for designs where interviewers are not confounded with sampling units. This requires random allocation to interviewers within (groups of) primary sampling units (as was done for a sub-sample of BHPS wave 2) or repeat measurement for a sub-sample (as was done on the 2001 English House Condition Survey). This therefore requires planning in advance of data collection and has an associated cost. Some longitudinal studies attempt estimation of interviewer errors periodically, but not frequently. In the absence of such direct assessment, general assessments can be made based upon information regarding the distribution of number of interviews per interviewer, supplemented with external estimates of correlated interviewer variance.

It should, of course, be noted that errors due to instruments, respondents and interviewers interact. This conceptual division is convenient as it reminds us that all three sources are important. However, in practice the exact cause of errors can not always be pinpointed. Quantitative estimates can often be made of respondent errors (through validation exercises or reliability measures) and interviewer errors (through empirical estimation based on sampling theory). However, this should not lead us to conclude that the "blame" lies somehow with the respondents or the interviewers.

3.7 Processing errors

After data collection (e.g. interviewing) is completed, a number of subsequent processes take place before data becomes available to analysts. Key processes include coding, data entry, data management and production of derived variables. Errors can be introduced by any of these processes.

Coding is an important source of *variance*. Coders, like interviewers, are not robotic. To perform their task well, they rely on subjective judgement. As they do not behave identically, variation is introduced (Kalton and Stowell, 1979). Even small differences in behaviour can have a large impact on the variance of estimates, due to the large number of cases typically dealt with by each coder (Bushnell, 2000). As with interviewer variance, coder variance can only be measured directly if a sub-sample of cases are either randomly allocated to coders or coded by multiple coders. However, unlike interviewing, repeat coding of a subsample is relatively inexpensive and coder errors are often assessed in this way. At the very least, an indirect assessment can be made via knowledge of the distribution of the number of cases coded per coder.

Other post-collection processes should not introduce error to survey estimates. If they do, it will be the result of mistakes such as keying errors, linking/merging errors etc. These can not be directly identified (if they were, they would be corrected) so quality assessment of this component is typically limited to *process* quality indicators (for example the use of double-keying, quality controls on linkage processes and so on).

3.8 Overall accuracy

To produce an estimate of overall accuracy for any particular survey estimate, each component must be separately estimated and the estimates combined. Typically, some components can be estimated well on a routine basis (e.g. sampling variance) others can be estimated well but infrequently (e.g. interviewer and coder variance) and others can only be estimated crudely (e.g. coverage errors). It is therefore important that any assessment of overall accuracy reports clearly and fully the methods used for estimation of the components.

4. Timeliness

The timeliness of survey outputs (data, documentation, reports) refers to the delay between the time reference point (or end of reference period) and the date when the outputs become available. This has two main components. One is an absolute measure (the length of the delay); the other is a relative measure (the actual delay relative to the delay that was planned/announced/anticipated by users. The latter is particularly important in situations where users may be scheduling resources or activities in anticipation of outputs becoming available. A distinction between primary and secondary users may also be relevant.

Timeliness can be assessed first, by relating actual delays in the release of outputs to the time-dependency of the uses to which the data are put and second, by evaluating the extent to which planned/announced release dates are met.

Other aspects of timing affect survey quality. These include the choice of time reference point or reference period and the timing of data collection exercises. However, these aspects are subsumed within the dimension of relevance (section 2 above).

5. Accessibility

This dimension of quality refers to the ease with which relevant outputs can be obtained. This encompasses the ease with which the existence of outputs can be established, the suitability of the form or medium through which they are accessed and any barriers to use of the outputs, such as costs or restrictions. These aspects can be assessed qualitatively and/or through user consultation exercises.

Accessibility will be influenced by a number of aspects of study management and organisation, including database management strategies, publicity, dissemination and outreach activities and user support facilities. Assessments of these aspects can serve as indirect indicators of accessibility.

6. Interpretability

This refers to the availability of supplementary information and metadata necessary to interpret and utilise outputs appropriately. This includes details of underlying concepts, variables and classifications, definitions of derived variables, contextual data, methodology of data collection and indications of statistical accuracy. These aspects can be assessed qualitatively through examination of documentation, publications and other available information and through user consultation exercises. The availability of a user support service can also aid interpretability.

Specific issues under this heading include: whether imputed values are identified as such in the data available to users; whether methods of imputation and weighting are fully documented; whether appropriate contextual information is provided, for example regarding policy and societal changes, how the timing of these relate to waves of the survey, and what is known about the impact of the policy changes on the survey measures.

7. Coherence

This dimension refers to the extent to which data/reports from different sources can be brought together within an interpretative framework and over time. In the context of longitudinal studies, there are two aspects to this. This first is coherence across waves of a study. The second is coherence between the study and other studies (both longitudinal and cross-sectional). The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology amongst surveys likely to be compared or otherwise combined in analysis and across waves of a longitudinal study. It should be noted that coherence is sometimes in conflict with relevance.

8. Costs

Important components of the cost of a survey include:

- Monetary cost to funders;
- Monetary cost to users;
- Opportunity costs;
- Costs to sample members in terms of time, burden and intrusiveness.

References

Brackstone, G. (1999) Managing data quality in a statistical agency, *Survey Methodology* 25:2, 139-149.

Bushnell, D. (2000) The impact of coding on data quality, *Survey Methods Newsletter* 20:1, 16-19.

Collins, M. (1980) Interviewer variability: a review of the problem, *Journal of the Market Research Society* 22:2, 77-95.

Dodd, T. (1987) A further investigation into the coverage of the Postcode Address File, *Survey Methodology Bulletin* 21, 35-40

Eurostat (1999) *Assessment of the Quality in Statistics: Standard Quality Report*. Doc. Eurostat/A4/Quality/00/General/Standard Report. Luxembourg: Eurostat.

Foster, K. (1993) The electoral register as a sampling frame, *Survey Methodology Bulletin* 33, 1-7

Foster, K. (1994) The coverage of the Postcode Address File as a sampling frame, *Survey Methodology Bulletin* 34, 9-18

Groves, R.M. (1989) *Survey Errors and Survey Costs*. New York: John Wiley.

Holt, T and Jones, T. (1999) Quality work and conflicting quality objectives, in *Quality Work and Quality Assurance within Statistics* (pp. 15-24). Luxembourg: EC/Eurostat.

Kalton, G. and Stowell, R. (1979) A study of coder variability, *Journal of the Royal Statistical Society Series C (Applied Statistics*), 28:3

Krosnick, J. and Fabrigar, (2001) Designing Great Questionnaires: Insights from Psychology. Oxford: Oxford University Press.

Lynn, P. and Taylor, B. (1995) On the bias and variance of samples of individuals: a comparison of the electoral registers and postcode address file as sampling frames, *Journal of the Royal Statistical Society Series D (The Statistician*), 44:2, 173-194

O'Muircheataigh, C. and Campanelli, P. (1999) A multilevel investigation of the role of interviewers in survey non-response, *Journal of the Royal Statistical Society Series A (Statistics in Society*), 162:3, 437-446

Schuman, H. and Presser, S. (1996) Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. 2nd edition. Thousand Oaks, CA: Sage

Statistics Canada (1998) *Statistical Quality Guidelines (3rd edition)*. Ottawa: Statistics Canada.