Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2014)
177, *Part* 1, *pp.* 3–38

On information quality

Ron S. Kenett

KPA, Raanana, Israel, University of Turin, Italy, and New York University–Poly, USA

and Galit Shmueli

Indian School of Business, Gachibowli, India

[Received August 2009. Final revision August 2012]

Summary. We define the concept of *information quality 'InfoQ'* as the potential of a data set to achieve a specific (scientific or practical) goal by using a given empirical analysis method. InfoQ is different from data quality and analysis quality, but is dependent on these components and on the relationship between them. We survey statistical methods for increasing InfoQ at the study design and post-data-collection stages, and we consider them relatively to what we define as InfoQ. We propose eight dimensions that help to assess InfoQ: *data resolution*, *data structure*, *data integration*, *temporal relevance*, *generalizability*, *chronology of data and goal*, *construct operationalization* and *communication*. We demonstrate the concept of InfoQ, its components (what it is) and assessment (how it is achieved) through three case-studies in on-line auctions research. We suggest that formalizing the concept of InfoQ can help to increase the value of statistical analysis, and data mining both methodologically and practically, thus contributing to a general theory of applied statistics.

Keywords: Data; Data analytics; Data mining, Statistical modelling; Study design; Study goal

1. Introduction

Statistics and data mining are disciplines that are focused on extracting knowledge from data. They provide a toolkit for testing hypotheses of interest, predicting new observations, quantifying population effects and summarizing data efficiently. In these fields, measurable data are used to derive knowledge. However, a clean, exact and complete data set, which is analysed professionally, might contain no useful information for the problem under investigation. Hand (2008) noted that

> 'statisticians working in a research environment . . . may well have to explain that the data are inadequate to answer a particular question'.

In contrast, a very 'dirty' data set, with missing values and incomplete coverage, can contain useful information for some goals. In some cases, clean data can even be misleading. This paper focuses on assessing the utility of a particular data set for achieving a given analysis goal by employing statistical analysis or data mining. We call this concept *information quality*, InfoQ. We propose a formal definition of InfoQ and provide guidelines for its assessment. Our objective is to offer a general framework that applies to empirical research and the practice of statistical analysis. Our discussion considers statistical and other approaches for maximizing InfoQ at the study design stage and at the post-data-collection stage.

Address for correspondence: Galit Shmueli, Indian School of Business, Gachibowli, Hyderabad 500 032, India.
E-mail: galit_shmueli@isb.edu

 0964–1998/14/177003

Increasing and assessing InfoQ require looking at the data, data analysis and goal, as well as at the relationships between them. To formalize the concept, we use the following notation and definitions: g, a specific analysis goal; X, the available data set; f, an empirical analysis method; U, a utility measure.

We use subscript indices to indicate alternatives. For example, to convey K different goals of analysis we use $g_1, g_2, \ldots, g_K$; J different methods of analysis are denoted $f_1, f_2, \ldots, f_J$.

The approach that we propose is 'top down', in the sense that first the goals are mapped, the data are obtained, the potential analysis method and performance measures are set, and then the utility of the data is evaluated within this context. We survey existing approaches and integrate them into the more general concept of InfoQ.

The paper proceeds as follows: Section 1.1 describes each of the four components of InfoQ and Section 1.2 provides a formal definition of the concept of InfoQ. In Section 2, we introduce three examples from on-line auction research, which are used throughout the paper to illustrate the different concepts and ideas. Section 3 describes related concepts such as *data quality* and *analysis quality*, and integrates them into the more general framework of InfoQ. Section 4 surveys statistical methods for increasing InfoQ at the study design stage, and Section 5 surveys the same at the post-data-collection stage. Section 6 proposes eight dimensions that characterize InfoQ and guidelines for assessing them. We conclude in Section 7 with a discussion and directions for further research.

1.1. Description of information quality components

1.1.1. Analysis goal g

Data analysis is used for a variety of purposes. Three general classes of goals are causal explanation, prediction and description (see Shmueli (2010) and Shmueli and Koppius (2011)). Causal explanation includes questions such as 'Which factors cause the outcome?' and 'What factor settings optimize the outcome?'. Predictive goals include forecasting future values of a series and predicting the output value for new observations given a set of input variables. Descriptive goals include quantifying and testing for population effects by using data summaries, graphical visualizations, statistical models and statistical tests. Deming (1953) introduced the distinction between *enumerative studies*, which are aimed at answering the question 'how many?' and *analytic studies*, which are aimed at answering the question 'why?'. Later, Tukey (1977) proposed a classification of exploratory and confirmatory data analysis. Our use of the term 'goal' generalizes all of these different types of goals and goal classifications. For examples of such goals in the context of customer satisfaction surveys see Kenett and Salini (2011).

1.1.2. Data X

The term 'data' includes any type of data to which empirical analysis can be applied. Data can arise from different collection tools: surveys, laboratory tests, field and computer experiments, simulations, Web searches, observational studies, social networks and more. Data can be univariate or multivariate and of any size (from a single observation in case-studies to 'big data' with zettabytes). They can also contain semantic unstructured information in the form of text or images with or without a time dimension.

1.1.3. Empirical analysis method f

We use *data analytics* and *empirical analysis* in a broad sense. These include statistical models and methods (parametric, semiparametric and non-parametric), data mining algorithms,

graphical methods and operations research methods (e.g. simplex optimization) where problems are modelled and parameterized.

1.1.4. Utility U

The extent to which the analysis goal is achieved is typically measured by some performance measure. We call this measure 'utility'. In predictive studies popular utility measures are predictive accuracy and lift; in descriptive studies, goodness-of-fit measures are common, whereas in explanatory modelling statistical power and strength-of-fit measures are most common.

1.2. Definition of information quality

Following Hand's (2008) definition of statistics as 'The technology of extracting meaning from data', we consider the utility of applying a technology f to a resource X for a given purpose g. In particular, we focus on the question 'What is the potential of a particular data set to achieve a particular goal using a given empirical analysis method?'. To formalize this question, we define the concept of information quality as

$$\text{InfoQ}(f, X, g) = U\{f(X|g)\}.$$

InfoQ is determined by the quality of its components g ('quality of goal definition'), X ('data quality'), f ('analysis quality') and U ('quality of utility measure') and by the relationships between them.

2. Examples from on-line auctions research

To illustrate the various definitions, concepts and arguments that are introduced in this paper, we use examples from the field of on-line auctions research, which have become a major electronic marketplace. Some of the large on-line auction Web sites, such as eBay, provide data on closed and on-going auctions, triggering a growing body of research in academia and in practice. A few popular analysis goals have been

(a) determining factors affecting the final price of an auction (e.g. Lucking-Reiley *et al.* (2007)),
(b) predicting the final price of an auction (e.g. Ghani and Simmons (2004) and Wang *et al.* (2008)),
(c) descriptive characterization of bidding strategies (e.g. Bapna *et al.* (2004) and Roth and Ockenfels (2002)),
(d) comparing behavioural characteristics of auction winners *versus* fixed price buyers (e.g. Angst *et al.* (2008)) and
(e) building descriptive statistical models of bid arrivals or bidder arrivals (e.g. Borle *et al.* (2006) and Shmueli *et al.* (2007)).

We next illustrate the InfoQ-components g, X, f and U with three case-studies.

2.1. Case 1: effect of a reserve price on the final auction price

Econometricians are interested in determining factors that affect the final price of an on-line auction. Although game theory provides an underlying theoretical causal model of price in off-line auctions, the on-line environment differs in substantial ways. We consider a study by Katkar and Reiley (2006) who investigated the effect of two types of reserve prices on the final

auction price. A reserve price is a value that is set by the seller at the start of the auction. If the final price does not exceed the reserve price, the auction does not transact. On eBay, sellers can choose to place a public reserve price that is visible to bidders, or an invisible secret reserve price (bidders see only that there is a reserve price but do not know its value). InfoQ, in the context of this study, consists of asking the question 'Given the data collected on a set of auctions, what is their potential to allow quantifying the difference between secret and public reserve prices by using regression modelling?'. Next, we examine each InfoQ-component for this study.

2.1.1. Study goal g

Quantify the effect of using a secret *versus* public reserve price on the final price of an auction.

2.1.2. Data X

Katkar and Reiley (2006) conducted a 'field experiment' by selling Pokémon cards on eBay. They auctioned 25 identical pairs of Pokémon cards in week-long auctions during a 2-week period in April 2000, where each card was auctioned twice: once with a public reserve price and once with a secret reserve price. The resulting data included the complete information on all 50 auctions.

2.1.3. Empirical analysis f

Katkar and Reiley (2006) used linear regression to test for the effect of a private or public reserve price on the final auction price and to quantify it.

2.1.4. Utility U

Katkar and Reiley (2006) used statistical significance (the p-value) of the regression coefficient to assess the presence of an effect for a private or public reserve price. They used the regression coefficient value for quantifying the magnitude of the effect (they concluded that 'a secret-reserve auction will generate a price $0.63 lower, on average, than will a public-reserve auction').

2.2. Case 2: forecasting the final price of an on-going auction

On any given day, thousands of auctions take place on line. Forecasting the price of on-going auctions is beneficial to buyers, sellers, auction houses and third parties. For potential bidders, price forecasts can be used for deciding whether, when and how much to bid. For sellers, price forecasts can help to decide whether and when to post another item for sale. For auction houses and third parties, services such as seller insurance can be offered with adjustable rates. The literature on auction price forecasting is mainly based on 'static' models which use only information that is available at the start of the auction. More recent work has integrated dynamic information that changes during the auction. Wang *et al.* (2008) developed a dynamic forecasting model that uses information that is available at the time of prediction. Their model has been used for a variety of products (electronics, contemporary art, etc.) and across different auction Web sites (see Jank and Shmueli (2010), chapter 4). In what follows, we briefly describe the study of Wang *et al.* (2008) in terms of the InfoQ-components.

2.2.1. Study goal g

Create a predictive model for accurately forecasting the price of an on-going auction, given the information on the auctioned item, the seller, the auction parameters, the bids from the auction start to the time of prediction and information on similar auctions that recently transacted.

2.2.2. Data X

Wang *et al.* (2008) used a set of 190 closed 7-day auctions of *Microsoft Xbox* gaming systems and *Harry Potter and the Half-blood Prince* books sold on eBay.com in August–September 2005. For each auction, the data included the bid history (bid amounts, time stamps and bidder identification) and information on the product characteristics, the auction parameters (e.g. the day of the week that the auction started) and bidder and seller information.

2.2.3. Empirical analysis f

The forecasting model is based on representing the sequences of bids from each auction by a smooth curve (using functional data analysis). Then, a regression model for the price at time t includes four types of predictors:

(a) static predictors (such as product characteristics),
(b) time varying predictors (such as the number of bids by time t),
(c) price dynamics (estimated from the price curves) and
(d) price lags.

Their model for the price at time t is given by

$$y(t) = \alpha + \sum_{i=1}^{Q} \beta_i x_i(t) + \sum_{j=1}^{J} \beta_j \mathrm{D}^j y(t) + \sum_{l=1}^{L} \eta_L y(t-l),$$

where $x_1(t), \ldots, x_Q(t)$ is the set of static and time varying predictors, $\mathrm{D}^{(j)} y(t)$ denotes the jth derivative of price at time t and $y(t-l)$ is the lth price lag. The h-step-ahead forecast, given information up to time T, is given by

$$\bar{y}(T+h|T) = \hat{\alpha} + \sum_{i=1}^{Q} \hat{\beta}_i x_i(T+h|T) + \sum_{j=1}^{J} \hat{\gamma}_j \bar{\mathrm{D}}^{(j)} y(T+h|T) + \sum_{l=1}^{L} \hat{\eta}_L \bar{y}(T+h-1|T).$$

2.2.4. Utility U

Katkar and Reiley (2006) measured forecast accuracy on a hold-out set of auctions by using the mean absolute percentage error, computed in two different ways: one comparing the true and forecasted price curves generated by functional data analysis (method MAPE_1), and another comparing the forecasted curves and the actual current auction prices (method MAPE_2).

2.3. Case 3: quantifying consumer surplus in eBay auctions

Classical microeconomic theory uses the notion of *consumer surplus* as the welfare measure that quantifies benefits to a consumer from an exchange. Marshall (1920), page 124, defined consumer surplus as

> 'the excess of the price which he (a consumer) would be willing to pay rather than go without the thing, over that which he actually does pay ...'.

Despite the growing research interest in on-line auctions, little is known about quantifiable consumer surplus levels in such mechanisms. On eBay, the winner is the highest bidder, and she or he pays the second highest bid. Whereas bid histories are publicly available, eBay never reveals the highest bid. Bapna *et al.* (2008a) set out to quantify consumer surplus on eBay by using a unique data set which revealed the highest bids for a sample of almost 5000 auctions. They found that, under a certain assumption, 'eBay's auctions generated at least $7.05 billion in total consumer surplus in 2003'.

2.3.1. Study goal g

Estimate the consumer surplus generated in eBay in 2003.

2.3.2. Data X

Since eBay does not disclose the highest bid in an auction, Katkar and Reiley (2006) used a large data set from Cniper.com, which is a Web-based tool that is used by many eBay users for placing a 'last minute bid'. Placing a bid very close to the auction close ('sniping') is a tactic for winning an auction by avoiding the placement of higher bids by competing bidders. The Cniper data set contained the highest bid for all the winners. The authors then merged the Cniper information with the eBay data for those auctions and obtained a data set of 4514 auctions that took place between January and April 2003. Their data set was also unique in that it contained information on auctions in three different currencies and across all eBay product categories.

2.3.3. Empirical analysis f

Katkar and Reiley (2006) computed the median surplus by using the sample median with a 95% bootstrap confidence interval. They examined various subsets of the data and used regression analysis to correct for possible biases and to evaluate robustness to various assumption violations. For example, they compared their sample with a random sample from eBay in terms of the various variables, to evaluate whether Cniper winners were savvier and hence derived a higher surplus.

2.3.4. Utility U

The precision of the estimated surplus value was measured via a confidence interval. The bias due to non-representative sampling was quantified by calculating an upper bound.

3. Quality of components of information quality

3.1. Goal definition

Translating a scientific or practical goal into a statistical goal is challenging and requires close collaboration between the data analyst and the domain expert. Defining the statistical goal inappropriately negatively affects InfoQ. A well-defined statistical goal is one that properly reflects the scientific or practical goal. Although a data set can be useful for scientific goal g_1 it can be completely useless for scientific goal g_2. Kimball (1957) coined the term 'error of the third kind' to denote 'giving the right answer to the wrong question'. Tukey (1962) commented

> 'Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise'.

3.2. 'Data quality' and U(X|g)

The same data can contain high quality information for one purpose and low quality information for another. This has been recognized and addressed in several fields. Mallows posed the 'zeroth problem' (Mallows, 1998) asking 'how do the data relate to the problem, and what other data might be relevant?'. In database engineering and management information systems, the term 'data quality' refers to the usefulness of queried data to the person querying it. Wang *et al.* (1993) gave the following example:

> 'Acceptable levels of data quality may differ from one user to another. An investor loosely following a

stock may consider a ten minute delay for share price sufficiently timely, whereas a trader who needs price quotes in real time may not consider ten minutes timely enough.'

These issues focus on $U(X|g)$, which differs from InfoQ by excluding the empirical analysis component. Several official agencies have also focused on $U(X|g)$. In the UK, for instance, the Department of Health uses a management information systems type of definition of data quality with respect to the quality of medical and healthcare patient data in the National Health Service (UK Department of Health, 2004).

A similar concept is the *quality of statistical data* which has been developed and used in European official statistics and international organizations such as the International Monetary Fund and the Organisation for Economic Co-operation and Development. This concept refers to the usefulness of summary statistics that are produced by national statistics agencies and other producers of official statistics. In other words, f is equivalent to computing summary statistics (although this operation might seem very simple, it is nonetheless considered 'analysis', because it is in fact estimation). Quality is evaluated in terms of the usefulness of the statistics for a particular goal. The Organisation for Economic Co-operation and Development uses seven dimensions for quality assessment: *relevance*, *accuracy*, *timeliness and punctuality*, *accessibility*, *interpretability*, *coherence* and *credibility* (see Giovanni (2008), chapter 5). In the context of survey quality, official agencies such as Eurostat, the National Center for Science and Engineering Statistics and Statistics Canada have created quality dimensions for evaluating the quality of a survey for the goal g of obtaining 'accurate survey data' as measured by U equivalent to the mean-squared error MSE (see Biemer and Lyberg (2003)). For example, Eurostat's quality dimensions are *relevance of statistical concept*, *accuracy of estimates*, *timeliness and punctuality in disseminating results*, *accessibility and clarity of the information*, *comparability*, *coherence* and *completeness* (for the National Center for Science and Engineering Statistics guidelines and standards see `www.nsf.gov/statistics`). InfoQ builds on these definitions and provides an operational approach for assessing them.

3.3. Analysis f

Analysis quality refers to the adequacy of the empirical analysis in light of the data and goal at hand. Analysis quality reflects the adequacy of the modelling with respect to the data and for answering the question of interest. Godfrey (2008) described low analysis quality as 'poor models and poor analysis techniques, or even analyzing the data in a totally incorrect way'. Statistics education is aimed at teaching high quality analysis. Techniques for checking analysis quality include residual analysis and cross-validation. Analysis quality depends on the expertise of the analyst and on the empirical methods and software that are available at the time of analysis.

3.4. Utility U

Performance measures must depend on the goal at hand, on the nature of the data and on the analysis method. For example, a common error in various fields is using R^2 for measuring predictive accuracy (see Shmueli and Koppius (2011)). Another error is relying solely on p-values for testing hypotheses with very large samples, which is a common practice in several fields that now use hundreds of thousands or even millions of observations (see Lin *et al.* (2009)).

4. Methods for increasing information quality at the study design stage

It is useful to distinguish between causes affecting data quality *a priori* and *a posteriori*. *A priori* causes are known during the study design stage and before data collection. They result, for

Table 1. Statistical strategies for increasing InfoQ given *a priori* causes at the design stage

	Strategies for increasing InfoQ	*A priori causes*
Design of experiments	Randomization; blocking; replication; linking data collection protocol with appropriate design	Resource constraints; impossible runs
Clinical trials	Randomization; blocking; replication; linking data collection protocol with appropriate design; blinding; placebo	Resource constraints; ethics; safety
Survey sampling	Reducing non-sampling errors (e.g. pretesting questionnaire, reducing non-response) and sampling errors (e.g. randomization, stratification, identifying target and sampled populations)	Resource constraints; ethics; safety
Computer experiments	Randomization; blocking; replication; linking data collection protocol with appropriate design; space filling designs	Impossible or difficult to obtain real data; time and costs associated with computer simulation

example, from known limitations of resources (e.g. the sample size), ethical, legal and safety considerations (e.g. an inability to test a certain drug on certain people in a clinical trial). *A posteriori* issues (the focus of Section 5) result from faults in the mechanism generating or collecting the data and are discovered (or not) after the data have been collected (e.g. data entry errors, measurement error and intentional data manipulation). Consider a measured data set X and a true data set $X^* \neq X$. We denote data that are affected by *a priori* causes by $X = \eta_1(X^*)$, by *a posteriori* causes by $X = \eta_2(X^*)$ and by both causes by $X = \eta_2\{\eta_1(X^*)\}$.

In this section we describe existing approaches for increasing InfoQ under different scenarios of *a priori* data issues and related InfoQ-decreasing constraints. Table 1 summarizes the strategies and constraints. The next sections expand on each point.

4.1. Statistical design of experiments

Experiments are considered the gold standard for inferring causality, yet experimentation can be resource intensive. The aim of the design of experiments (DOE) is to collect data in the most efficient way for answering a causal question of interest.

Objectives g of DOE are typically classified into *screening* (identifying main factors affecting a response), *comparing* (testing the effect of a single factor on a response, often in the presence of other nuisance factors), *optimizing* (finding a subset of factor space that optimizes the response) and *robustifying* (reducing the sensitivity of a response to noise variables, in the subset of factor space identified as optimal). Given one of these objectives and typical resource constraints, experimental factor level combinations are selected and an experimental array is chosen. These considerations assume that $X = \eta_1(X^*)$. Then, assuming adequate statistical analysis, the data that are generated by the experiment are expected to yield high InfoQ in terms of answering the question of interest at the required level of type I and II errors (U).

Note that data collected through a design for achieving g_1 should lead to high InfoQ relative to that objective, but they might have low InfoQ for $g_i \neq g_1$. For instance, data from a study that was designed to screen several dozen factors might be of low InfoQ for *comparing* in terms of the chosen factors and their levels.

In the on-line auction experimental study (Section 2.1), the researchers' choice of item to auction (Pokémon cards), the experimental design (selling 25 identical pairs of Pokémon cards,

each card auctioned twice: once with a public and once with a secret reserve) and the experimental setting (for example, all auctions were 7-day auctions and started on Sunday between 7 and 9 p.m. Eastern Standard Time, and the seller rating was kept at 0) were directly aimed at achieving high InfoQ for answering the particular *comparative* question. In contrast, the same data would be of low InfoQ for a screening study to determine the main factors affecting the final price, because many potential factors such as the duration of the auction, start and end day of the week and seller rating were purposely held constant.

The three main principles of experimental design, *randomization*, *blocking* and *replication*, assume that the data collection is under the control of the experimenter. However, once the data have been collected all that can be done is to evaluate whether and how these principles have been achieved by auditing the protocol of the experiment execution.

4.2. Clinical trials and experiments with human subjects

A clinical trial is defined as 'a prospective study comparing the effect and value of intervention(s) against a control in human beings' (Friedman *et al.* (1999), page 2). In clinical trials, key strategies are *randomization* and *double blinding* (where both doctors and patients do not have knowledge about the treatment assignment; sometimes triple blinding is applied, where even the data analyst is unaware of the meaning of the labels of various analysed groups). The goal is to generate unbiased data that can then be used to evaluate or compare the intervention effects. Although the design of clinical trials and of experiments involving humans has its roots in classic DOE, the human factor (e.g. patients and doctors) introduces two important differentiating factors: ethics and safety, which further limit the level of InfoQ of the generated data $X = \eta_1(X^*)$. Ethical considerations constrain the experimental design and can lower InfoQ. For instance, some treatment combinations might not be ethical, some sequences of runs might not be ethical or even the lack of treatment can be unethical. The early termination of a trial through futility analyses is a reasonable strategy in some cases. Clinical trials typically require a control group, yet not providing treatment or providing a placebo might not be ethical in some cases. Moreover, the strategy of *randomization* is ethically debated (Friedman *et al.* (1999), page 45). InfoQ is also potentially decreased by the human factor of compliance of patients and physicians with the treatment regime mandated by the experimental design. This can limit the types of answerable research questions. In extreme cases on non-compliance,

> 'an investigator [interested in comparing interventions] may not be able to compare interventions, but only intervention strategies'

(Friedman *et al.* (1999), page 3). When the purpose of the trial is comparing interventions, human responses can be affected not only by the treatment but also by psychological and other effects. Hence, *placebos* and *blinding* are widely employed.

Two further factors that decrease InfoQ for comparing interventions are safety issues and the need for informed consent. Safety considerations can affect InfoQ by impacting the study design, for instance by restricting the dosage to relatively low levels. The need to obtain informed consent from participants in clinical trials creates a constraint on the ability to draw 'objective' results due to impacting psychological effects and of levels of compliance.

4.3. Design of observational studies: survey sampling

The statistical literature includes methodology for designing observational studies including sample surveys. Sampling methodology aims to achieve high precision and low bias estimates, within resource constraints. Designing a survey study consists of determining sampling issues

such as sample size, sampling scheme and sampling allocation, as well as addressing other issues such as non-response and questionnaire design for reducing non-sampling errors (measurement bias and selection bias). InfoQ is influenced by both sampling and non-sampling errors, relative to the goal at hand. *Individual surveys*, where the opinions of respondents are sought, differ from *households and establishment surveys*, where respondents are asked to provide an aggregate representative response. An individual survey with high InfoQ can be of low InfoQ when the opinion sought is that of a company rather than a personal opinion. Another example is surveys measuring the rate of unemployment (such as the Current Population Survey in the USA) but are not designed to produce statistics about the number of jobs held.

The first step in designing a survey study is generating a clear statistical statement about the information desired as well as a clear definition of the target population. These two factors, which determine g, affect not only the data analysis, but also the data collection instrument. As in DOE and clinical trials, the study design must take into account resource constraints and ethical, legal and safety considerations. In all cases, when human subjects are involved as experimental units or surveyed participants, a special approval process is usually required to carry out a study. Institutional review boards are in charge of approving, monitoring and reviewing studies that involve human subjects. These review boards' mission is to protect participants' safety and wellbeing as well as to validate the goals of the proposed study to assure that the study design will achieve sufficient InfoQ. As in clinical trials, InfoQ is potentially limited by such constraints.

Related to the on-line auction examples that were presented in Section 2, consider a survey study aimed at comparing behavioural traits of auction winners who placed a bid *versus* those who paid the 'buy-it-now' price (which is an option in many on-line auctions that allows purchasing an item at a fixed price before the bidding begins). Angst *et al.* (2008) surveyed winners in eBay auctions to test whether 'competitiveness, impulsiveness, and level of hedonistic need' distinguish bidders from fixed price buyers. To obtain data with high InfoQ they tried to reduce non-sampling errors (e.g. by using previously validated scales in the questionnaire, and sending multiple follow-ups) and sampling errors (by choosing a sample of auctions for a popular product during a limited time period). Several limitations reduced InfoQ in this study: non-sampling issues include a response rate of 27% (113 usable questionnaires) and a change in eBay's policy during the survey period that led to a shift from Web surveys to e-mail surveys (thereby introducing a 'survey type' effect). Sampling issues relate to the generalizability from the sample to the larger population, given the small sample size and that a single product was chosen. For more on customer surveys see Kenett and Salini (2012).

4.4. Computer experiments

In computer experiments, a computer runs a computationally intensive stochastic or deterministic model that simulates a scientific phenomenon (e.g. a computational fluid dynamics model). DOE is then used to collect response values for a set of input values for building a statistical model ('metamodel') of the response in terms of the input variables. Statistical methods such as 'space filling designs' are used to generate data with high InfoQ to help to design robust systems and products. Computer experiments range from very basic simulation to complex dynamic systems. This variety is reflected by the wide range of levels of proposed 'fidelity' and Bayesian-based inference (Huang and Allen, 2005). Validating and calibrating computer simulations are non-trivial tasks, and one major and not uncommon risk is using the wrong simulation model (see Bayarri *et al.* (2007)). A wrong model will obviously lead to low InfoQ. For more on computer experiments see Bates *et al.* (2006).

Table 2. Strategies for increasing InfoQ given *a posteriori* causes at the post-data-collection stage

	Strategies for increasing InfoQ	*A posteriori causes*
Data cleaning and preprocessing	Imputation, other handling of missing values; advanced technologies for data collection, transfer, storage; detecting and handling outliers and influential observations	Data entry errors, measurement error and intentional data manipulation
Reweighting and bias adjustment	Reweighting observations in explanatory studies	Selection bias
Meta-analysis	Combine results of studies	'File drawer' bias; agenda-driven bias; Simpson's paradox

5. Methods for increasing information quality in the post-data-collection stage

This section describes existing approaches designed to increase InfoQ at the post-data-collection stage, under various scenarios of $X = \eta_2\{\eta_1(X^*)\}$. Table 2 summarizes the main points.

5.1. Data cleaning and preprocessing

Data 'cleanliness' has long been recognized by statisticians as a serious challenge. Hand (2008) commented that 'it is rare to meet a data set which does not have quality problems of some kind'. Godfrey (2008) noted that

> 'Data quality is a critically important subject. Unfortunately, it is one of the least understood subjects in quality management and, far too often, is simply ignored.'

Consider a measured data set X and a true data set $X^* \neq X$. The data quality literature includes methods for 'cleaning' X to achieve X^* and guidelines for data collection, transfer and storage that reduce the distance between X and X^*. Denote data quality procedures (cleaning, avoiding errors, etc.) by $h(\cdot)$. We distinguish between two general types of procedures $h(X)$ and $h(X|g)$. $h(X)$ focuses on procedures that generate or clean X to minimize its distance from X^* without considering anything except the data set itself. Advanced data recording devices (such as scanners and radio-frequency identification readers), data validation methods, data transfer and verification technologies and robust data storage, as well as more advanced measurement instruments, have produced 'cleaner' data (Redman, 2007) in terms of the distance between X and X^*. Management information systems type data quality (see Section 3.1) focuses on $h(X)$ operations.

In contrast, $h(X|g)$ focuses on quality procedures that generate or clean X conditional on the goal g. One example is classic statistical data imputation (Little and Rubin, 2002), where the type of imputation is based on the assumed missing data generation mechanism, and conditional on the purpose of minimizing bias (which is important in explanatory and descriptive studies). Another example is a method for handling missing predictor values in studies with a predictive goal, by Saar-Tsechansky and Provost (2007). Their approach builds multiple predictive models using different subsets of the predictors, and then uses for each new observation the model that excludes predictors that are missing for that observation. A third example is a data acquisition algorithm that was developed by Saar-Tsechansky *et al.* (2009) for data with missing response Y labels. The algorithm chooses the predictor values or missing response labels to collect, taking into consideration a predictive goal (by considering the cost and contribution to predictive accuracy).

Another $h(X|g)$ 'data cleaning' strategy is the detection and handling of outliers and influential observations. The choice between removing such observations, including them in the analysis or otherwise modifying them is goal dependent.

Does data cleaning always increase InfoQ? For $X \neq X^*$ we expect $\text{InfoQ}(f, X, g) \neq \text{InfoQ}(f, X^*, g)$. In most cases, data quality issues degrade the ability to extract knowledge ($\text{InfoQ}(f, X, g) < \text{InfoQ}(f, X^*, g)$). Missing values and incorrect values often add noise to our limited sample signal. Yet, sometimes X^* is just as informative or even more informative than X when conditioning on the goal and, hence, choosing $h(X) = X$ is optimal. For example, when the goal is to predict the outcome of new observations given a set of predictors, missing predictor values can be a blessing if they are sufficiently informative of the outcome (Ding and Simonoff, 2010). An example is missing data in financial statements which can be useful for predicting fraudulent reporting.

5.2. Reweighting and bias adjustment

A common approach that is aimed at correcting data for selection bias, and especially in survey data, is reweighting or adjustment. Using weights is aimed at reducing bias at the expense of increased variance, in an effort to maximize MSE. In other words, $h(X)$ is chosen to maximize $U[f\{h(X|g)\}] = \text{MSE}$. Yet, there is disagreement between survey statisticians regarding the usefulness of reweighting data, because 'weighted estimators can do very badly, particularly in small samples' (Little, 2009). When the analysis goal is estimating a population parameter, and f is equivalent to estimation, adjusting for estimator bias is common. For the propensity score approach, see Mealli *et al.* (2011).

In the eBay consumer surplus example (Section 2.3), Katkar and Reiley (2006) proposed a bias-corrected estimator of consumer surplus in common value auctions (where the auctioned item has the same value to all bidders), which is based on the highest bid.

5.3. Meta-analysis

In meta-analysis 'data' refer to statistical results of a set of previous studies studying the same research question. Statistical methodology is then used to combine these results for obtaining more precise and reliable results, i.e. for increasing InfoQ. *A posteriori* causes that decrease InfoQ include 'file drawer' bias, where studies that do not find effects remain unpublished and do not become factored into the meta-analysis, agenda-driven bias, where the researcher intentionally chooses a non-representative set of studies to include in the analysis, and unawareness of Simpson's paradox, which arises due to the aggregation of studies.

6. Assessing information quality

6.1. Assessing quality: the case of data quality

In marketing research and in the medical literature, data quality is assessed by defining the criteria of *recency*, *accuracy*, *availability* and *relevance* (Boslaugh, 2007; Patzer, 1995). The first three are characteristics of the data and relate indirectly to the analysis goal. Recency refers to the time difference between the study of interest and the data collection. Accuracy refers to the data quality. Availability describes the information in the data that are made available to the analyst. Relevance refers to the relevance of the data to the analysis goal: whether the data contain the required variables in the right form, and whether they are from the population of interest. These four criteria, although related to InfoQ, consider X and g, but exclude f and U.

In the management information systems literature, quantitative data quality criteria are often embedded in the database directly. Wang *et al.* (1993) proposed incorporating data quality characteristics in the database itself (for example, the date of cells gives recency information; the source of cells gives credibility information).

As already mentioned in Section 3.2, the European Commission's Eurostat agency uses seven dimensions of survey quality: relevance of statistical concept, accuracy of estimates, timeliness and punctuality in disseminating results, accessibility and clarity of the information, comparability, coherence and completeness.

6.2. Eight dimensions of information quality

Taking an approach that is similar to data quality assessment, we define eight dimensions for assessing InfoQ that consider and affect not only X and g, but also f and U. With this approach we provide a decomposition of InfoQ that can be used for assessing and improving research initiatives or *ex post* evaluations.

6.2.1. Data resolution

Data resolution refers to the *measurement scale* and *aggregation level* of X. The measurement scale of the data should be carefully evaluated in terms of its suitability to the goal, the analysis methods to be used and the required resolution of U. Given the original recorded scale, the researcher should evaluate its adequacy. It is usually easy to produce a more aggregated scale (e.g. two income categories instead of 10), but not a finer scale. Data might be recorded by multiple instruments or by multiple sources. To choose between the multiple measurements, supplemental information about the reliability and precision of the measuring devices or sources of data is useful. A finer measurement scale is often associated with more noise; hence the choice of scale can affect the empirical analysis directly.

The data *aggregation level* must also be evaluated relative to g. For example, consider daily purchases of over-the-counter medications at a large pharmacy. If the goal of the analysis is to forecast future inventory levels of different medications, when restocking is done weekly, then weekly aggregates are preferable to daily aggregates owing to less data recording errors and noise. However, for the early detection of outbreaks of disease, where alerts that are generated a day or two earlier can make a significant difference in terms of treatment, then weekly aggregates are of low quality. In addition to data frequency, the aggregation level is also important: for purposes of inventory medication level information is required, whereas for disease outbreak detection medications can be grouped by symptoms, and the symptom-aggregated daily series would be preferable.

In the on-line auctions example, bid times are typically recorded in seconds, and prices in a currency unit. On eBay.com, for example, bid times are reported at the level of seconds (e.g. August 20th, 2010, 03.14.07 Pacific Daylight Time) and prices at the dollar and cent level (e.g. $23.01). The forecasting model by Wang *et al.* (2008) uses bid times at second level and cent level bid amounts until the time of prediction to produce forecasts of price in cents for any second during the auction. In contrast, the forecasting model by Ghani and Simmons (2004) produces forecasts of the final price in terms of $5 intervals, using only information that is available at the start of the auction.

The concept of *rational subgroup* that is used in statistical process control is a special case of aggregation level. The rational subgroup set-up determines the level of process variability and the type of signals to detect. If the rational subgroup consists of measurements within a short period of a production process, then statistical process control will pick up short-term out-of-control

signals, whereas rational subgroups spread over longer periods will support detecting longer-term trends and out-of-control signals (see Kenett and Zacks (1998)). Using our notation, f is equivalent to statistical process control, X is the data, g_1 is the short-term signal, g_2 is the long-term signal and U is equivalent to good alerting behaviour.

6.2.2. *Data structure*

Data structure relates to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data collection mechanism. Data types include structured numerical data in different forms (e.g. cross-sectional, time series and network data) as well as unstructured, non-numerical data (e.g. text, text with hyperlinks, audio, video and semantic data). The InfoQ-level of a certain data type depends on the goal at hand. Bapna *et al.* (2006) discussed the value of different 'data types' for answering new research questions in electronic commerce research:

> 'For each research investigation, we seek to identify and utilize the best data type, that is, that data which is most appropriate to help achieve the specific research goals'.

An example from the on-line auction literature is related to the effect of 'seller feedback' on the auction price. Sellers on eBay receive numerical feedback ratings and textual comments. Although most explanatory studies of price determinants use the numerical feedback ratings as a covariate, a study by Pavlou and Dimoka (2006) showed that using the textual comments as a covariate in a model for price leads to much higher R^2-values (U) compared with using the numerical rating.

Corrupted and missing values require handling by removal, imputation, data recovery or other methods, depending on g. Wrong values might be treated as missing values when the purpose is to estimate a population parameter, such as in surveys where respondents intentionally enter wrong answers. Yet, for some goals, intentionally wrong values might be informative (see Section 5.1).

6.2.3. *Data integration*

Integrating multiple sources and/or types of data often creates new knowledge regarding the goal at hand, thereby increasing InfoQ. An example is the study estimating consumer surplus in on-line auctions (Section 2.3), where data from eBay (X_1) that lacked the highest bid values were combined with Cniper.com data (X_2) that contained the missing information. Estimating consumer surplus was impossible by using either X_1 or X_2, and only their combination yielded the sufficient InfoQ. In the auction example of Pavlou and Dimoka (2006), textual comments were used as covariates.

New analysis methodologies, such as functional data analysis and text mining, are aimed at increasing InfoQ of new data types and their combination. For example, in the on-line auction forecasting study (Section 2.2), functional data analysis was used to integrate temporal bid sequences with cross-sectional auction and seller information. The combination allowed more precise forecasts of final prices compared with models based on cross-sectional data alone. The functional approach has also enabled quantifying the effects of different factors on the *price process* during an auction (Bapna *et al.*, 2008b).

Another aspect of data integration is linking records across databases. Although record linkage algorithms are popular for increasing InfoQ, studies that use record linkage often employ masking techniques that reduce risks of identification and breaches of privacy and confidentiality. Such techniques (e.g. removing identifiers, adding noise, data perturbation and micro-aggregation) can obviously decrease InfoQ, even to the degree of making the combined data set

useless for the goal at hand. Solutions, such as 'privacy-preserving data mining' and 'selective revelation', are aimed at utilizing the linked data set with high InfoQ without compromising privacy (see, for example, Fienberg (2006)).

6.2.4. Temporal relevance

The process of deriving knowledge from data can be put on a timeline that includes the data collection, data analysis and study deployment periods as well as the temporal gaps between these periods (Fig. 1). These different durations and gaps can each affect InfoQ. The data collection duration can increase or decrease InfoQ, depending on the study goal, e.g. studying longitudinal effects *versus* a cross-sectional goal. Similarly, uncontrollable transitions during the collection phase can be useful or disruptive, depending on g. For this reason, on-line auction studies that collect data on fashionable or popular products (which generate large amounts of data) for estimating an effect try to restrict the data collection period as much as possible. The experiment by Katkar and Reiley (2006) (which was mentioned in Section 2.1) was conducted over 2 weeks in April 2000. The data in Wang *et al.* (2008) were collected in the non-holiday months of August and September 2005. In contrast, a study that is interested in comparing preholiday with post-holiday bidding or selling behaviour would require collection over a relevant period. The gap between data collection and analysis, which coincides with the *recency* criterion in Section 6.1, is typically larger for secondary data (data that were not collected for the purpose of the study). In predictive modelling, where the context of prediction should be as close as possible to the data collection context, temporal lags can significantly decrease InfoQ. For instance, a 2010 data set of on-line auctions for iPads on eBay will probably be of low InfoQ for forecasting or even estimating current iPad prices because of the fast changing interest in electronic gadgets.

Another aspect affecting temporal relevance is analysis timeliness, or the timeliness of $f(X|g)$. Raiffa (1970), page 264, called this an 'error of the fourth kind: solving the right problem too late'. Analysis timeliness is affected by the nature of X, by the complexity of f and ultimately by the application of f to X. The nature of a data set (size, sparseness, etc.) can affect analysis timeliness, and in turn affect its utility for the goal at hand. For example, computing summary statistics for a very large data set might take several hours, thereby deeming InfoQ low for the purpose of realtime tasks (g_1) but high for retrospective analysis (g_2). The computational complexity of f also determines analysis time: Markov chain Monte Carlo estimation methods and computationally intensive predictive algorithms take longer than estimating linear models or computing summary statistics. In the on-line auction price forecasting example, the choice of a linear forecasting model was needed for producing timely forecasts of an on-going auction. Wang *et al.* (2008) used smoothing splines to estimate price curves for each auction in the data set—information which is then used in the forecasting model. Although smoothing splines do not necessarily produce monotone curves (as would be expected of price curves in eBay-type

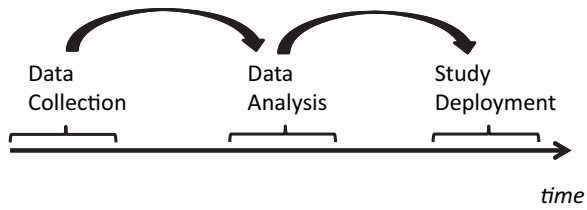

Fig. 1. Temporal durations and gaps that affect InfoQ: feedback arrows indicate the cyclic process of data collection and analysis

auctions), this method is much faster than monotone smoothing splines, and therefore generates higher InfoQ for realtime forecasting applications. Temporal relevance and analysis timeliness obviously depend on the availability of software and hardware as well as on the efficiency of the analysis team.

6.2.5. *Chronology of data and goal*

The choice of variables to collect, the temporal relationship between them and their meaning in the context of g critically affect InfoQ. We must consider the retrospective *versus* prospective nature of the goal as well as its type in terms of causal explanation, prediction or description (Shmueli, 2010). In predictive studies, the input variables must be available at the time of prediction. In contrast, in explanatory models causal arguments determine the relationship between dependent and independent variables. The term 'endogeneity', or reverse causation, can occur when a causal input variable is omitted from a model, resulting in biased parameter estimates. Endogeneity therefore yields low InfoQ in explanatory studies, but not necessarily in predictive studies, where omitting input variables can lead to higher predictive accuracy (see Shmueli (2010)). Related is the Granger causality test (Granger, 1969) aimed at determining whether a lagged time series X contains useful information for predicting future values g of a time series Y by using a regression model f.

In the on-line auctions context, the level of InfoQ that is contained in the 'number of bidders' for models of auction price depends on the study goal. Classic auction theory specifies the number of bidders as an important factor influencing price (generally, the more bidders the higher the price). Hence, data on number of bidders is of high quality in an explanatory model of price, yet for the purpose of forecasting the price of on-going on-line auctions, where the number of bidders is unknown until the end of the auction, the InfoQ of 'number of bidders', even if available in a retrospective data set, is very low. For this reason, the forecasting model by Wang *et al.* (2008) from Section 2.2 excludes number of bidders or number of bids and instead uses the cumulative number of bids until the time of prediction.

6.2.6. *Generalizability*

The utility of $f(X|g)$ is dependent on the ability to generalize f to the appropriate population. Two types of generalizability are statistical and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern or model f to other populations, or applying f estimated from one population to predict individual observations in other populations.

Determining the level of generalizability requires careful characterization of g. For instance, for inferring about a population parameter, statistical generalizability and sampling bias are the focus, and the question of interest is 'What population does the sample represent?' (Rao, 1985). In contrast, for predicting the values of new observations, the question of interest is whether f captures associations in the training data X (the data that are used for model building) that are generalizable to the to-be-predicted data. Another type of generalization in the context of ability testing is the concept of *specific objectivity* (Rasch, 1977). Specific objectivity is achieved if outcomes of questions in a questionnaire that is used to compare levels of students are independent of the specific questions and of other students. In other words, the purpose is to generalize from data on certain students answering a set of questions to the population of outcomes, irrespective of the particular responders or particular questions. The

type of required generalizability affects the choice of f and U. For instance, data-driven methods are more prone to overfitting, which conflicts with scientific generalizability. Statistical generalizability is commonly evaluated by using measures of sampling bias and goodness of fit. In contrast, scientific generalizability for predicting new observations is typically evaluated by the accuracy of predicting a hold-out set from the to-be-predicted population to protect against overfitting.

The on-line auction studies from Section 2 illustrate the different generalizability types. The 'effect of reserve price on final price' study (Katkar and Reiley, 2006) is concerned with statistical generalizability. Katkar and Reiley (2006) designed the experiment so that it produces a representative sample (see Section 4.1). Their focus is then on standard errors and statistical significance. The forecasting study by Wang *et al.* (2008) is concerned with generalizability to new individual auctions. They evaluated predictive accuracy on a hold-out set. The third study on 'consumer surplus in eBay' is concerned with statistical generalizability from the sample to all eBay auctions in 2003. In particular, because the sample was not drawn randomly from the population, Bapna *et al.* (2008a) performed a special analysis, comparing their sample with a randomly drawn sample (see appendix B in Bapna *et al.* (2008a)).

6.2.7. Construct operationalization

Constructs are abstractions that describe a phenomenon of theoretical interest. Measurable data are an operationalization of underlying constructs. For example, 'psychological stress' can be measured via a questionnaire or by physiological measures, such as cortisol levels in saliva (Kirschbaum and Hellhammer, 1989); 'economic prosperity' can be measured via income or by unemployment rate. The relationship between the underlying construct χ and its operationalization $X = \theta(\chi)$ can vary, and its level relative to g is another important aspect of InfoQ. The role of construct operationalization is dependent on $g(X = \theta(\chi|g))$, and especially on whether the goal is explanatory, predictive or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as X is assumed to measure χ, the variable is considered adequate. Using the example above, both questionnaire answers and physiological measurements would be acceptable for measuring psychological stress. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variable is critical. Predicting psychological stress as reported in a questionnaire (X_1) is different from predicting levels of a physiological measure (X_2). Hence, the InfoQ in predictive studies relies more heavily on the quality of X, whereas in explanatory studies InfoQ relies more on the adequacy of X for measuring χ.

Returning to the on-line auction context, the consumer surplus study relies on observable bid amounts, which are considered to reflect an underlying 'willingness-to-pay' construct for a bidder. The same construct is operationalized differently in other types of studies. In contrast, in price forecasting studies the measurable variable of interest is auction price, which is always defined very similarly. An example is the work by McShane and Wyner (2011), showing that, for purposes of predicting temperatures, theoretically based 'natural covariates' are inferior to 'pseudoproxies' that are lower dimension approximations of the natural covariates. Descriptive tasks are more similar to predictive tasks in the sense of the focus on the observable level. In descriptive studies, the goal is to uncover a signal in a data set (e.g. to estimate the income distribution or to uncover the temporal patterns in a time series). Because there is no underlying causal theory behind descriptive studies, and because results are reported at the level of the measured variables, InfoQ relies, as in predictive tasks, on the quality of the measured variables rather than on their relationship to an underlying construct.

6.2.8. Communication

Effective communication of the analysis $f(X|g)$ and its utility U directly affects InfoQ. Common communication media are visual, textual and verbal presentations and reports. Within research environments, communication focuses on written publications and conference presentations. Research mentoring and the refereeing process are aimed at improving communication (and InfoQ) within the research community. Research results are communicated to the public via articles in the popular media, interviews on television and conferences such as `www.ted.com`, and more recently through blogs and other Internet media. Here the risk of miscommunication is much larger. For example, the 'consumer surplus in eBay auctions' study was covered by public media. However, the main results were not always conveyed properly by journalists. For example, the nytimes.com article (`http://bits.blogs.nytimes.com/2008/01/28/tracking-consumer-savings-on-ebay`) failed to mention that the study results were evaluated under different assumptions, thereby affecting generalizability. As a result, some readers doubted the study results (for example 'is the Cniper sample skewed[?]'). In response, one of the study co-authors posted a reply on line to improve communication.

In industry, communication is typically done via internal presentations and reports. The failure potential of O-rings at low temperatures that caused the Nasa shuttle Challenger disaster was ignored despite being known to engineers. Yet, the engineers failed to communicate the results of their analysis: the 13 charts that were circulated to the teleconferences did not clearly show the relationship between the temperature in 22 previous launches and the 22 recordings of O-ring conditions (see Tufte (1992)). In our notation, the meaning of f equivalent risk analysis and its implications were not properly communicated.

6.3. Rating-based information quality assessment

Similar to the use of 'data quality' dimensions by statistical agencies for evaluating data quality, we evaluate the eight InfoQ-dimensions to assess InfoQ. This evaluation integrates different aspects of a study and assigns an overall InfoQ-score. The broad perspective of InfoQ-dimensions is designed to help researchers to enhance the added value of their studies. Kenett *et al.* (2010) applied this approach to evaluate research proposals of graduate students at the Faculty of Economics of the University of Ljubljana in Slovenia. The InfoQ-report provided important feedback to the students that helped to make their work more relevant and more comprehensive.

Assessing InfoQ by using quantitative metrics can be done in several ways. We present a rating-based approach that examines a study report and scores each of the eight InfoQ-dimensions. A coarse-grained approach is to rate each dimension on a 1–5-scale, with '5' indicating 'high' achievement in that dimension. The ratings Y_i, $i = 1, \ldots, 8$, can then be normalized into a *desirability function* for each dimension, $0 \leqslant d(Y_i) \leqslant 1$, which are then combined to produce an overall InfoQ-score by using the geometric mean of the individual desirabilities:

$$\text{InfoQ} = \{d_1(Y_1)\, d_2(Y_2) \ldots d_8(Y_8)\}^{1/8}.$$

The geometric mean is smaller than the arithmetic mean and applies only to positive elements. Overall desirability functions, with geometric means of individual desirabilities, provide an effective and efficient way to achieve overall optima (Derringer and Suich, 1980). For more on desirability and integrated measures see Figini *et al.* (2010). In the next section we apply this approach to the eBay auctions experiment example.

6.4. Example: information quality assessment of on-line auctions experimental data

As described in Section 2.1, Katkar and Reiley (2006) investigated the effect of two types of

reserve price on the final auction price on eBay. Their data X came from an experiment selling 25 identical pairs of Pokémon cards, where each card was auctioned twice, once with a public reserve price and once with a secret reserve price. The data consist of complete information on all 50 auctions. Katkar and Reiley used linear regression f to test for the effect of private or public reserve on the final price and to quantify it. The utility U was statistical significance to evaluate the effect of private or public reserve price, and the regression coefficient for quantifying the magnitude of the effect. They concluded that

> 'a secret-reserve auction will generate a price \$0.63 lower, on average, than will a public-reserve auction'.

We evaluate the eight InfoQ-dimensions on the basis of Katkar and Reiley (2006). A more thorough evaluation would have required interaction with them and access to their data. For demonstration purposes we use a 1–5-scale and generate an InfoQ-score based on a desirability function with $d(1) = 0$, $d(2) = 0.25$, $d(3) = 0.5$, $d(4) = 0.75$ and $d(5) = 1$.

6.4.1. Data resolution

The experiment was conducted over 2 weeks in April 2000. We therefore have no data on possible seasonal effects during other periods of the year. Data resolution was in US cents but individual bids were dropped and only the final price was considered. Other time series (e.g. the cumulative number of bids) were also aggregated to create end-of-auction statistics such as 'total number of bids'. Given the general goal of quantifying the effect of using a secret *versus* public reserve price on the final price of an auction, it seems that the data were somewhat restrictive. The 2-week data window allows for good control of the experiment but limits data resolution for studying a more general effect. Hence we rate the data resolution as $Y_1 = 4$ ('high').

6.4.2. Data structure

The data included only information on the factor levels that were set by Katkar and Reiley (2006) and the three outcomes final price, whether the auction transacted and the number of bids received. These data were either set by the experimenters or collected from the auction Web site. Although time series data were potentially available for the 50 auctions (e.g. the series of bids and cumulative number of bidders), Katkar and Reiley (2006) aggregated them into auction totals. Textual data were available, but not used. For example, bidder usernames can be used to track individual bidders who placed multiple bids. With respect to corrupted data, one auction winner unexpectedly rated the sellers, despite their request to refrain from doing so (to keep the rating constant across the experiment). Luckily, this corruption did not affect the analysis owing to the study design. Another unexpected source of data corruption was eBay's policy on disallowing bids below a public reserve price. Hence, the total number of bids in auctions with a secret reserve price could not be compared with the same measure in public reserve price auctions. Katkar and Reiley resorted to deriving a new 'total serious bids' variable, which counts the number of bids above the secret reserve price.

Given the level of detailed attention to the experimental conditions, but the lack of use of available time series and textual data, we rate this dimension as $Y_2 = 4$ (high).

6.4.3. Data integration

Katkar and Reiley (2006) analysed the 2-week data in the context of an experimental design strategy. The integration with the DOE factors was clearly achieved. No textual or other semantic data seemed to have been integrated. We rate this dimension as $Y_3 = 4$ (high).

6.4.4. Temporal relevance

The short duration of the experiment and the experimental design assured that the results would not be confounded with the effect of time. The experimenters tried to avoid confounding the results with a changing seller rating and therefore actively requested winners to avoid rating the seller. Moreover, the choice of Pokémon cards was aligned with timeliness, since at the time such items were in high demand. Finally, because of the retrospective nature of the goal, there is no urgency in conducting the data analysis shortly after data collection. We rate this dimension as $Y_4 = 5$ ('very high').

6.4.5. Chronology of data and goal

The causal variable (secret or public reserve) and the blocking variable (week) were determined at the auction design stage and manipulated before the auction started. We rate this dimension as $Y_5 = 5$ (very high).

6.4.6. Generalizability

The study is concerned with statistical generalizability: do effects that were found in the sample generalize to the larger context of on-line auctions? One possible bias, which was acknowledged by Katkar and Reiley (2006), is the seller's rating of 0 (indicating a new seller) which limits the generalizability of the study to more reputable sellers. In addition, they limited the generality of their results to low value items, which might not generalize to more expensive items. We rate this dimension as $Y_6 = 3$ ('acceptable').

6.4.7. Construct operationalization

Katkar and Reiley (2006) considered two theories that explain the effect of a secret *versus* public reserve price on the final price. One is psychological, where bidders can become 'caught up in the bidding' at low bid amounts and end up bidding more than they would have if the bidding had started higher. The second theory is a model of rational bidders: 'an auction with a low starting bid and a high secret reserve can provide more information to bidders than an auction with a high starting bid'. Although these two theories rely on operationalizing constructs such as 'information' and 'caught up in the bidding', Katkar and Reiley limited their study to eBay's measurable reserve price options and final prices. We rate this dimension as $Y_7 = 3$ (acceptable).

6.4.8. Communication

This research study communicated the analysis via a paper published in a peer-reviewed journal. Analysis results are presented in the form of a scatter plot, a series of estimated regression models (estimated effects and standard errors) and their interpretation in the text. We assume that Katkar and Reiley (2006) made additional dissemination efforts (for example, the paper is publicly available on line as a working paper). The paper's abstract is written in a non-technical and clear way, and can therefore be easily understood not only by academics and researchers but also by eBay participants. The main communication weakness of the analysis is in terms of visualization, where plots would have conveyed some of the results more clearly. We therefore rate this dimension as $Y_8 = 4$ (high).

6.4.9. Information quality score

On the basis of these informal assessments representing expert opinions derived from the single publication on the auction experiments, we obtain its InfoQ-score of 0.73, i.e. high. The relatively

weak points are generalizability and concept operationalization; the strong points are temporal relevance and chronology of data and goal. An effort to review the scores with some perspective of time proved these scores to be robust even though expert opinions will tend to differ, to a large or minimal extent. To derive consensus-based scores, one can ask a number of experts to review the case (say 3–5) and compare their scores. If the scores are consistent, one can derive a consistent InfoQ-score. If they show discrepancies, one would conduct a consensus meeting of the experts where the reasoning behind their score is discussed and some score reconciliation is attempted. If a range of scores remains then the InfoQ-score can be presented as a range of values.

7. Discussion and future directions

This paper defines the general concept of InfoQ. The assessment of InfoQ is designed to help researchers and practitioners to ensure the value of data analysis in terms of the stated goal. The question of 'the potential of a particular data set to achieve a particular goal using a given empirical analysis method' often arises in practice, yet there has not been a formalization or direct discussion of InfoQ within the larger framework of statistical methodology. Hence researchers and analysts must resort to improvising solutions, which often leads to a waste of time, effort and other resources. An example is the evaluation of polygraph validity by the National Academies committee (Fienberg, 2003). The committee, charged with evaluating the validity of polygraph usage for employment screening, relied on 57 crime-related (true and mock) studies to determine the accuracy of polygraphs. They concluded that

> 'Because the studies of acceptable quality all focus on specific incidents, generalization from them to uses for screening is not justified'.

In summarizing the quality of the data that were available to them for assessing polygraph validity, they write:

> 'The general quality of the evidence for judging polygraph validity is relatively low: the substantial majority of the studies most relevant for this purpose were below the quality level typically needed for funding by the National Science Foundation or the National Institutes of Health'.

This work is aimed at providing a more streamlined approach to InfoQ and its assessment. We propose eight dimensions of InfoQ that are conditional on the goal g and affect the data X, analysis f and utility U. In their discussion of the role of statisticians in the 'Post-Financial Meltdown era', Hoerl and Snee (2009) addressed the issue of data collection and its value:

> 'While statisticians almost always vote for collecting data, we need to remind ourselves that data cost money. The organization should be cost effective in the data it collects, making sure it knows ahead of time how the data will be used. Management will generally support data collection when it understands the value of the problem being addressed, and how the data will help solve the problem.'

Moreover, when considering the type of data to use for a study, one can compare potential data sets in terms of their InfoQ and choose accordingly. In addition, InfoQ-assessment can assist in prioritizing and ranking studies or projects, what Hoerl and Snee (2009) called 'effective prioritization—working on the right things'. Awareness of InfoQ helps to guide the empirical analysis and the final reporting in maintaining a close focus on the final study goal.

Sometimes during the study the goal changes or new goals are added. Cox (2009) noted that

> 'Objectives may be redefined, hopefully improved, and sometimes radically altered as one proceeds'.

Similarly, Friedman *et al.* (1999) commented that, in clinical trials,

> 'One would like answers to several questions, but the study should be designed with only one major question in mind'.

Therefore, often multiple questions will be answered by using data that were collected through a design for answering a single primary question. Assessing InfoQ relative to other questions is therefore crucial. For example, the evaluation of adverse effects is important, yet not the primary goal of a clinical trial and hence

> 'clinical trials have inherent methodological limitations in the evaluation of adverse effects. These include inadequate size, duration of follow-up, and restricted participant selection'

(Friedman *et al.* (1999), page 182).

Finally, in light of the burning need to strengthen the relationship between the functional problem and the statistical problem, as well as the functional and statistical solutions, InfoQ can serve as a powerful translation tool. We believe that statisticians should take a leadership role in assessment, research and education of InfoQ. Pregibon, Head of the Statistics Group at Google's New York Engineering office, noted that

> 'it is essential that managers be able to translate business or other functional problem into the appropriate statistical problem before it can be "handed off" to a technical team'

(the foreword in Shmueli *et al.* (2010)). Focusing on InfoQ before, during and after an analysis will put the statistician in a more integrated and communicable relationship with domain experts.

Our discussion of InfoQ as a crucial component in the empirical analysis framework calls for further discussion and research in various directions. A few directions are described next.

In assessing InfoQ, we proposed a rating-based approach. Future research is needed on specific implementations such as investigating the reliability of ratings across different raters. One alternative approach for assessing InfoQ is to use pilot samples before collecting the full data set. For example, early studies in biosurveillance were aimed at determining the usefulness of tracking prediagnostic data for detecting outbreaks of disease earlier than more traditional diagnostic measures. At first, small amounts of data from various sources were examined, and their potential for early detection evaluated (Wagner *et al.*, 2001). Another example is futility studies in clinical trials, which are conducted half way through the trial, to determine whether the trial has the potential to generate useful results. If the outcomes up to that point indicate that no statistical significance can be achieved at the end of the trial, the committee recommends halting the trial. The disadvantage of pilot studies, however, is that they consume time and delay the actual study.

Another assessment approach is to use exploratory data analysis and especially data visualization using software that provides interactivity and a range of visualizations. These techniques can support the analyst in exploring and determining, with 'freehand format', the level of InfoQ in the data. Exploratory data analysis is often conducted iteratively and used to detect salient features and outliers, to trigger further investigations and, in some cases, to collect additional data. Other exploratory tools that are useful for assessing InfoQ, which were termed 'exploratory models' by De Veaux (2009), include classification and regression trees, cluster analysis and data reduction techniques.

Although we discussed several dimensions of InfoQ, there are others that might be considered and new dimensions might evolve over time. For example, in today's environment, an important aspect of InfoQ is related to privacy and confidentiality of data. In some areas, InfoQ could include a measure of risk in terms of confidentiality or human subjects. Analysing individual records might have higher InfoQ in terms of answering the question of interest, but aggregate

data might be preferable in terms of confidentiality. Hence, not only can new dimensions arise, but so also can constraints that can limit the potential level of InfoQ.

Although we discussed InfoQ in the context of the study design stage (Section 4) and post-data-collection stage (Section 5), there are various issues that should be considered during the study implementation stage. We touched on compliance of patients and physicians in clinical trials and its effect on InfoQ. However, future discussion is needed regarding the effect of other study conduct practices on InfoQ.

Another direction for future research is the effect of technological advances such as cloud computing on InfoQ. Although technological advances have clearly improved data quality, their effect on analysis quality is mixed. On the one hand, advances in statistical software have made statistical analysis and data mining very easy to perform by non-experts, and therefore very easy to abuse. Commonly used tools are graphical interfaces that allow the user to 'drag and drop' a statistical method or data mining algorithm on a data set, or simply to run a statistical method by choosing it from a drop-down menu. On the other hand, technological advances have also contributed to improving analysis quality by enabling computationally intensive methods and the derivation of new statistical methodologies (e.g. functional data analysis and Markov chain Monte Carlo sampling). In sum, technological advances have increased the variability of analysis quality, by advancing the use of statistical and data mining methods across a wider range of users, and by advancing the level of complexity of modelling. Given that InfoQ is conditional on data quality and analysis quality, technological advances clearly affect InfoQ as well. Moreover, there might be technological advances that influence InfoQ directly. For example, recent developments in data visualization and visualization tools by artists, human–computer interaction experts and other non-statistician researchers have created new and different ways of looking at and interpreting data. For data analysts, an important question is whether the use of such methods or tools increases or decreases InfoQ, and whether the contribution to InfoQ justifies the extra resources required. These are just a few directions for future InfoQ research. Such research is necessarily multidisciplinary and has the potential to expand the scope of application and the influence of statistical work in science, business and industry.

Acknowledgements

In preparing this paper we benefited from inputs, comments and suggestions of several people including Sir David Cox, Blan Godfrey, Shirley Coleman, Irad ben Gal, John Shade, Douglas Montgomery, Charles Tapiero and David Steinberg. Their contributions are gratefully acknowledged. We also thank the referees and the Joint Editor for their helpful suggestions which greatly improved this paper.

References

Angst, C. M., Agarwal, R. and Kuruzovich, J. (2008) Bid or buy?: individual shopping traits as predictors of strategic exit in on-line auctions. *Int. J. Electron. Commrce*, **13**, 59–84.

Bapna, R., Goes, P., Gopal, R. and Marsden, J. R. (2006) Moving from data-constrained to data-enabled research: experiences and challenges in collecting, validating and analyzing large-scale e-commerce data. *Statist. Sci.*, **21**, 116–130.

Bapna, R., Goes, P., Gupta, A. and Jin, Y. (2004) User heterogeneity and its impact on electronic auction market design: an empirical exploration. *Mangmnt Inform. Syst. Q.*, **28**, 21–43.

Bapna, R., Jank, W. and Shmueli, G. (2008a) Consumer surplus in online auctions. *Inform. Syst. Res.*, **19**, 400–416.

Bapna, R., Jank, W. and Shmueli, G. (2008b) Price formation and its dynamics in online auctions. *Decsn Supp. Syst.*, **44**, 641–656.

Bates, R., Kenett R., Steinberg D. and Wynn, H. (2006) Achieving robust design from computer simulations. *Qual. Technol. Quant. Mangmnt*, **3**, 161–177.

Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C.-H. and Tu, J. (2007) A framework for validation of computer models. *Technometrics*, **49**, 138–154.
Biemer, P. and Lyberg, L. (2003) *Introduction to Survey Quality*. Hoboken: Wiley.
Borle, S., Boatwright, P. and Kadane, J. B. (2006) The timing of bid placement and extent of multiple bidding: an empirical investigation using ebay online auctions. *Statist. Sci.*, **21**, 194–205.
Boslaugh, S. (2007) *Secondary Data Sources for Public Health: a Practical Guide*. Cambridge: Cambridge University Press.
Cox, D. R. (2009) Personal communication with first author.
Deming, W. E. (1953) On the distinction between enumerative and analytic studies. *J. Am. Statist. Ass.*, **48**, 244–255.
Derringer, G. and Suich, R. (1980) Simultaneous optimization of several response variables. *J. Qual. Technol.*, **12**, 214–219.
De Veaux, R. D. (2009) Successful exploratory data mining in practice. *JMP Explorer Series*. (Available from `http://www.williams.edu/Mathematics/rdeveaux/success.pdf`.)
Ding, Y. and Simonoff, J. (2010) An investigation of missing data methods for classification trees applied to binary response data. *J. Mach. Learn. Res.*, **11**, 131–170.
Fienberg, S. E. (2003) *The Polygraph and Lie Detection*. New York: National Academy Press.
Fienberg, S. E. (2006) Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation. *Statist. Sci.*, **21**, 143–154.
Figini, S., Kenett, R. S. and Salini, S. (2010) Integrating operational and financial risk assessments. *Qual. Reliab. Engng Int.*, **26**, 887–897.
Friedman, L. M., Furberg, C. D. and DeMets, D. L. (1999) *Fundamentals of Clinical Trials*, 3rd edn. New York: Springer.
Ghani, R. and Simmons, H. (2004) Predicting the end-price of online auctions. *Int. Wrkshp Data Mining and Adaptive Modelling Methods for Economics and Management, Pisa*.
Giovanni, E. (2008) *Understanding Economic Statistics*. Geneva: Organisation for Economic Co-operation and Development Publishing.
Godfrey, A. B. (2008) Eye on data quality. *Six Sigma Forum Mag.*, **8**, 5–6.
Granger, C. W. J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.
Hand, D. J. (2008) *Statistics: a Very Short Introduction*. Oxford: Oxford University Press.
Hoerl, R. W. and Snee, R. D. (2009) Post-financial meltdown: what do the services industries need from us now? *Appl. Stochast. Mod. Bus. Indstry*, **25**, 522–526.
Huang, D. and Allen, T. T. (2005) Design and analysis of variable fidelity experimentation applied to engine valve heat treatment process design. *Appl. Statist.*, **54**, 443–463.
Jank, W. and Shmueli, G. (2010) *Modeling Online Auctions*. Hoboken: Wiley.
Katkar, R. and Reiley, D. H. (2006) Public versus secret reserve prices in eBay auctions: results from a Pokémon field experiment. *Adv. Econ. Anal. Poly*, **6**, no. 2, article 7.
Kenett, R. S., Coleman, S. and Ograjenšek, I. (2010) On quality research: an application of InfoQ to the PhD research process. In *Proc. European Network for Business and Industrial Statistics 10th A. Conf. Business and Industrial Statistics, Antwerp, Sept. 12th–16th.*
Kenett, R. S. and Salini, S. (2011) Modern analysis of customer surveys: comparison of models and integrated analysis (with discussion). *Appl. Stochast. Mod. Bus. Indstry*, **27**, 465–475.
Kenett, R. S. and Salini, S. (2012) *Modern Analysis of Customer Satisfaction Surveys: with Applications using R*. Chichester: Wiley.
Kenett, R. S. and Zacks, S. (1998) *Modern Industrial Statistics: Design and Control of Quality and Reliability*. San Francisco: Duxbury.
Kimball, A. W. (1957) Errors of the third kind in statistical consulting. *J. Am. Statist. Ass.*, **52**, 133–142.
Kirschbaum, C. and Hellhammer, D. H. (1989) Salivary cortisol in psychobiological research: an overview. *Neuropsychobiology*, **22**, 150–169.
Lin, M., Lucas, H. and Shmueli, G. (2009) Too big to fail: large samples and the P-value problem. *Working Paper RHS 06-068*. Smith School of Business, University of Maryland, College Park.
Little, R. (2009) Weighting and prediction in sample surveys. *Working Paper 81*. Department of Biostatistics, University of Michigan, Ann Arbor.
Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. New York: Wiley.
Lucking-Reiley, D., Bryan, D., Prasad, N. and Reeves, D. (2007) Pennies from ebay: the determinants of price in online auctions. *J. Industrl Econ.*, **55**, 223–233.
Mallows, C. (1998) The zeroth problem. *Am. Statistn*, **52**, 1–9.
Marshall, A. (1920) *Principles of Economics*, 8th edn. London: MacMillan.
McShane, B. B. and Wyner, A. J. (2011) A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1000 years reliable? *Ann. Appl. Statist.*, **5**, 5–44.
Mealli, F., Pacini, B. and Rubin, D. B. (2011) Statistical inference for causal effects. In *Modern Analysis of Customer Satisfaction Surveys: with Applications using R* (eds R. S. Kenett and S. Salini), pp. 173–192. Chichester: Wiley.

Patzer, G. L. (1995) *Using Secondary Data in Marketing Research*. Greenwood.
Pavlou, P. A. and Dimoka, A. (2006) The nature and role of feedback text comments in online marketplaces: implications for trust building, price premiums, and seller differentiation. *Inform. Syst. Res.*, **17**, 392–414.
Raiffa, H. (1970) *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Reading: Addison-Wesley.
Rao, C. R. (1985) Weighted distributions arising out of methods of ascertainment: what population does a sample represent? In *A Celebration of Statistics: the ISI Centenary Volume* (eds A. C. Atkinson and S. E. Fienberg), pp. 543–569. New York: Springer.
Rasch, G. (1977) On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Dan. Yearb. Philos.*, **14**, 58–93.
Redman, T. (2007) Statistics in data and information quality. In *Encyclopedia of Statistics in Quality and Reliability* (eds F. Ruggeri, R. S. Kenett and F. Faltin). Hoboken: Wiley.
Roth, A. E. and Ockenfels, A. (2002) Last-minute bidding and the rules for ending second-price auctions: evidence from ebay and Amazon on the internet. *Am. Econ. Rev.*, **92**, 1093–1103.
Saar-Tsechansky, M., Melville, P. and Provost, F. (2009) Active feature-value acquisition. *Mangmnt Sci.*, **55**, 664–684.
Saar-Tsechansky, M. and Provost, F. (2007) Handling missing features when applying classification models. *J. Mach. Learn. Res.*, **8**, 1625–1657.
Shmueli, G. (2010) To explain or to predict? *Statist. Sci.*, **25**, 289–310.
Shmueli, G. and Koppius, O. R. (2011) Predictive analytics in information systems research. *Mangmnt Inform. Syst. Q.*, **35**, 553–572.
Shmueli, G., Patel, N. R. and Bruce, P. (2010) *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, 2nd edn. Hoboken: Wiley.
Shmueli, G., Russo, R. P. and Jank, W. (2007) The BARISTA: a model for bid arrivals in online auctions. *Ann. Appl. Statist.*, **1**, 412–441.
Tufte, R. E. (1992) *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.
Tukey, J. W. (1962) The future of data analysis. *Ann. Math. Statist.*, **33**, 1–67.
Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading: Addison-Wesley.
UK Department of Health (2004) A strategy for NHS information quality assurance—consultation draft. Department of Health, London. (Available from `www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4125508`.)
Wagner, M. M., Tsui, F. C., Espino, J. U., Dato, V. M., Sittig, D. F., Caruana, R. A., McGinnis, L. F., Deerfield, D. W., Druzdzel, M. J. and Fridsma, D. B. (2001) The emerging science of very early detection of disease outbreaks. *J. Publ Hlth Mangmnt Pract.*, **7**, 51–59.
Wang, S., Jank, W. and Shmueli, G. (2008) Explaining and forecasting online auction prices and their dynamics using functional data analysis. *J. Bus. Econ. Statist.*, **26**, 144–160.
Wang, R.Y., Kon, H. B. and Madnick, S. E. (1993) Data quality requirements analysis and modeling. *9th Int. Conf. Data Engineering, Vienna*.

Comments on the paper by Kenett and Shmueli

Paul Biemer (*RTI International, Durham*)

My compliments go to the authors on a well-written paper and for providing a very useful framework for evaluating the quality of a set of data, the statistical analysis and study report. The concept of InfoQ is an important and novel contribution to data science; however, as the authors note, the ideas of a quality framework and its components are not new. The InfoQ framework draws inspiration from the quality frameworks that have been developed by Statistics Canada, Eurostat and some European countries, notably Statistics Sweden where the concept originated (Felme *et al.*, 1976). For example, the Eurostat framework consists of seven dimensions (Eurostat, 2009). These are relevance (which is similar to *data resolution* in the InfoQ model), accuracy (similar to *data structure*), timeliness and punctuality (similar to *temporal relevance* and *chronology of data and goal*), accessibility and clarity (similar to *communication*), completeness (another facet of *data resolution*) and comparability and coherence (which are similar to *data integration*). The InfoQ dimensions of *generalizability* and *construct operationalization* are related to the concepts of representativity and validity respectively that are embedded in Eurostat's accuracy dimension.

What is different about the InfoQ framework is that it extends Eurostat's concept of 'total survey quality' to encompass the key aspects of quality for a statistical study—i.e. 'total study quality'. In the same way that the survey quality framework has been useful for both public and private statistical offices for evaluating the quality of their survey products, InfoQ should be useful for the evaluation of other statistical studies such as clinical trials and laboratory and field experiments. InfoQ also provides a certain level of rigour that is absent from the survey quality model. As an example, the elegant mathematical definition of InfoQ that

appears in Section 1.2—i.e. InfoQ$(f, X, g) = U\{f(X|g)\}$—is clear, concise and comprehensive. It conveys that InfoQ is not only an attribute of the data set; it is also an attribute of the analysis and reporting of the results.

Like the survey quality framework, the value of the InfoQ framework is its potential for considerably improving statistical products. The recent experiences of Statistics Sweden (Biemer *et al.*, 2012) in applying the survey quality framework to a number of their statistical products provide relevant lessons learned for InfoQ as well. To set the stage briefly, in 2011, the Swedish Ministry of Finance directed Statistics Sweden to develop a system of quality evaluations for their key statistical products. This system was to include metrics that reflect current quality as well as changes in quality over time. Later that year, a quality review process based on Statistics Sweden's total survey quality framework was pilot tested on eight statistical programmes: the annual municipal accounts, the consumer price index, the Foreign Trade of Goods Survey, the Labour Force Survey, the national accounts, the Structural Business Survey, the business register and the total population register. The review was limited to accuracy only, but it is now being expanded to relevance, accessibility, comparability and coherence, and timeliness.

The following points summarize some of the important lessons learned from these evaluations.

(a) Prior experience with quality reviews suggested that using internal (Statistics Sweden) evaluators would not yield unbiased assessments. Instead, Statistics Sweden found it critically important to engage external evaluators in the review process who are experts in national statistics and survey quality.
(b) Although objective quality assessments are always the goal, some level of subjectivity is unavoidable in any evaluation process. However, subjectivity can be minimized if detailed guidelines are provided regarding what constitutes each level of quality, from poor to excellent. An example of quality guidelines can be found in Biemer *et al.* (2012).
(c) Evaluating an entire dimension of quality directly can be quite difficult and unreliable. It is better to decompose the dimension further into mutually exclusive and exhaustive components (i.e. subdimensions). As an example, the accuracy review assessed the errors from eight potential sources (i.e. specification, frame, non-response, measurement, data process, sampling, model or estimation, and revision) emphasizing those sources of error that provided the greatest risks to data quality. Similarly, accessibility can be evaluated along six subdimensions:
    (i) ease of data access,
    (ii) documentation (including metadata),
    (iii) quality reports,
    (iv) alternative forms of dissemination,
    (v) user support and
    (vi) timeliness of supporting documentation.
(d) Even these decompositions proved to be too coarse for achieving reliable ratings and further decomposition of each subdimension was needed. Each subdimension was rated according to five quality criteria. As an example, the criteria applied to accuracy were
    (i) knowledge of the source of error,
    (ii) communication of the risk of the source of error to data users,
    (iii) expertise within the programme team to mitigate the risks of error from this source,
    (iv) compliance with standards and best practices in the design, data collection, data processing and the estimation process, and
    (v) planning and achievement towards risk mitigation.

    Similar criteria were developed for the other quality dimensions and their subdimensions.
(e) Finally, it is important that quality reviews be based on complete knowledge of the products and processes to be evaluated. This can only be accomplished by providing detailed and comprehensive documentation of the processes involved in producing the estimates from each programme. For example, assessing accuracy for the Labour Force Survey statistics required complete documentation of the survey design, data collection and data processing procedures, estimation methodology and data dissemination approaches. In addition, the reviewers required reports on prior quality investigations, current planning documentation and evidence of progress towards risk mitigation activities.

On the basis of the experiences of Statistics Sweden, important strengths of the InfoQ-system are that it comprehensively defines study quality in a standardized way. Standardization permits reproducibility of the assessment results; but, more importantly, it admits comparisons of quality both across studies and

within studies across time. For example, in the Statistics Sweden model, the accuracy of the eight products was compared by error source and criterion. These comparisons provided important information on where resources and improvement efforts are most needed to improve quality. In addition, when the evaluation process is repeated in future years, standardization allows the prior years' evaluations to be used as baselines for assessing quality improvement.

Areas where InfoQ can be strengthened can also be discerned from the Statistics Sweden experience. Like their survey quality counterparts, the InfoQ-dimensions are too broad and complex to be evaluated directly. An important improvement would be to decompose each dimension into subdimensions and then to develop specific criteria by which each subdimension can be judged. Guidelines to aid the judges in assigning quality ratings are essential for reliable assessments of these criteria. In addition, the expertise of the judges and the adequacy of the documentation that is available to the judges should also be emphasized in the InfoQ-model. The examples in Section 6.4 illustrate why this is important. Incomplete documentation on the Katkar and Reiley study would not only lead to inaccurate quality assessments, but also unreliable assessments where two judges rating the InfoQ-dimensions could arrive at very different ratings due to their incomplete knowledge of the study details.

Notwithstanding these shortcomings, InfoQ is an important step in defining and communicating the quality of a statistical study. It focuses much needed attention on the need to improve the quality of scientific data and the findings from investigations that rely on their analysis. In that regard, it is a significant contribution to the field of statistics. With a few additions and refinements, InfoQ could be a useful tool for study designers, data analysts, researchers and other users of the results of scientific investigations. It could benefit from the experience of having been applied in a real situation similar to the work at Statistics Sweden on the survey quality model.

Barry Schouten (*Statistics Netherlands, The Hague, and Utrecht University*)

General remarks

The paper by Ron Kenett and Galit Shmueli comes exactly at the right time. Pressing budgets, an increasing perceived response burden among enterprises and an increasing demand for statistical data are making researchers and analysts turn their view towards secondary data. At the same time the emergence of so-called organic data or 'big data' on the Internet and in commercial databases has opened up a new field of information ready to explore. The authors rightfully state that it is paramount that we take a structured look at the value of this information for different research purposes. Doing so, we can detect strengths and weaknesses and adapt methodology or search for supplemental data.

I very much appreciate the initiative taken by the authors and believe that they have identified a variety of important ingredients to quality of information. In what follows, I shall briefly review the framework proposed. Furthermore, I have applied the eight InfoQ-dimensions to two data sets that I have analysed in the past and I shall briefly discuss the outcomes of this study. First, though, I would like to make two side remarks.

A decisive ingredient to information quality is the goal or goals that researchers have set when starting an analysis. From my own experience and looking at analyses done by others, I conclude that research goals may not be that rigorously defined and/or stated beforehand. They should of course be well defined to assess the fitness for use of data, but often data exploration and analysis sharpen the mind of the researcher and goals become formed interactively. As such I believe that an assessment of the InfoQ-dimensions may actually be helpful in deriving more specific and elaborated analysis goals. Still, I suspect that the framework is only powerful when researchers have well-defined goals.

Then there is the use of multiple data sets. The framework set out by the authors is most useful when it is applied to secondary data. Most researchers, however, will recognize that the secondary data may have too little information and may require a number of strong assumptions to make inference. They will often turn to multiple data sets, e.g. a range of registry data or a combination of registry and survey data. In such a setting the assessment of information quality becomes much more difficult as weaknesses of some data may be overcome by strengths of other data. The framework should be extended to this setting as it is more realistic than a single set of data.

The concepts and framework

In my mind the concepts data quality, analysis quality and information quality are slightly different from how they are set out by the authors. I would say that there are data and there are metadata. Metadata can be structural when they describe the design and specification of the data to be measured, and they can be descriptive when they are about data content of one or more specific realizations. I

would say that data quality is about the distance between ideal descriptive metadata and true descriptive metadata, and information quality is about the distance between desired structural metadata and the true structural metadata. Since terminology is ambiguous, it may be easier to state that data quality is about the data that one intended to have and information quality is about the data that one desired to have.

The authors do not use the term metadata but give a clear account of how such metadata arise from research goals. For primary data research goals are translated to structural metadata and information quality may be high, but still there may be a gap due to all kinds of practical, ethical or logistical restrictions. For secondary data, obviously, the researcher or analyst is searching for data with high information quality, but the gap may be much larger. Data quality and information quality could be assessed without a specific goal but would then become very descriptive and exhaustive. This is clear from the eight proposed InfoQ-dimensions which are not specific to a certain goal. Only when a goal is defined will data quality and information quality really have focus.

Different methods are used to improve data quality and information quality. Data processing, editing, imputing and weighting from my point of view are about reducing the gap between the data at hand and the data that one intended to have. These statistical methods bring in new, external, possibly longitudinal, data and aim at improving data quality. Data analysis is about bridging the gap between intended and desired data. Researchers construct models that make some simplifying assumptions based on experience and knowledge about underlying causal relations. In other words, they bring in external information.

The concept 'analysis quality', which was coined by the authors, remains somewhat vague. I believe that this term could refer both to the extent to which data quality is improved and to the extent to which information quality is improved. The skills, experience and knowledge of the researcher or analyst are of course important ingredients to improving quality, but in many cases the analysis can simply not be done without accepting the limitations of the available data and information.

Does this different view on the concepts of data quality, analysis quality and information quality change the utility of InfoQ? I do not think so. The eight InfoQ-dimensions all refer to structural metadata. Also from a different viewpoint, they could be used. They give a structured look at the gap between desired and intended data and a starting point to improve information quality.

Application of the eight InfoQ-dimensions

The proof of the pudding is in the eating, they say. I have applied the eight InfoQ-dimensions to two very different data sets: recordings of electric current through ion channels in plant cells and the respondent data to the Labour Force Survey (LFS).

Ion channels are pores in cell membranes that regulate the transport of various ions that are vital to the cell. Since the ions have a charge, the transport can be observed as an electric current by clamping a pipette to the cell, which is called a patch. Ion channel recordings typically consist of time series of current at a certain sampling rate. The time series show noisy jumps between various current levels. The ion channel is believed to reside in various physical states and jumps between the states depend on the presence of various physical stimuli. The recordings were made by biologists and analysed by using statistical models. For the statistician the recordings may be viewed as secondary data. The goal of the analysis was to estimate the number of states and the transition probabilities between the states in time. In the analysis so-called hidden Markov models were used to describe jumps between states of the ion channels.

The LFS is a monthly survey that collects data on employment status, type of profession and educational level about people aged 15 years and older. The LFS is conducted in virtually all countries by national statistical institutes. The primary users are governments, but there is a wide variety of other, *ad hoc* users. The main goal is, however, the estimation of the percentage of the labour force population that is unemployed. This statistic is produced by means of so-called generalized regression estimators of the response to the survey.

I shall run briefly through the eight InfoQ-dimensions.

Data resolution

Data resolution plays an important role in ion channel recordings as it is the sampling rate and noise levels that determine the signal-to-noise ratio. Although the ion channel current is continuous, it is observed only at certain time points. The levels of noise in the experimental set-up may be high, obscuring and masking most of the visual signal. The models need to be extended to account for the time censoring and noise. For the LFS, data resolution has a less apparent meaning, but it may be viewed as the historical detail in

the data. Ideally, the full career of a respondent up to the time of interview should be recorded. However, respondent memory and interview duration limit the interview to a short time window.

Data structure
The ion channel recordings are a time series of electric current. They do not have missing data but the length of the series as well as the levels of noise vary greatly from one experiment to another. It must be expected that part of the recordings cannot be used owing to deteriorating clamps to the cell. To the statistician, it is unknown what the levels of noise are and to what part of the recording attention should be restricted. The LFS consists of closed questions with prescribed wording, answer categories and questionnaire routing. The wording and choice of the answer categories are very influential. The LFS may have up to 50% missing people from the sample. Much research is invested in the questionnaire design and the analysis and reduction of non-response.

Data integration
For the analysis of the ion channel recordings, this dimension does not seem applicable. For the LFS, data integration is a very important issue. LFS sample and respondent data must be linked to other government registry data. For this reason, it is imperative that respondents are identified over time.

Temporal relevance
In the ion channel recordings also the temporal relevance dimension does not play a role as behaviour of plant cells can be assumed to be stable in time. Clearly, for the LFS temporal relevance is important as statistics form the input to policies. Since it was recognized early on that memory effects may decrease the quality of data, the survey is conducted at a monthly frequency and in a panel setting. Still, secondary users may find the temporal relevance low as the reference period of the survey is very short. They must apply longitudinal analysis methods to account for censoring.

Chronology of data and goal
The statistical goals for the ion channel experiments were set after the data had been collected, but the design leaves little room for manipulation. For the LFS the goals are set by the sponsors. Secondary users must accept definitions and usually combine the LFS with other data.

Generalizability
The analysis of ion channel recordings is linked to certain cells in certain plants. This clearly limits the generalizability. The LFS attempts to provide detailed statistics about the full population and many subdomains of the population. As such it is designed to be general.

Construct operationalization
The current through an ion channel functions as a surrogate for the real behaviour of interest: the state in which a channel resides. This is not uncommon in biological experiments, but the operationalization requires models with strong assumptions about the relationship between observed current and underlying phenomena. It also requires modelling parts of the signal that are considered to be noise. For the LFS the definition of statistical concepts, the wording of questions and the classifications of variables are strongly regulated. An interview using a questionnaire itself demands a strong standardization and abstraction from the phenomena that it tries to measure.

Communication
The descriptive metadata of ion channel recordings are available in digital reports, but they lack information about the nature and magnitude of noise, and the quality of the patch clamp during the experiment. The LFS provides detailed reports on methodology and classifications. When the LFS is combined with other data sets, then again a detailed report is provided.

The dimensions data resolution, temporal relevance and communication were the easiest to interpret and to apply to the two examples. The dimensions generalizability and construct operationalization were easy to interpret but difficult to apply. They are strongly affected by the goal(s), which may not be that sharply defined. The dimensions data structure and data integration comprise multiple features and seem to have multiple subdimensions. The dimension chronology of data and goals was least meaningful to me.

The application of the eight InfoQ-dimensions to the two examples may not be that interesting to read by itself. However, the message that may be taken from the application is that the dimensions can be applied to greatly varying data. From the two applications I concluded that the eight InfoQ-dimensions do provide a structured look at the metadata.

Paul A. Smith and Jacqui Jones (*Office for National Statistics, Newport*) (Reproduced with the permission of the Controller of Her Majesty's Stationery Office and the Office for National Statistics)

Introduction

Kenett and Shmueli (KS) put forward their case for a quality framework for applied statistics, as a guide to the ability to achieve an analysis goal with an assemblage of data and a choice of analysis method. We are grateful to the Editors for this opportunity to contribute to the discussion of this paper. There is much of value in KS's train of thought—it is sometimes unclear what problem a study is attempting to solve; inappropriate analytical techniques are sometimes used, resulting in inappropriate conclusions; and, as KS say, many studies are used to solve problems for which they were never intended. So some way to evaluate the quality of the whole of an analysis, from the source of data to the inference, seems highly desirable.

KS formalize their case by proposing a composite concept of 'information quality' InfoQ, conceptually defined as a function of the analysis goal g, the available data X (we would have preferred 'observed data' as a term, since data may be available but unused), the analysis method f ('a technology') and a utility measure U.

Quality in official statistics

Our experience of quality assessment comes from official statistics, where there is general agreement on different dimensions of quality (e.g. accuracy, relevance, coherence—Statistics Canada (2002), Eurostat (2003) and Office for National Statistics (2007)). The purpose of these dimensions is to convey information to users about the usefulness of the official statistics for their particular analysis. Survey quality components are now relatively mature and, although there are variations on classifications of the components, there is substantial convergence. That is not to say that there are no alternative views, and the total survey error concept (Groves, 2004) provides a framework for expressing several aspects of survey quality in mean-squared error terms. It is nevertheless difficult to come up with a single composite measure for quality; combining the dimensions though visual displays of the different dimensions can help with interpretation, and Smith and Weir (2006) used principal components to examine the variation between some different survey quality measures. KS refer to these dimensions of survey quality but only in passing. They have not embedded this substantial body of literature in their approach. We shall consider KS's approach largely by comparison with the survey quality.

Information quality

KS set out a challenging objective—rather than providing quality information about some estimates for further use (e.g. in the official statistics case, further analysis or decision making) they seek to summarize the quality with respect to the final goal. Within the paper, though, they seem to provide multiple definitions for InfoQ:

(a) in Section 1.1, 'the extent to which the analysis goal is achieved' is measured by U;
(b) in Section 1.2 InfoQ is 'determined by the quality of its components', g, X, f and U;
(c) in Section 6 InfoQ has eight dimensions.

Definition (a) sounds like the problem solved in one step!—if we could indeed measure the extent to which the goal has been achieved by the utility. The problem here is to define and measure U appropriately, which is difficult or impossible in most situations.

Definition (b) has some appropriate elements. Data quality is certainly an important component (more of this later), and analysis quality is also important, though it has multiple aspects—choosing the best or most appropriate analysis from a range of options, applying the chosen method correctly and doing all the appropriate checking that assumptions that are important to the chosen method hold, for example. However, we struggled with 'quality of the goal definition' . . . the goal is what is needed, and in some cases it may be *appropriately* quite vaguely defined (for example in data mining it could be 'are there any interesting relationships between the available variables?'); maybe the *specificity* of the goal is a contributing factor, but how do you measure the *quality* of a goal?

Utility measures the extent to which the analysis goal is achieved, but this seems to depend on the type of goal—if as in Section 6.4 utility is defined as 'statistical significance to evaluate the effect of private or public reserve price', then taken at face value a low significance (α, the probability of a type I error) means high utility, and a high significance means low utility. But what if there is in fact no relationship? Then we expect α to be large, but the utility of the data and analysis for answering the goal (with the answer being 'There is no (significant) effect') may be high. What then is the *quality* of the utility function? If it says how well the utility function U measures how well the goal has been achieved, there will be a need for a

gold standard utility function U^* to compare it against. And if we had that we would not need the utility function U. So, although a concept of utility seems natural, it does not seem to be implementable over all types of goal in its current form.

Definition (c) is completely different, but it may be an attempt to define the components of (b) more precisely—Table 3 relates the dimensions to KS's earlier presentation of InfoQ as we deduce them, and to the survey quality framework. Dimensions in Sections 6.2.1–6.2.4 and 6.2.7 largely deal with data quality (sometimes evaluated in the context of the goal), and Sections 6.2.5 and 6.2.6 seem to cover some components of utility. Section 6.2.8 introduces a whole new layer around communication, which does not seem to us to affect the utility of achieving the goal unless the goal is to *communicate* the solution to a problem. But the dimensions do not seem to cover the analysis quality (whether the chosen f is a good choice from the set of possible $\{f\}$), or goal quality (if that is a well-defined concept). And not all components of InfoQ fit in this framework; in their discussion KS acknowledge that other dimensions may be defined, and there seems no counterpart to the concept of comparability in the survey quality framework (Eurostat, 2003). This undermines the approach, which should provide a framework in which all the important attributes of quality can be placed. Otherwise there is no standard and it is difficult to see why this is the right framework to use.

KS go further and attempt to summarize this information into a single variable, the 'InfoQ-score' using a subjective scoring approach, and a geometric mean of the scores. This is perhaps helpful, though it would be nice to see some assessment of what is lost in the summarizing and whether the resultant score conforms with expectations—in particular the similarity of dimensions 1 and 7 suggests that associated aspects of quality will receive more weight, and the lack of a dimension dealing explicitly with analysis quality suggests that this aspect will be underrepresented in the summary score.

In the rating-based assessment (Section 6.3) the scale is not defined; by comparison with values in Section 6.4, $5 \equiv$ very high (not high as in Section 6.3), $4 \equiv$ high and $3 \equiv$ acceptable. What do 2 and 1 represent? When this is 'normalized' (*rescaled* would be a better word, since no attempt is made to adjust for variability) to $\{1, 0.75, 0.5, 0.25, 0\}$, a geometric mean is an inappropriate summary, as, whenever a single attribute scores 1 (equal to 0 normalized), the geometric mean will be 0 regardless of the other scores. If a score of 1 means 'this is so bad that regardless of the other attributes the goal is unattainable' then just possibly this might be appropriate, but that seems a very extreme lower end to the scale. Otherwise, why should we bother to normalize? Aside from the zero problem, the geometric mean of the original scores is the same as the geometric mean of the normalized scores except for a constant scale factor.

So, we seek an unambiguous characterization of InfoQ—is it a multi-dimensional vector, or a scalar quantity and, if it is multi-dimensional, what are its elements—expressed in such a way that they span the required information?

Using InfoQ-components to improve studies

Sections 4 and 5 contain several ideas for using thinking about components of InfoQ to improve the effectiveness of a study at the design stage, and at the post-data-collection stage, in particular. There is plenty of potential for ensuring that a designed study is good for the questions that it is intended to answer, and we especially liked the comments about simulations in Section 4.4, as it is easy to assume that they do things that they do not really! The ideas in these sections are good guides for what to consider when doing a study, and we commend these comments.

We find it strange, however, that this discussion is restricted to the design and post-data-collection phases of a study; the data collection itself (as KS acknowledge in Section 7) is also critical in a study, and good practices here will increase InfoQ as surely as in the other phases of a study. There is also a long history of work on process quality from Deming onwards, and there would be a case for considering the quality of processes as an element of the quality also.

Conclusion

The ideas in KS are important, and the construction of a suitable framework in which they can be discussed, and where further work can be done, seems to be a valuable tool, perhaps especially in evidence-based policy making, where the quality of evidence to answer a particular question is not often discussed. Unfortunately, the InfoQ-framework does not meet this goal because it is not clearly defined and does not span the important measures—some are missing. To become a generally useful approach the concepts, definitions and rating method that are included in KS's approach need to be developed so that they are clearer and more easily applied in the wide range of situations which they ambitiously try to cover. Even with these enhancements, the application of their proposed approach would have limited value in relation

Table 3. Our interpretation of the relationships between the dimensions and components of InfoQ as presented by KS, and their nearest analogues from the survey quality literature (based on Eurostat (2003) and Office for National Statistics (2007))

Eight dimensions of InfoQ (Section 6.2)	*Quality of InfoQ-components (Section 1)*	*Comments*	*Survey quality—nearest analogue*
1. Data resolution—refers to the measurement scale and aggregation level of X	A property of the data X, so part of data quality	The relationship between aggregation and errors or noise is not clear—in the weekly stocking example it may be much easier to spot errors in daily data; whether a finer scale has more noise or less depends on the signal-to-noise ratio in the measurement method	*Accuracy*
2. Data structure—relates to types of data and data characteristics (e.g. corrupted and missing data)	A property of the data X, so part of data quality	This seems to cover several different characteristics—data type (e.g. cross-sectional or time series; numeric, nominal ordinal, qualitative), but also whether there are missing data, or other data problems (we are not clear what 'corrupt data' are in this context)	Types of data not covered; errors and missing data are in the non-sampling error component of *accuracy*
3. Data integration—using multiple sources	A property of the integrated data X, so part of data quality; Zhang (2012) gives a framework for quality of multiple-source data	Can generate errors, e.g. data matching and false matches	*Accuracy, coherence*
4. Temporal relevance—temporal gaps between data collection, data analysis and study deployment periods	An interaction between a property of the data X and the goal g affecting utility U	It would be nice to expand this slightly to cover *temporal focus*—whether the data refer to a precise time, or whether they have been collected or aggregated over a period	*Timeliness*
5. Chronology of data and goal	An element of the utility U	This seems to refer to whether the study is well designed for the goal, specifically in the time dimension	*Relevance?*
6. Generalizability—the utility of $f(X\|g)$ is dependent on the ability to generalize f to the appropriate population	An element of the utility U	This is a much wider question than 'what population does the sample represent'—more whether the achieved data and analysis are credible as an approximation to the target population and goal	*Relevance*

(*continued*)

Table 3 (*continued*)

Eight dimensions of InfoQ (Section 6.2)	*Quality of InfoQ-components (Section 1)*	*Comments*	*Survey quality—nearest analogue*
7. Construct operationalization	A determinant of the data quality	This seems very closely related to the measurement scale as defined in the first dimension; it would be better to have one dimension covering both aspects	*Accuracy*
8. Communication	None	This is a whole new concept, unless the goal is defined as to communicate the conclusion, rather than to make one; otherwise, although important, communication does not seem to affect the utility as defined here; we suggest dropping this 'dimension'	*Accessibility?*

to the production of official statistics. It may, as they note in the paper, have wider application in the assessment of academic and research grant statistical studies. The elements nevertheless provide a starting point for study design, and consideration of these issues at the design stage is likely to bring benefits for anyone attempting to answer particular questions.

(The views in this discussion are the authors' and do not necessarily represent those of the Office for National Statistics.)

Authors' reply

In this work we were motivated by the need for a comprehensive methodology for evaluating the quality of information that is generated by statistical research or projects. All three discussants have reinforced this perception and gracefully contributed to elucidate information quality further in general and InfoQ in particular. Our response begins with specific points that are related to the three discussions and concludes with some general insights that build on these points.

Dr Biemer addresses the aspect of *evaluating* InfoQ, on the basis of considerable experience regarding evaluating quality of survey studies by organizations such as Statistics Canada, Statistics Sweden and Eurostat. He describes lessons learned from Statistics Sweden's recent experience with evaluating 'total survey quality' which are indeed relevant in the context of InfoQ-evaluation. One challenge is the difficulty of constructing measures that capture the diverse aspects of concepts such as total survey quality and InfoQ, which are operationable and comprehensive. In particular, breaking down dimensions into subdimensions and even finer points makes evaluation more viable, although there is always the danger of not seeing the forest for the trees. Biemer's emphasis on the need to decompose the InfoQ-dimensions further is an important direction for further research.

The second point relates to *who* can be an effective evaluator. Biemer mentions the need for engaging external evaluators for reducing subjectivity while at the same time requiring evaluators to have good knowledge of the study and data analysis protocol. In the context of InfoQ, where the data, analysis and context are tightly coupled, it is critical that evaluators are familiar not only with the data analysis process (or potential data analysis approaches) but also with issues that are related to the data (collection, quality, etc.) and—very importantly—the larger context of the study, including the ultimate goals of the analysis, how its results will be communicated and used. The choice of evaluator also depends on who is carrying out the InfoQ-evaluation. For instance, InfoQ, like total survey quality, can be addressed from the point of institutions providing data (or summary statistics), such as central bureaus of statistics and/or of organizations consuming data, such as public services, businesses and industry. A related point that Biemer

makes is the importance of adequate documentation; in the InfoQ-context, this means documentation for each component: goal, data, analysis and utility.

Dr Biemer points to the usefulness of InfoQ as a standardizing evaluation mechanism that permits not only reproducing assessment results but also comparing studies as well as assessing changes in quality over time. Such comparisons (and reproducible assessments) are indeed the major motivation of our pursuit of InfoQ. In the paper, we alluded to using InfoQ for prioritizing and ranking studies as well as for assessing and improving research initiatives by considering InfoQ at different stages of the study.

Although InfoQ-evaluation plays an important role, we emphasize that the InfoQ-concept itself and awareness of InfoQ as a guiding concept are crucial for improving InfoQ: the awareness helps to guide the empirical analysis while maintaining a close focus on the study goal and context.

Professor Schouten provided a first-hand testimonial of how InfoQ is used and his experience adds to the experience that we have accumulated with students and colleagues over several dozen case-studies. His point about the clarity of research goals is very valid. In some studies, goals are explicitly formulated only late in the research, sometimes adapting the original goals to match what was actually done. As Schouten points out, awareness of InfoQ can help to elicit sharper research goals *ab initio* by considering InfoQ as early as the planning phase of a study. Regarding multiple data sets, this is accounted for in the data structure and data integration dimensions. When one deals with multiple data sets these dimensions can be non-trivial, as demonstrated by Professor Schouten's examples.

Describing InfoQ as being about 'the data that one desired to have' can generate an *InfoQ-benchmark* with which an existing data set can be compared. The ideal benchmark can then be used in the evaluation for each InfoQ-dimension. One important *caveat*, however, is that 'ideal' is subjective and limited by the imagination of the analyst (as well his or her preconceptions due to experience and knowledge). The era of 'big data' has taught us that real data create many opportunities that were not previously imagined. Nevertheless, the invocation of metadata is valid and thought provoking and is pointing to an excellent direction for further research into the formulation of InfoQ-dimensions. We also note his reference to 'big data analytics' and expand on it later in this rejoinder.

With respect to 'analysis quality', discussants Smith and Jones provide the clear explanation: 'the quality of analysis… [is] whether the chosen f is a good choice from the set of possible $\{f\}$'. Professor Schouten describes two studies which he examined through the InfoQ-dimensions. As we ourselves experienced, by going through this process many insights are revealed. His analysis illustrates the generality of the InfoQ-framework for any type of study that involves data analysis. However, such an analysis is typically absent from published studies, yet it sharpens and tightens the relationship between the data analysis and its context. The two examples also further illustrate the crucial importance of understanding not only the particular statistical goal but also the overarching study goal. For example, in the ion channels recording study, we must understand the purpose of estimating the number of states and transition probabilities. Similarly, in the Labour Force Survey, different stakeholders might use the unemployment estimates for different purposes, and therefore InfoQ would be different for each stakeholder. The dimension 'chronology of data and goals', which Professor Schouten found least meaningful, is a dimension that directly requires considering the overarching study goal.

Dr Smith and Dr Jones provided an important contribution by attempting to map InfoQ-dimensions to survey quality characteristics. In discussing various possible definitions of InfoQ they point to interesting issues that we dealt with in a straightforward way. The mathematical definition of InfoQ that we provide is $\text{InfoQ}(f, X, g) = U\{f(X|g)\}$. This corresponds to what InfoQ stands for. The eight InfoQ-dimensions represent how high InfoQ is achieved. In other words, our definitions of InfoQ is their definition labelled '(a)'. What they label as definition '(c)' corresponds to criteria for evaluating the level of InfoQ achieved. For clarity, let us address the issues of the InfoQ-concept and definition separately from InfoQ-evaluation. In terms of InfoQ-definition: Smith and Jones's point regarding utility measures being dependent on the type of goal is very well taken. This is exactly why we emphasize the need for goal elicitation so that we can clarify the utility of the analysis conducted on a certain data set. As to 'quality of the goal definition', this refers to the difficult step of mapping the overarching study goal(s) to the particular data analysis goals g. Professor Schouten pointed out the challenge of adequately defining the data analysis goal and the usefulness of InfoQ in eliciting this process early in the study. Finally, we agree that the InfoQ-dimensions are not set in stone. They are designed to provide guidelines for asking useful questions for evaluating InfoQ and for designing high infoQ studies, in a similar fashion to quality dimensions in surveys.

As to InfoQ-evaluation: Smith and Jones's comment about the application of geometric means to combined utility functions is appropriate. We referred to commonly applied techniques for scoring several

dimensions and achieve a unique scalar. With the expanded use of InfoQ we expect that more sophisticated multivariate techniques will be used, as suggested by the discussants.

The combined points made in these three discussions highlight several issues that are worth expanding. In what follows, we provide additional aspects of InfoQ along these lines.

InfoQ and 'big data' analytics

Big data analytics are gaining significant attention in various disciplines. The definition of big data analytics includes volume (memory size, number of transactions and number of records), velocity (batch, realtime and streams) and variety (structured, unstructured and semistructured). These three 'V's map to five of the InfoQ-dimensions (data resolution, data structure, data integration, temporal relevance and chronology of data and goal). Three additional InfoQ-dimensions (generalizability, construct operationalization and communication) are not covered in the big data analytics literature and, in this sense, InfoQ can be considered a valuable addition to this domain. For more on this topic see Russom (2011).

InfoQ and meta-analysis

A recent article in the *New York Times* (Chang, 2012) has brought up, once again, the issue of irreproducible research (Banks, 2011). The article describes studies relying on meta-analysis (i.e. secondary data) for comparing properties of organic and ordinary grown food, which reach contrasting conclusions. We suggest that considering InfoQ in these meta-analysis-based studies would have helped to structure the research approach and to avoid the confusion that is created by these parallel publications. For more on meta-analysis with implicit reference to InfoQ-dimensions see Negri (2012).

InfoQ and statistical education

From the three discussions we see the need to emphasize the pedagogical value of InfoQ. We suggest that InfoQ should be incorporated in the basic statistics curriculum in an effort to bridge the gap between what is typically taught in academia and how things look in practice (Kenett and Thyregod, 2006). At the graduate level, InfoQ can be applied to help to develop research proposals and to assess statistically based thesis and research studies (Kenett *et al.*, 2010).

The InfoQ-approach was designed to motivate researchers and practitioners to look more carefully into the influence of statistical work. The eight InfoQ-dimensions, with possible elaboration into more detailed subdimensions and expanded evaluation methods, can help to create such an impact. We thank discussants Biemer, Schouten, and Smith and Jones for their inputs and commentary which provide rich directions for further research in what one might call a main component of a theory of applied statistics (Kenett, 2012).

References in the comments

Banks, D. (2011) Reproducible research: a range of response. *Statist. Polit. Poly*, **2**, article 4.

Biemer, P. P., Trewin, D., Bergdahl, H., Japec, L. and Pettersson, Å. (2012) A tool for managing product quality. *Eur. Conf. Quality in Official Statistics, Athens*. (Available from `http://www.q2012.gr/articlefiles/sessions/3.2_Biemer_Bergdahl_Tool%20for%20Managing%20Product%20Quality.pdf`.)

Chang, K. (2012) Parsing of data led to mixed messages on organic food's value. *New York Times*. (Available from `http://www.nytimes.com/2012/10/16/science/stanford-organic-food-study-and-vagaries-of-meta-analyses.html`.)

Eurostat (2003) *Standard Quality Report*. Luxembourg: Eurostat. (Available from `http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/STANDARD_QUALITY_REPORT_0.pdf`.)

Eurostat (2009) *Handbook for Quality Reports*. Luxembourg: Eurostat. (Available from `http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/EHQR_FINAL.pdf`.)

Felme, S., Lyberg, L. and Olsson, L. (1976) *Data Quality Assurance* (in Swedish). Stockholm: Liber.

Groves, R. (2004) *Survey Errors and Survey Costs*. Hoboken: Wiley.

Kenett, R. S. (2012) A note on the theory of applied statistics. KPA, Raanana. (Available from `http://ssrn.com/abstract=2171179`.)

Kenett, R. S., Coleman, S. and Ograjenšek, I. (2010) On quality research: an application of InfoQ to the PhD research process. In *Proc. European Network for Business and Industrial Statistics 10th A. Conf., Business and Industrial Statistics, Antwerp, Sept. 12th–16th.*

Kenett, R. S. and Thyregod, P. (2006) Aspects of statistical consulting not taught by academia. *Statist. Neerland.*, **60**, 396–412.

Negri, E. (2012) Meta-analysis. In *Statistical Methods in Healthcare* (eds F. Faltin, R. S. Kenett and F. Ruggeri). Chichester: Wiley.

Office for National Statistics (2007) Guidelines for measuring statistical quality. Office for National Statistics, London. (Available from http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/measuring-statistical-quality-in-output-production-order.pdf.)

Russom, P. (2011) Big data analytics. *Best Practices Report.* Data Warehouse Institute.

Smith, P. and Weir, P. (2006) Characterisation of quality in sample surveys using principal components analysis. In *Statistical Data Editing*, vol. 3, *Impact on Data Quality*. New York: United Nations.

Statistics Canada (2002) Statistics Canada's quality assurance framework. Statistics Canada, Ottawa. (Available from http://unstats.un.org/unsd/dnss/docs-nqaf/Canada-12-586-x2002001-eng.pdf.)

Zhang, L.-C. (2012) Topics of statistical theory for register-based statistics and data integration. *Statist. Neerland.*, **66**, 41–63.