
**Journal of Official Statistics**
**vol. 30, i. 3 (2014)**

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 381–415, http://dx.doi.org/10.2478/JOS-2014-0022

# A System for Managing the Quality of Official Statistics

*Paul Biemer*[1], *Dennis Trewin*[2], *Heather Bergdahl*[3], *and Lilli Japec*[4]

This article describes a general framework for improving the quality of statistical programs in organizations that provide a continual flow of statistical products to users and stakeholders. The work stems from a 2011 mandate to Statistics Sweden issued by the Swedish Ministry of Finance to develop a system of quality indicators for tracking developments and changes in product quality and for achieving continual improvements in survey quality across a diverse set of key statistical products. We describe this system, apply it to a number of products at Statistics Sweden, and summarize key results and lessons learned. The implications of this work for monitoring and evaluating product quality in other statistical organizations are also discussed.

*Key words:* Total survey error; process control; GDP; quality indicators; statistical standards.

## 1. Introduction

Official statistics include the data and estimates that are published by national statistical offices (NSOs) and other public organizations on the major areas of society and the economy. They provide both quantitative and qualitative information on economic and social development, national productivity, living conditions, health, education, transportation, the environment, and many other areas of national interest. Credibility and confidence in the statistics depends to a large extent on the quality of official statistics. If the quality is suspect, the NSO's reputation as an independent, objective source of trustworthy information could be undermined. Therefore, managing the quality of statistical products is a key objective for all NSOs.

Quality is a vague concept that has become over-used in the literature and a more precise definition is required for the purposes of this article. Here, we define the quality of official statistics in terms of five dimensions that reflect their fitness for use by data users and other constituents. These dimensions, which will be described in more detail subsequently, are: Accuracy, Relevance/Contents, Timeliness & Punctuality, Comparability & Coherence, and Accessibility & Clarity. This article considers all five dimensions but primarily focuses on Accuracy or *data* quality which is considered fundamental to product quality. After providing a brief background for this work, the

[1] RTI International, P.O. Box 12194 Research Triangle Park, NC 27709-2194 North Carolina 27709, U.S.A. Email: ppb@rti.org
[2] Former Australian Statistician, Canberra, Australian Capital Territory, Australia. Email: dennistrewin@grapevine.net.au
[3] Statistics Sweden, SE-70189 Örebro, Sweden. Email: heather.bergdahl@scb.se
[4] Statistics Sweden, P.O. Box 24300, SE-10451 Stockholm, Sweden. Email: lilli.japec@scb.se

article considers a process for continually monitoring, evaluating, and improving quality over time across a diverse set of key data products.

NSOs world-wide are struggling to maintain high quality products as operating budgets continue to decline (see, for example, Struijs et al. 2013; Nealon and Gleaton 2013; Seyb et al. 2013). In fact, the March 2013 issue of the Journal of Official Statistics (JOS 2013) was devoted to cost-effective system architectures for producing high-quality statistics. In 2011, with guidance and support from Statistics Sweden we developed a structured, systematic approach for guiding the quality improvements in the agency's statistical programs and assessing the effects of these improvements on product quality. Referred to as ASPIRE (*A S*ystem for *P*roduct *I*mprovement, *R*eview, and *E*valuation), this approach provides a comprehensive framework for systematically evaluating all dimensions of quality with the primary focus on Accuracy. ASPIRE is quite general and can be applied in essentially any NSO or other statistical organization that supplies a continuous flow of statistical data to a community of users such as economists, researchers, government planners, and policy developers.

ASPIRE comprises an exhaustive inventory of potential risks to data quality for the products being reviewed and evaluates the organization's efforts to understand and mitigate these risks through evaluation studies and process improvements, assigning higher priorities where there are higher risks. The approach imposes a high standard of excellence on products based upon the best practices in the field while objectively and consistently rating products against well-specified quality standards or criteria. The ASPIRE framework provides an integrated approach to quality and risk whilst bringing rigour and heightened objectivity to assessments that might otherwise be based on subjectivity and intuition.

ASPIRE incorporates a number of unique features that may be considered new and innovative in the survey evaluation literature. First, ASPIRE goes beyond assessments that are based solely on compliance with statistical standards. Rather, it encourages continual improvements (both incremental and breakthrough improvements) in areas that represent the highest risks to data quality and thus motivates product excellence. Second, it provides numerical scores by error source, by criterion, and overall error sources and criteria that reflect product and process quality and that can be used for comparisons across time and products. Finally, ASPIRE provides a graphical presentation that can be readily understood by workers, managers, and administrators at all levels. It can communicate a general overview of quality simultaneously across numerous products or be used to "drill down" to view the evaluation details by product, by error source, by criterion level, or by any combinations of the three. Cost optimization is not the goal of ASPIRE; however, it does provides valuable information for cost-benefit analysis.

The first implementation of ASPIRE (referred to as Round 1) was conducted in 2011 for eight key statistical products at Statistics Sweden. This review provided a baseline for measuring improvements for these products in subsequent ASPIRE rounds. In 2012, Round 2 of ASPIRE was conducted for the same eight products while two additional products received an initial review. A third ASPIRE round on these ten products was completed in November 2013. This article presents the theory underlying the ASPIRE methodology, describes the process and its components, and mostly uses the experiences from Round 1 and 2 implementations to illustrate the application of ASPIRE. Further

refinements to the methodology were made in Round 3 but these were relatively minor in nature.

The next section provides an overview of the literature on quality of official statistics and lays the theoretical foundations for ASPIRE. Section 3 describes the ASPIRE approach in some detail including the basic criteria used in the evaluations, scoring system, and methods for ascertaining risks. Section 4 describes how ASPIRE was applied to a number of products at Statistics Sweden in 2011 and 2012 and summarizes some of the key results. Finally, we conclude the article with a discussion of the ASPIRE approach based upon our experience to date and plans for future implementations and evaluations of the methodology.

## 2. Total Quality

### *2.1. Product, Process and Organizational Quality*

NSOs and other statistical organizations have a long history of addressing various aspects of quality. The concept of quality has evolved over the years to become increasingly complex (Lyberg 2012). Today, we might view quality on three different levels, product, process and organization (Lyberg et al. 1998; Lyberg and Biemer 2008), each with its own set of assessment approaches. These quality levels can only be summarized here; however, Lyberg et al. (1998) describes them in some detail.

Product quality refers to the acceptability of a product (for example, an estimate of the unemployment rate) for its intended uses (for e.g., to monitor job loss/growth in the economy). Improvements in product quality are made by improving the processes generating the product. Thus process quality refers to the ability of survey processes to generate data and other statistical products of high quality. It is important that NSOs possess the knowledge, skills, and appropriate control systems to sustain and improve process quality. Organizational quality refers to the ability of the organization to consistently develop and maintain high quality processes. These three quality levels do not exist independently. Rather, organizational quality is required to achieve quality at the process level which is required for consistent product quality.

As an example, Statistics Sweden's Labour Force Survey (LFS) produces monthly estimates of the unemployment rate whose accuracy can be described in terms of error components that comprise the total mean squared error (MSE) of the estimate – an indicator of product data quality. Reductions in the MSE can only be achieved through process improvements such as more effective follow up of nonrespondents, improved interviewing, better estimation approaches, and so on (i.e., improved process quality). These improvements are possible because the organization possesses the knowledge, skills, and management structure to design and implement improved processes that result in real quality improvements.

The early literature on survey quality focused on product data quality (Accuracy) and the MSE as the primary indicator. Starting with the development of sampling theory in the 1930s and 1940s (Neyman 1934, 1938; Stephan 1948; Hansen et al. 1953) the focus obviously was on minimizing and controlling sampling errors. But it was also recognized early on that other error sources could affect the survey results – for example, the

interviewers and the nonrespondents (Deming 1944). In the 1960s, the importance of minimizing all error sources was stressed by some researchers; particularly, Dalenius (1967), Hansen et al. (1967) and Kish (1962). In order to estimate separate error components, evaluation studies were carried out, especially at the U.S. Census Bureau. Large evaluation studies, however, are expensive and of limited use for improving quality in real time because their findings may lag behind those of the main survey by many months. Standardizing and controlling processes that are known to affect product quality such as sampling, interviewing and coding, therefore became an important part of statistics production. The basic idea is that by continuously improving key survey processes, the overall process approaches an ideal state – that is, one that is stable and repeatable with minimal variation (Biemer and Lyberg 2003). A number of standards, guidelines and recommended practices have been developed over the years spanning from 1970 until today (U.S. Bureau of the Census 1974; Gonzales et al. 1975; U.S. Office of Management and Budget 2002; Eurostat 2005; International Standards Organization 2006; Statistics Canada 2009) all aiming at reducing errors and unnecessary variation. These efforts led to the so-called total survey error approach to survey design (Andersen et al. 1979).

In the late 1970s, the concept of survey quality was broadened via the so-called quality frameworks developed within the survey community (see Subsection 2.2), from encompassing not only Relevance and Accuracy, but also other dimensions of quality. In the 1990s, many survey organizations, influenced by the Total Quality Management (TQM) movement (Groves and Lyberg 2010), started to work on improvement projects. The importance of using process data (later named paradata; see Couper and Lyberg 2005) to evaluate and control process quality was stressed by Morganstein and Marker (1997). To view process quality as key to product quality was a new way of thinking in the survey community but in the private sector the concept of Six Sigma had already started to develop at Motorola in 1985. Also Deming's (1986) emphasis on statistical process control as a means for continuous improvement had large effects on how quality was perceived. Six Sigma (Breyfogle 2003) has become a toolbox for improvement projects, much like TQM, but with a strategic focus and a standardized method for process improvement and control. It turns out that it can also be very useful for improving survey processes.

Outside the survey community in the late 1980s and early 1990s, frameworks for evaluating organizations that strive for excellence were developed, for example, the Baldrige Performance Excellence Program (2013) and the European Foundation for Quality Management (EFQM 2013). These frameworks emphasize customer focus and results, and recognize the importance of leadership, people, partnership and strategy in order for an organization to reach excellence. Other important features of these frameworks are continuous improvement, which they share with Six Sigma and Kaizen, deployment of good practices and external evaluations. Some survey organizations such as the Czech Republic, Statistics Finland and Statistics Sweden have adopted one of these frameworks, namely EFQM.

In the auditing field the Committee of Sponsoring Organizations of the Treadway Commission (COSO) developed a framework to assess and improve internal control systems in the 1990s (COSO 2013) and later a framework to assess and improve enterprise risk management (COSO 2004). Both frameworks stress the importance of risks being

assessed in terms of likelihood and impact. The importance of risk assessment has so far been largely neglected in survey research (Eltinge et al. 2013).

Recently, Kenett and Shmueli (2014) developed a new framework for evaluating the quality of a generic statistical study that includes the dataset, the statistical analysis and the study report which they refer to as InfoQ or Information Quality. InfoQ provides a general framework applicable to data analysis in a broader sense than product quality. Rather it is "the potential of a dataset to achieve a specific goal using a given empirical analysis method." InfoQ framework identifies and examines relationships among the analytic objectives, the data available to achieve those objectives, the analysis of the data, and the ability of the analysis to achieve the objectives. Similar to the quality frameworks for official statistics, InfoQ provides eight dimensions used to deconstruct InfoQ as an approach for assessing it.

As Biemer (2014) notes, InfoQ can be regarded as a general framework that encompasses the survey total quality framework as a special case. Further, the development of InfoQ emphasizes the need for new practical tools for assessing quality in order to inform and caution data users regarding the limitations of a data analysis. In that regard, ASPIRE makes important contributions to data user knowledge and education about survey errors and their potential effects on statistical inference.

Thus, ASPIRE integrates many of the main ideas from the literature and frameworks mentioned above into a tool that will help product managers in survey organizations continually improve product quality. It does not rely solely on evaluations of MSE components for assessing Accuracy; yet it provides a practical, feasible approach to minimizing total survey error. In addition, the process facilitates the communication of quality improvements to stakeholders and users and greatly enhances an organization's ability to set clear goals for continual quality improvement.

As shown in the following, ASPIRE is not only applicable to surveys, but essentially any program that produce statistical products. By "statistical product" we mean virtually any data output that is used for statistical purposes including estimates, data sets, frames, registers, administrative databases, data tables, and indices. A major advantage of ASPIRE's generality in this regard is the consistency of the criteria, guidelines, ratings and definitions across the diverse assortment of statistical products found within NSOs.

### 2.2. *Dimensions of Product Quality*

To most statisticians and data analysts, good quality is synonymous with estimates having small mean squared errors (MSEs). The smaller the MSE, the more accurate are the estimates and the better are statistical inferences. As noted above, Deming (1944) recognized that quality should go beyond accurate estimates and should also encompass Relevance (Deming 1944). Over the years, the definition of quality has expanded to encompass other dimensions that are important to data users such as Timeliness, Comparability and Accessibility. This period also saw the development of so-called quality frameworks for official statistics whose use has expanded by new developments in survey methodology, technology and system architectures.

As an example, accessing data sets through the Internet is now common place and, for users, ease of access (i.e., Accessibility) is an important component of quality.

Decision-making in society has become more complex and global resulting in demands for harmonized and comparable statistics across countries and surveys (i.e., Comparability and Coherence). The timeliness of official statistics such as employment figures (i.e., the Timeliness dimension) often drives financial markets. Thus, quality frameworks for official statistics have been established to accommodate all these demands.

Several quality frameworks have been developed – each consisting of a number of quality dimensions. As an example, the quality framework developed by Eurostat (2009) consists of six dimensions: Relevance, Accuracy, Timeliness and Punctuality, Accessibility and Clarity, Comparability, and Coherence. This is essentially the framework adopted for the current report after combining the latter two dimensions into one dimension. Similar frameworks have been developed by, among others, Statistics Canada (Brackstone 1999), Statistics Sweden (Statistiska centralbyrån 2001), the UK Office for National Statistics (ONS 2007), the Organization for Economic Cooperation and Development (OECD 2011) and the International Monetary Fund (IMF 2003).

The work presented in this article emphasizes the Accuracy component of product quality. Biemer and Lyberg (2003) viewed accuracy as the dimension to be optimized in a survey while the other dimensions (the so-called *user dimensions*) can be treated as constraints during the design and implementation phases of production. They argued that sufficient Accuracy is essential for the other quality dimensions to be relevant. However, there are examples where accurate data may lose much of their utility if, for example, they are released too late to affect important decision-making or if they are presented in ways that are difficult for the user to access or interpret. As an example, surveys designed for the surveillance of disease outbreaks must be very timely if diseases are to be effectively contained. Accuracy may be secondary to timeliness in that case or there may be trade-offs involved where accuracy must be compromised to some extent for the sake of timeliness.

ASPIRE can help inform trade-offs among quality dimensions when assessments of these dimensions are incorporated into the evaluation framework. As discussed in Subsection 4.4, extensions of ASPIRE to include the user dimensions have been tested but more work is needed. However, this preliminary work was successful at identifying several important quality trade-offs and providing critical information needed for reconciling conflicting user and producer dimensions of quality.

### 2.3. *Accuracy*

For survey products, data accuracy is achieved by minimizing *total survey error* (TSE) which is the totality of error that can arise in the design, collection, processing, and analysis of survey data. (The term, TSE, could be generalized as "total *product* error" to acknowledge that ASPIRE's applications transcend survey products; however, we will use the traditional terminology in this article but note its limitations to describe some of the applications that follow.) A few error sources (such as measurement and data processing errors) are common to almost all surveys; however, other sources of error are dependent upon the survey design, type of data collected, and processing system used to develop the survey products. The ASPIRE system assesses accuracy by first decomposing the total error for a product into a number of error components that hold some appreciable risks to quality for the product. These risks are evaluated in the ASPIRE approach as well as

the steps that have been taken in the design and production stages to contain or mitigate these risks.

To identify the relevant error components, we let $\hat{Y}$ denote a survey estimate (or product) that is subject to errors from a number of sources. One can conceive of an "error-free" version of $\hat{Y}$ denoted by $Y$ which would result if the processes producing $\hat{Y}$ were error free including no sampling error (i.e., a complete census). Thus, the difference, i.e., $\hat{Y} - Y$, i.e., the total survey error, is due to all the errors in the processes that produce $\hat{Y}$, both sampling and nonsampling errors.

The ASPIRE model for surveys decomposes the total survey error into sampling error and seven nonsampling error components, viz., frame error, nonresponse, measurement error, data processing error, modelling/estimation error, revision error, and specification error. *Frame error* (denoted by $\varepsilon_{\text{frame}}$) arises in the process of constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. It includes the inclusion of non-population members (*overcoverage*), exclusions of population members (*undercoverage*), and duplication of population members, which is another type of overcoverage error. Frame error also includes errors in the auxiliary variables associated with the frame units (sometimes referred to as *content error*) as well as missing values for these variables. As examples, information on company size, industry, location, contact name, and address may be missing or erroneous for some enterprises on a business frame or register, thus potentially increasing costs and other errors (for example, sampling and modelling errors)

*Nonresponse error* ($\varepsilon_{\text{nonresponse}}$) encompasses both unit and item nonresponse. *Unit nonresponse* occurs when a sampled unit does not respond to any part of a questionnaire. *Item nonresponse* occurs when the questionnaire is only partially completed because an interview was prematurely terminated or some items that should have been answered were skipped or left blank. *Measurement error* ($\varepsilon_{\text{measurement}}$) includes errors arising from respondents, interviewers, imperfect survey questions and other factors which affect survey responses. *Data processing error* ($\varepsilon_{\text{data processing}}$) includes errors in editing, data entry, coding, computation of weights, and tabulation of the survey data. *Modelling/estimation error* ($\varepsilon_{\text{model/estimation}}$) combines the error arising from fitting models for various purposes such as imputation, derivation of new variables, adjusting data values or estimates to conform to benchmarks, and so on.

Preliminary estimates are published for some key statistics in order to address user needs for timely data. For example, quarterly GDP estimates based on preliminary data are published in order to provide government leaders and other important users with timely, albeit approximate, information on national economic performance. Preliminary estimates may be available one month after the end of a quarter; final estimates may be delayed until the end of the following year or later. Obviously, the utility of the preliminary estimates depends substantially on how close they are to the final, official estimates that are ultimately released. *Revision error* is the difference between a preliminary, published estimate and the final revised estimate and is an important component of the total error for some products.

To see why, let $\hat{Y}_P$ denote the preliminary, published estimate of the parameter $Y$ and let $\hat{Y}$ denote the final estimate. Then the total error in $\hat{Y}_P$ is given by $\hat{Y}_P - Y$ which can be rewritten as $(\hat{Y}_P - \hat{Y}) + (\hat{Y} - Y)$ where $\hat{Y}_P - \hat{Y}$ is the revision error and $\hat{Y} - Y$ is the total error in the final published estimate as described above. Because NSOs are quite interested in reducing the error in all published estimates, not just the revised ones, we focus on both

preliminary and revised estimates in our evaluation of Accuracy. Furthermore, considering revision error as a distinct error source reflects the view that large revisions, regardless of their reasons, are undesirable from the user's perspective and should be avoided. Thus, an important quality goal for any statistical agency is to reduce the size of the revisions which is facilitated by emphasizing revision error whenever it is applicable.

Note, however, that revision error is somewhat unusual because it reflects the combination of all other error sources on the preliminary estimate. For example, the preliminary estimate may differ from the final estimate as a result of late respondents (i.e., nonrespondents at the preliminary deadline) whose characteristics may be estimated or imputed in the preliminary estimate while their reported values are used in the final estimate. Likewise, revisions may correct for other nonsampling errors such as measurement, data processing, or modelling/estimation errors that are identified after the preliminary deadline. In this way, revision error may account for error sources that have already been considered in the assessment of data quality for the revised estimate.

For this review, our primary interest with regard to revision error is on the magnitude of the error – that is, the difference $\hat{Y}_P - \hat{Y}-$ and the steps that could be taken to reduce it and/or its impact on data users. As such, we have not decomposed revision error into its associated subcomponents (nonresponse error, data processing errors, etc.) because these error sources are considered in great detail in the evaluation of the final estimates. Nevertheless, separately decomposing revision error may still be very important in some cases to understand the impact of error sources on revision error that may be distinct from those affecting the final estimates.

For most products, a seventh nonsampling error source – referred to as *specification error* – is also applicable. Specification error arises when the observed variable, $y$, differs from the desired construct, $x$ – that is, the construct that data analysts and other users prefer. In survey literature, for example Biemer (2011), $x$ is often referred to as a *latent* variable representing the true, unobservable variable and $y$ is often referred to as an indicator of $x$. As an example, in the European statistics for Foreign Trade of Goods (FTG), the invoice value of goods is collected from enterprises ($y$) while the statistical value ($x$) (i.e., the cost of goods at the border of the reporting country excluding costs incurred after crossing the border) is preferred for most statistical uses of the data. Thus, specification error may be defined as the difference between $y$ and $x$ (see, for example, Biemer and Lyberg 2003).

Specification error biases the estimates of population parameters. Let $X$ denote the true population parameter which is a function of $x$. Then the total survey error (TSE) in a preliminary estimate can be written as

$$\hat{Y}_P - X = (\hat{Y}_P - \hat{Y}) + (\hat{Y} - Y) + (Y - X) \tag{1}$$

where $(\hat{Y}_P - \hat{Y})$ is the revision error, $(\hat{Y} - Y)$ is a combination of errors from multiple sources; specifically $\hat{Y} - Y = \varepsilon_{\text{sampling}} + \varepsilon_{\text{frame}} + \varepsilon_{\text{nonresponse}} + \varepsilon_{\text{measurement}} + \varepsilon_{\text{data processing}} + \varepsilon_{\text{model/estimation}}$, and $(Y - X)$ is the specification error. Likewise, the TSE in the final estimate, $\hat{Y}$, is just the right side of (1) with the revision error term omitted.

Under this model, the total survey error of an estimate includes specification error as well as the other aforementioned sampling and nonsampling errors. Thus, the specification error in the aggregate, $\hat{Y}$, is essentially the difference between the expected value of $\hat{Y}$

conditioned on the concepts implied by the survey instrument ($Y$) and the population parameter under the preferred or true concept ($X$). Some would argue that specification error should be part of the Relevance/Contents dimension. However, our view is that it is part of total survey error and, thus, should be considered a component of Accuracy.

## 3. The ASPIRE Model

The ASPIRE model borrows heavily from the quality assurance literature (see, for example, Juran and Godfrey 1999 and Breyfogle 2003) whose core principle rests on the identification, reduction, and elimination of suboptimal processes as well as the literature on continual improvement or Kaizen (Imai 1986). As a corollary to this principle, Lyberg et al. (1998) argue that improvements in survey processes aimed at reducing error risks (i.e., the probability that important errors will occur) will often produce products with reduced error to the extent that the risks are actually reduced. As an example, data collection processes designed using best practices and state of the art knowledge can achieve lower risks of measurement error and nonresponse, particularly if these processes are routinely monitored for compliance with the design specifications. While continual process improvement is often desirable, it may not always lead to product improvements. For example, some methods for increasing response rates (such as incentives) can actually lead to an increase in the nonresponse bias (see, for example, Keeter et al. 2000; Curtin et al. 2000; Merkle and Edelman 2002).

Thus, an essential ingredient of process improvement is to conduct experiments that directly measure the effects of alternative designs and processes on one or more components of the total error. Such experiments can provide quantitative evidence that the processes implemented actually reduce the errors from the targeted error source compared to the tested alternatives. As an example, the estimation of bias has been used effectively for comparing modes of data collection, alternative incentives, questionnaire design alternatives, and so on. However, this approach may be impracticable for TSE reduction across dozens of surveys generating hundreds of statistical products. It may not even be feasible for a single survey given the many potential sources of error whose effects may interact and vary considerably over the many estimates (products) generated by the survey.

Often the final survey design is a compromise that balances the TSE across many competing objectives; for any particular objective, it may be suboptimal. This "compromised design" phenomenon is not unique to surveys; rather it arises quite often in industrial quality control as well (see, for example, Michalek et al. 2006; Karsak 2004.) Given these complexities, the process improvement principles embodied in ASPIRE provide a feasible and effective approach for achieving product quality improvements across the wide range of products produced by the typical NSO.

ASPIRE is a system for assessing the risks of error from each potential source of error in a product and rating progress that has been made to reduce this risks according to clearly specified evaluation criteria. Its primary goals are to:

(a) identify the current, most important threats or risks to the quality of a product,
(b) apply a structured, comprehensive approach for rating the efforts aimed at reducing these risks, and

(c) identify areas where future efforts are needed to continually improve process and product quality focussing on those high risk error sources where ratings are relatively low.

We believe that product quality will improve to the extent that ASPIRE achieves these three goals. ASPIRE is quite general in that it can be applied to a specific statistical estimate such as the monthly unemployment rate, a range of products produced by a data collection program such as the estimates from a survey of local government agencies, or a frame or register such as the business register or master address frame, or a compilation of a number of statistical inputs such as estimates of gross domestic product (GDP). ASPIRE is also comprehensive in that it considers the errors in official statistics arising from all major error sources from the design of the data collection to final publication or data release.

The ASPIRE model assesses product quality by first decomposing the total error for a product into major error components. It then evaluates the potential (or risks) for these error sources to affect data quality (referred to as "the risks of poor quality") according to five evaluation criteria. Clearly specified and sufficiently detailed guidelines have been developed that are used to evaluate the risks with acceptable inter-rater reliability.

As previously noted, ASPIRE can be customized so that it considers only those error sources that pertain to a specific statistical product. For example, sampling error would not apply to products from the Swedish municipal accounts collection (referred to as RS) which does not employ sampling. As discussed in the next section, the model also accommodates the risk variation across error sources so that a product's overall quality is affected more by the error sources that pose greater error risks. For example, in the RS, revision error was judged as "low risk" because preliminary and final data releases seldom differ appreciably. Moreover, RS data users claim they are seldom affected by such revisions. On the other hand, data processing error is of high risk in the RS due to the amount of editing data receive and the potential for editing error to substantially affect the final estimates.

### *3.1. Assessing Error Risks*

A critical element of the ASPIRE rating system is the assessment of error risk which involves assigning a risk rating to each error source according to its potential impact on product quality. For this purpose, it is important to distinguish between two types of risk: *residual* (or "current") risk and *inherent* (or "potential") risk. *Residual risk* reflects the likelihood that the survey process will produce a serious, impactful error *despite* the current efforts that are in place to reduce or mitigate the risk. *Inherent* risk is the likelihood of such an error *in the absence of* current efforts toward risk mitigation. In other words, inherent risk reflects the expected impact of errors from the error source if efforts to maintain current, residual error were suspended.

As an example, for a survey process that places a high burden on respondents (e.g., lengthy interview or complex data collection protocol), the risk of nonresponse and thus, nonresponse error may be considered inherently high. However, these error risks can be reduced by various data collection strategies such as multiple follow-up attempts, incentives, enhanced interviewer training on techniques for averting refusals, and so on.

Postsurvey adjustments may further reduce the risk of nonresponse bias. Thus, although inherent risk for the survey process is high, the residual may be moderate or low.

One may view the inherent risk rating for an error source as an indicator of the need for measures to control the errors from that source in the process. The greater the inherent risk the greater the need for approaches that will reduce it. The residual risk rating may be regarded as an indicator of the effectiveness of these measures to limit the error from a specific source. It therefore follows that inherent risks should be stable over time. Changes in the survey taking environment that alter the potential for error in the absence of risk mitigation can alter inherent risks, but such environmental changes occur infrequently and usually evolve gradually. On the other hand, residual risks are more transient as they depend upon risk mitigation activities which can change over time or may become less effective. As an example, nonresponse rates may increase over time as contact and refusal aversion strategies that were once effective become less so, thus increasing the residual risk of nonresponse error.

There are some similarities with the ASPIRE approach and those outlined in the program evaluation and risk management literature. Program evaluation consists of methods for collecting and analyzing data in order to address questions about the effectiveness and efficiency of projects, policies and programs (Rossi et al. 2004); for example, an evaluation of the effectiveness of establishing community health centers in low income areas at reducing the need of long hospital stays or expensive emergency room use. Consistent with most program evaluation systems (see McDavid et al. 2013), there is an underlying model and methodology and a performance management system. However, program evaluations often rely on experiments or quasi-experiments that compare the program outcomes with counterfactual outcomes – designs that seldom arise with NSO product evaluations. With respect to risk management, the literature uses the concepts of intrinsic and residual risks, usually uses templates to support the risk analysis, values risks in terms of both impact and likelihood, and relies on a range of risk assessment tools (see Barkley 2004; International Standards Organization 2009). Notwithstanding these commonalities, ASPIRE is the only system to incorporate a total error framework while still remaining accessible to NSO executives who may have very limited knowledge of the complex programs being evaluated.

As shown in the next section, the inherent risk for an error source directly affects a product's overall score because it determines the weight attributed to an error source in computing a product's average rating. While residual risk does not directly affect a product's score, it still plays an important role in the evaluation in two ways. An increase in residual risk from the prior evaluation could suggest that efforts to reduce the inherent risks of error have become less effective. Thus, the product's rating relative to risk mitigation would deteriorate accordingly. In addition, residual risk helps clarify the meaning and facilitate the assessment of inherent risk.

### *3.2. Evaluation Criteria*

In addition to decomposing total error for a product into its component sources and identification of the risks associated with each source, the ASPIRE model evaluates the potential for these error sources to affect data quality according to five evaluation criteria,

viz., Knowledge of Risks, Communication with Users, Available Expertise, Compliance with Standards and Best Practices, and Achievement Towards Risk Mitigation or Improvement Plans. In Round 3, Communication with Users was extended to include data suppliers or providers as well as users. (For example, in the case of the National Accounts, these include departments responsible for key inputs to the GDP calculations such as the foreign trade and business statistics units.) The five criteria are given equal weight; however, differential weights could be used if desired. The guidelines currently used for evaluating these five criteria are shown in Appendix A.

A two-step rating process was used to assign ratings on a 10-point scale for each error source by criterion combination. First, a given criterion is assigned a qualitative rating of Poor, Fair, Good, Very Good, and Excellent based on the check list and subsequent discussions with the product area. Then, in step two, these qualitative ratings are then refined by choosing between low or high numerical point ratings within each of the five categories; for example, Poor (1 or 2), Fair (3 or 4), and so on to complete the 10-point scale. This is further described in the subsequent illustration.

A product's *error-source score* is the sum of its ratings (on a scale of 1 to 10) for the error source across the five criteria divided by the highest possible score attainable (which is 50 for most products) and then expressed as a percentage. A product's overall score, also expressed as a percentage, is then computed by the following formula:

$$\text{Overall Score} = \sum_{\text{all error sources}} \frac{(\text{error-source score}) \times (\text{error-source weight})}{10 \times (\text{number of criteria}) \times (\text{weight sum})} \tag{2}$$

where the "error-source weight" is either 1, 2, or 3 corresponding to an assessment of the source's inherent risk – 1 if low risk, 2 if moderate risk, and 3 if high risk – and "weight sum" is the sum of these "risk" weights over the product's applicable error sources.

The form of the overall score is somewhat arbitrary and other metrics could be used to summarize a product's overall rating. For example, as previously noted, it is possible to weight the five criteria differentially to reflect their relative importance. In addition, Kenett and Shmueli (2014) suggest a metric based upon the weighted geometric mean of scores which also has some desirable properties. Nonetheless, the current metric is intuitive while still providing a useful way to rank and compare products.

## 4. The Statistics Sweden Experience

As noted above, ASPIRE has been applied to seven key products at Statistics Sweden for three consecutive years (or rounds) and three products for the last two rounds. The quarterly and annual national accounts were considered together in the first round and then considered separately in the last two rounds. Table 1 lists the products and the error sources that were considered in the review for each. These products were considered "key" regarding their importance to the Swedish statistical system. In addition, together they span the breadth of statistical products offered by Statistics Sweden including; business and social surveys, registers, indices, and compilations. As shown in the table, eight products received an initial review in 2011(i.e., Round 1) and a second, follow up review in 2012 (Round 2) although quarterly and annual national accounts were considered separately in the second round. One product received its initial review in 2012. All ten

*Table 1. Products and Error Sources Evaluated in Rounds 1, 2, and 3*

| Product | Round | Error Sources |
|---|---|---|
| *Survey Products* | | Specification error |
| Foreign Trade of Goods (FTG) | 1,2,3 | Frame error |
| Labour Force Survey (LFS) | 1,2,3 | Nonresponse error |
| Annual Municipal Accounts (RS) | 1,2,3 | Measurement error |
| Structural Business Statistics (SBS) | 1,2,3 | Data processing error |
| Consumer Price Index (CPI) | 1,2,3 | Sampling error |
| Living Conditions Survey (ULF/SILC) | 2,3 | Model/estimation error |
| | | Revision error |
| *Registers* | | Specification error |
| Business Register (BR) | 1,2,3 | Frame: Overcoverage |
| Total Population Register (TPR) | 1,2,3 | Undercoverage |
| | | Duplication |
| | | Missing data |
| | | Content error |
| *Compilations* | | Input data error (up to four sources) |
| GDP | 1* | Compilation error |
| GDP by Production Approach | 2*,3* | Data Processing error |
| Annual | 2*,3* | Model/Estimation error |
| Quarterly | | Deflation/Reflation error |
| | | Balancing error |
| | | Revision error |

* Error sources were modified in Rounds 2 and 3 based upon the error model in Figure 1.

products were reviewed for a third time in November 2013 (Round 3). This section describes some key aspects of these reviews and reports on some of the key findings.

### *4.1. Implementing ASPIRE*

#### 4.1.1. Forming the Evaluation Team

A key issue in forming a program evaluation team is whether to use internal or external evaluators. As summarized in Conley-Tyler (2005), there are important advantages of each approach. Internal evaluators provide some costs advantages and may excel in their intimate knowledge of the specific products and processes to be evaluated. In addition, whereas highly capable external evaluators may be scarce, internal evaluators having high levels of program-specific expertise may be readily available. With regard to costs, Statistics Sweden's experience suggest that cost savings using internal evaluators would be small or nil for broad-based evaluations like ASPIRE once the labor costs devoted to maintaining consistency of ratings across multiple evaluations teams are considered.

On the other hand, external evaluators generally have greater "perceived objectivity" if not greater "real" objectivity – key issues for NSOs intending to make the evaluation results public. Conley-Tyler (2005) notes that external evaluators are more objective and willing to criticise processes, management, and the organization itself. In support of this

claim, Statistics Sweden's prior experiences using internal evaluators engaged in similar activities raised concerns about the objectivity of that approach.

With respect to relevant knowledge of the TSE paradigm, the expertise of external evaluators may be broader and their experiences of having worked in other organisations provide benchmarks for judging quality. Likewise, their knowledge of the total error in official statistics may be greater than that of the internal evaluators. Thus, another advantage to using a small group of external evaluators having broad knowledge to conduct all the evaluations is greater consistency in the ratings across the products.

The advantages of using external evaluators are even stronger for government programs where transparency and objectivity are critical. While transparent evaluation can be achieved by both internal and external evaluators, credibility and legitimation is much greater with external evaluators (Conley-Tyler 2005), particularly if they are recognized experts in both the TSE paradigm and in the functioning of the NSO's statistical programs. This could be the deciding factor for NSOs and other organisations receiving government funding.

In the end, Statistics Sweden opted for external evaluators (Biemer and Trewin) who were aided by two management liaisons (Bergdahl and Japec) who provided internal program context and support for the evaluation.

For each round of ASPIRE, three sets of activities were conducted which may be described as preinterview, interview, and postinterview activities.

#### 4.1.2. Preinterview Activities

a. *Background Reading and Preparation.* Several weeks prior to the onsite evaluation, each of the two external evaluators received an extensive set of materials for each of the products. Central among these was the "quality declaration" (if available) for each product. The quality declaration is a type of quality profile (Biemer and Lyberg, 2003) that documents key aspects of the design, data collection and production process for the product including the major error sources and what is currently known about them, descriptions of previous, current, or planned quality studies, and relevant information related to the user quality dimensions. Questionnaires, training manuals, and reports on recent studies related to quality were also included in the reading materials.
b. *Self-evaluations by Product Teams*. Also during this period, each product team was asked to complete a self-evaluation form that reflects the guidelines the external evaluators used to complete their initial evaluation of the product. In Rounds 2 and 3, the self-evaluations used the checklist format shown in Appendix B.

#### 4.1.3. Quality Interview

A face to face interview lasting about four hours was conducted by the external evaluators with each product team. One important purpose of this interview was to supplement and clarify the information provided in background reading materials and self-evaluations. During these discussions, inherent and residual risks levels (high, medium, and low) were assigned to each applicable error source. Once the risk levels were established, the evaluators separately considered each applicable error source to assign a rating for

each criterion using a simple five-point scale: poor, fair, good, very good, and excellent. At the conclusion of interview, the risk levels and criteria ratings were reviewed and further discussed. Any disputes were clarified and reconciled to the extent possible. Detailed minutes were kept to provide a record of the proceedings. Of particular importance, these minutes captured justifications for the ratings by error source and criterion.

#### 4.1.4. Postinterview Activities

Within a day or two following each interview, the evaluators reviewed the minutes, refined the ratings and resolved any inter-rater discrepancies. Apparent rating inconsistencies within and across products were identified and removed. These ratings and their written justifications were then shared with the product teams who were asked to correct any inaccurate or misleading information and dispute ratings they believe were not justified. This process yields the final ratings and justification narratives. These ratings constitute a major portion of the final report authored by the external evaluators.

Following Round 1, ASPIRE was improved in the following ways:

1. A number of enhancements were made to the rating process. Chief among these was the development of a criterion checklist that could be applied generically across the applicable error sources and products. Items in the checklist were sorted so that the criterion's rating usually followed directly from the last item affirmatively checked. The simple "yes/no" format eliminated much of the subjectivity in the self-evaluation process observed in Round 1. Appendix B shows one such checklist (for Knowledge of Risks).
2. Except for new products, the quality review focused on changes in knowledge, staffing, methodology, processing, planning, mitigation strategies, etc. that may have some implications for data quality. This emphasis reflects the goal of the second and third rounds which are to assess the changes in quality since Round 1.
3. Post-interview, face to face, debriefing meetings were held with product teams that wanted to appeal one of more of their ratings and/or discuss the written rating justifications and recommendations for improvement.
4. In the second round, user dimensions were also evaluated for two products (the Labour Force Survey and the Consumer Price Index) as described in Subsection 4.3.
5. The error sources used in Round 1 for the GDP were substantially revised following in-depth discussions with the National Accounts staff about the GDP production process. This necessitated revamping the criteria used to evaluate GDP data quality. Details regarding this approach are provided in Subsection 4.2.

#### 4.1.5. Illustration – Foreign Trade of Goods (FTG)

In this section, we illustrate how the steps of the process were executed for Statistics Sweden's survey of international trade or the FTG. The FTG collects information on the imports and exports of 9,000 different types of commodities by country of origin and destination for 250 countries resulting in almost two million statistical items being reported each month. The primary uses of the results of the survey include the trade in commodity components of the balance of payments statistics and the expenditure measure of GDP. It consists of two statistical systems: Extrastat (for countries outside the EU) and Intrastat (for EU countries).

In Round 1, measurement error was classified as high inherent risk for the FTG for a number of reasons including possible misclassification of commodities (more so for responses via paper forms than for electronic responses), data concerns regarding net weight (and other quantities) of shipments especially for textiles and chemicals, and errors resulting from the methods used to convert the invoice value to conceptually correct statistical value. At the other extreme, revision error was deemed to be low risk because the size of revisions tended to be relatively small and inconsequential to most users. The other error sources were given medium risk.

In Round 2, these risk ratings were revised based upon further discussions with internal data users such as the National Accounts staff. In particular, revision error was upgraded to high inherent risk after the potential effects of revision error on the GDP estimates were better understood. Likewise, data processing error was raised to high inherent risk after realising the extensive editing that is done in the FTG and the risk it poses to data quality without this editing. Frame error was downgraded to low risk when it was determined that the risk of overcoverage in the FTG frames (viz., the Business Register and National Tax Board VAT register) is much lower than originally thought. Theoretically, changes to inherent risks should only occur when (a) the design of a process undergoes a fundamental change; for example, rather than collecting EU export data directly from enterprises, exports are based upon imports from other EU countries or, (b) as in the case of FTG, the information upon which the current risk level was based is deemed incomplete or erroneous and, thus, the inherent risk level for the product should be corrected.

Note that sampling error is not applicable for the FTG because it employs a cut-off sample that includes all enterprises above a threshold value representing at least 95% of all imports and exports within the EU but there will be modelling error because certain assumptions are made to estimate the contribution of those enterprises below the threshold value.

With regard to quality ratings, processing error received the lowest score which in part reflects the FTG personnel's lack of knowledge at that time about the causes and extent of editing errors which have a high risk of error. In addition, the evaluators had concerns that lack of quality control in the keying of paper forms was a violation of ISO standards. In fact, the number of paper forms that are keyed was quite small (about 10% of all reports) which diminishes any risk of error from this source. Nevertheless, the paper transactions could comprise a sizeable percentage of trade for some commodities and pose an appreciable error risk in those situations.

Notwithstanding these concerns, FTG's overall quality score was among the highest in Round 1. Nonetheless, its rating for measurement error was fairly low and the evaluators provided several recommendations and strong encouragement to take initiatives that would increase that score in the coming year. The evaluators' guidance was apparently followed because important improvements to address measurement error were quite evident in Round 2. For example, communication with data users regarding accuracy, particularly measurement error, substantially improved as a result of enhancements to the quality declaration. In addition, several important studies were completed and documented in reports providing more information on measurement and other error sources.

In addition to these improvements, other quality improvements were made as follows:

- Swedish Customs adopted the FTG editing system for its programs improving the quality of data received by Statistics Sweden.
- Plans are in place to better understand the causes of revision error, its impact on important users such as the National Accounts, and some effective means for reducing it over time.
- An asymmetry study with Finland (i.e., a reconciliation of Swedish imports against Finnish exports and vice versa) was completed which focused on understanding the effects of coding error on trade statistics.
- Work has commenced to replace the current Excel-based macro-editing software with a much improved, flexible and professionally developed system.
- Use of the Statistics Sweden's "Standardized Methods and Toolbox" increased resulting in a number of improved practices.
- A new survey to calibrate the conversion of invoice value to statistical value was scheduled for completion (and, subsequently completed) in 2013.

The current and previous round's ratings are shown in Table 2 in graphical form and the changes are shown in Table 3. Similar tables were developed for Round 3 so that improvements over successive rounds could be shown.

### 4.2. *Error Sources Specific to the Gross Domestic Product*

In retrospect, the Round 1 evaluation of the GDP error was somewhat flawed as a result of attempting to force an error structure identical to that used for the surveys. The eight error sources that are applicable to other products cannot easily be applied to GDP considering its unique, extensive and complex error structure. Thus, in Round 2, ASPIRE was modified by tailoring it to more closely reflect the complex GDP error structure. Because of the time constraints, the focus of the Round 2 review was considerably narrower, focusing solely on the estimation of quarterly and annual GDP using the production approach. In addition, the error structure of the GDP estimation process was restructured to more precisely capture the GDP's major error sources. The same approach and error structure can be used as well for GDP compiled from the expenditure approach.

Figure 1 provides a flow diagram that attempts to capture the major activities associated with the estimation of GDP. As shown, the GDP estimation process incorporates two somewhat independent ways for estimating GDP. These are referred to as the production (shown on the left) and the expenditure approaches (shown on the right). Both approaches begin with a number of inputs that must be assembled, processed, and compiled to prepare them for the next step in the process. Each of these inputs is subject to error. The "Compile" stage includes data processing, which may be simply entering the inputs into an Excel spread sheet but may also include some editing as well as modelling/estimation especially when only proxy variables are available. This latter process may involve combining multiple inputs to create derived variables as well as modelling the data to reduce specification and other errors. For producing GDP in current prices, these compiled inputs proceed through an estimation stage which, for the production approach, involves

*Table 2. FTG Quality Ratings Matrix for Round 2 with Round 1 vs. Round 2 Scores by Error Source*

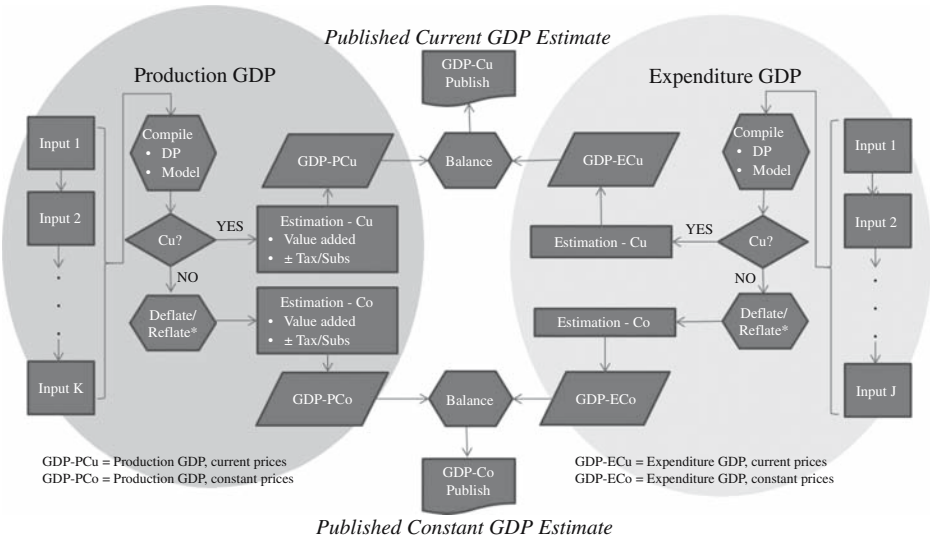
| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with Standards & Best Practices | Plan Towards Mitigation of Risks | Risk to Data Quality |
|---|---|---|---|---|---|---|---|---|
| Specification error | 58 | 58 | ○ | ○ | ◡ | ◡ | ○ | M |
| Frame error | 58 | 58 | ○ | ○ | ◡ | ○ | ◡ | L |
| Non-response error | 62 | 66 | ◡ | ◡ | ◡ | ○ | ◡ | M |
| Measurement error | 54 | 62 | ◡ | ○ | ◡ | ◡ | ○ | H |
| Data processing | 46 | 60 | ◡ | ◡ | ◡ | ◓ | ○ | H |
| Sampling error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Model/estimation | 66 | 80 | ◡ | ◡ | ◎ | ◎ | ◡ | M |
| Revision error | 62 | 76 | ◡ | ◡ | ◡ | ◎ | ◡ | H |
| Total score | 57,3 | 65,8 | | | | | | |

| Scores | | | | | Levels of Risk | | | Changes from round 1 | |
|---|---|---|---|---|---|---|---|---|---|
| ● | ◓ | ○ | ◡ | ◎ | H | M | L | | |
| Poor | Fair | Good | Very good | Excellent | High | Medium | Low | Improvements | Deteriorations |

*Table 3. FTG Round 1 to Round 2 Rating Changes and Corrections with Annotations*

| Error source | Score round 1 | Score round 2 | Knowledge of Risks | Communication to Users | Available Expertise | Compliance with Standards & Best Practices | Plans Towards Mitigation of Risks | Risk to Data Quality | *Correction from 2011 rating* / *Improvement from 2011 rating* Comments on changes |
|---|---|---|---|---|---|---|---|---|---|
| Specification error | 58 | 58 | 5 | ~~7~~5[1] | 7 | 7 | 5 | M | [1]Under the current guidelines, communication should have been "Good" last year, not "Very Good." |
| Frame error | 58 | 58 | ~~7~~5[1] | 5 | 7 | 5 | 7 | ~~M~~L[2] | [1]Corrects error in last years rating for Knowledge of Risks.<br>[2]Also, corrects risk level based upon intrinsic risk of frame error being low. |
| Non-response error | 62 | 66 | 7 | 5→7[1] | 7 | 5 | 7 | M | [1]Communication to users about nonresponse improved as a result of the QD. |
| Measurement error | 54 | 62 | 5→7[1] | 5 | 5→7[2] | 7 | 5 | H | [1]Knowledge of risks gained through writing the QD as well as preparation of the annexes to the SLA with the NA.<br>[2]Working relationship and closer cooperation between the col lection unit and the methods group as a result of the SLA. |
| Data processing error | 46 | 60 | 5→7[1] | 5→7[2] | 5→7[3] | 3 | 5→6[4] | ~~M~~H[5] | [1]Knowledge of risks gained through writing the QD as well as preparation of the documents "Improvements of the work on revisions in the Swedish good" and "Improving macro-editing in Intrastat."<br>[2]Likewise Communication has improved through both of the above mechanisms.<br>[3]Working relationship and closer cooperation between the col lection unit and the methods group as a result of the SLA.<br>[4]Some planning is underway for further improvements of editing and coding. Planning and discussions are underway to reduce the misclassification of goods by enterprises.<br>[5]Risk level was re-evaluated and elevated to H based upon the importance of editing to data quality. |
| Sampling error | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | |
| Model/estimation error | 66 | 80 | 7→8[1] | 5→7[2] | 7→9[3] | 7→9[4] | 7 | M | [1]Both Knowledge and Communication has improved evidenced by the document "Improvement of the distribution keys for the estimated trade in the Swedish Intrastat."<br>[2]Key staff have made national presentations in connection with the WG Quality Meetings elevating expertise.<br>[3]Swedish Customs adopted SCB's editing system which indicates state of the art systems.<br>[4]Plans are in place to study more sophisticate models for estimation under cutoff using VAT possibly using the Vat Information Exchange System (VIES). |
| Revision error | 62 | 76 | 5→7[1] | 5→7[1] | 7 | 7→9[2] | 7→8[3] | ~~L~~H[4] | [1]Knowledge and communication of risks improved through writing the QD as well as preparation of the documents "Improvements of the work on revisions in the Swedish goods."<br>[2]Compliance with standards and best practices enhanced through Standardized Toolbox. Above referenced document also provides evidence that best practices are being followed. Progress has been made to rapidly detect and repair causes of large revisions.<br>[3]Plans being developed to identify causes of revision error.<br>[4]The risk level was re-evaluated and elevated to H as a result of the impact on the NA statistics. |
| **Total score** | 57,3 | 65,8 | | | | | | | |

Note: *(Shaded cells denote either improvements (light) or deteriorations (dark) in ratings since Round 1. Corrections denoted by strikeouts with correct rating inserted. Footnotes describe reasons for the changes.*)




*Note: Some items follow the deflation process in the opposite direction and are complied starting with information on volume change from the previous year. The volume estimate is then reflated with the price index in order to come to the current price estimate. Items within the Energy sector is one such example.

*Fig. 1. High-Level Process Flow Diagram for Estimating Current and Constant Price GDP by Production and Expenditure Approaches*

adding taxes and deducting subsidies (subs). For constant prices, the current prices must be "deflated" using the appropriate price indices before adjustments for taxes and subsidies.

Both the production and expenditure approaches will produce interim estimates of GDP (both current and constant prices) which must then be "balanced" or forced into agreement as the economic theory dictates (see, for example, Lequiller and Blades 2006). This balancing process produces the preliminary estimates of GDP for both current (denoted by Cu in the exhibit) and constant (denoted by Co) prices. The latter differs from the former primarily by a deflation/reflation process that adjusts prices to a common base-year. The preliminary estimates are subsequently revised when additional data become available. Thus, the error sources associated with the GDP estimation process are as shown in Table 1, bottom panel.

In the evaluation of production GDP, considerable attention was given to the error in the inputs and their effects on the error in the GDP estimates. Priority was given to inputs that posed the greatest risk to GDP error. These were determined by the evaluators in collaboration with the National Accounts staff.

### *4.3. Overall Results for All Products for Round 2*

This section further illustrates some important uses of ASPIRE to compare the scores of all ten products in Round 2. Table 4 provides the overall scores for the six survey products and two registers and Table 5 provides the overall scores for the National Accounts only because the structure of their error sources is quite different from the other products. To facilitate the exposition of the results, the error sources were consolidated into a single list which appears in the first column of Table 4. The other columns of the table refer to the particular product being evaluated. For each product, the bold figures correspond to "High Risk" error sources, italic corresponds to "Medium Risk," and non-bold corresponds to

*Table 4. Product Error-Level, Overall Level, and Error Source-Level Ratings with Risk-Levels Highlighted and Comparisons to Round 1 Overall Ratings*

| Error Source | RS | CPI | FTG | LFS | SBS | LCS | BR | TPR | Error Source Mean Rating |
|---|---|---|---|---|---|---|---|---|---|
| *Specification error* | N/A | **68** | *58* | 70 | *54* | *34* | 66 | *46* | 57 |
| Frame error | 60 | *62* | 58 | 58 | *64* | **42** | 55 | 62 | 58 |
| Overcoverage | | | | | | | **56** | **56** | |
| Undercoverage | | | | | | | *46* | 60 | |
| Duplication | | | | | | | 63 | 70 | |
| Nonresponse error/Missing data | *52* | 55 | *66* | **52** | *70* | **40** | 48 | *66* | 56 |
| Measurement error/Content error | *58* | **62** | **62** | **56** | **52** | **46** | **46** | 58 | 55 |
| Data processing error | **48** | **76** | **60** | *62* | **60** | 42 | N/A | N/A | 58 |
| Sampling error | N/A | **66** | N/A | *78* | *84* | *54* | N/A | N/A | 71 |
| Model/estimation error | **38** | **52** | *80* | *60* | **60** | **38** | N/A | N/A | 55 |
| Revision error | 58 | N/A | **76** | N/A | **56** | N/A | N/A | N/A | 63 |
| **Round 2 Mean Rating** | 49,6 | 63,9 | 65,8 | 60,9 | 61,4 | 42,1 | 52,2 | 58,0 | 57 |
| **Round 1 Mean Rating** | 46,7 | 60,3 | 57,3 | 56,4 | 59,6 | N/A | 47,2 | 52,2 | 54 |
| Improvement | 2,9 | 3,6 | 8,5 | 4,5 | 1,8 | N/A | 5,0 | 5,8 | 2,5 |

*In this table,individual and mean ratings can be compared across products (columns) and by error source (rows) as well as. Note, for example, that the LCS and Measurement error/Content Error have the lowest average ratings. The FTG shows the greatest improvement from Round 1 to Round 2.*

**BOLD** = HIGH RISK

*ITALICS* = MEDIUM RISK

REGULAR FONT = LOW RISK

N/A = NOT APPLICABLE

*Table 5. Product Error-Level, Overall Level, and Error Source-Level Rating with Risk-Levels Highlighted for the National Accounts*

| Error Source | GDP Quarterly | GDP Annual |
|---|---|---|
| Input source (Average) | **53** | **66** |
| Structural Business Survey (SBS) | N/A | **66** |
| Index of Service Production (ISP) | **58** | N/A |
| Index of Industrial Production (IIP) | **58** | N/A |
| Merchanting Service of Global Enterprises | **42** | N.E. |
| Compilation error (modelling) | **48** | **48** |
| Compilation error (data processing) | **40** | **35** |
| Deflation error (including specification error) | **48** | **48** |
| Balancing error | **56** | **50** |
| Revision error | *56* | *54* |
| **Round 2 Mean Rating** | 50,5 | 49,9 |

**BOLD = HIGH RISK**
*ITALICS = MEDIUM RISK*
REGULAR FONT = LOW RISK
N/A = NOT APPLICABLE
N.E. = NOT EVALUATED

"Low Risk" error sources. The same applies to the second table for the two National Accounts products. Note that the interpretation of the error sources (see Subsection 2.3) and criteria may vary between surveys and registers.

Before discussing the results in Tables 4 and 5, a few cautions should be stated. There is a natural tendency to compare the overall scores across the products or to rank the products by their total score. The interpretation of such comparisons may not be straightforward for several reasons. First, the total score for a product reflects a weighting of the error sources by the risk levels which can vary considerably across products. Products with many high risk error sources, such as the National Accounts, may be at somewhat of a disadvantage in such comparisons because they must perform well in many high risk areas in order to achieve a high score. Second, the assessment of low, medium, or high risk is done within a product, not across products. Thus, it is possible that a high risk error source for one product could be of less importance to Statistics Sweden than a medium risk error source for another product if the latter product carries greater importance to Statistics Sweden. (For example, measurement error for the ULF/SILC may be somewhat lower priority than it is for the CPI.) Finally, the scores assigned to a particular error source for a product have an unknown level of uncertainty due to a number of factors. We believe rating consistency and reliability considerably improved with the development of the checklist as discussed above. Still, a difference of 2 or 3 points in the overall product scores may not be meaningful because an independent reassessment of the product could reasonably produce a new score that differs from the current score by that margin. Note further that, because of the very different approach taken in Round 2 for the National Accounts, comparisons to Round 1 for the GDP ratings are not meaningful.

Close inspection of scores in Tables 4 and 5 yield the following general observations:

- The average score for all products in Round 2 was 57 compared to 54 in Round 1 – a 5.6 percent improvement in the ratings. However, among products evaluated in both

*Table 6. User Dimensions and their Components*

| Timeliness & Punctuality | Accessibility & Clarity |
|---|---|
| • Timeliness of release of main aggregates<br>• Timeliness of release of detailed outputs (including microdata)<br>• Punctuality | • Ease of data access<br>• Documentation (including metadata)<br>• Availability of Quality Reports<br>• User support |
| **Comparability & Coherence** | **Relevance/Contents** |
| • Comparability across geography, populations, and other relevant domains<br>• Comparability across time (including impacts of redesign)<br>• Coherence with other relevant statistics (including use of standard classifications, frameworks, etc.) | • Inputs (content, scope, classification, etc.)<br>• Outputs (including microdata and other products) |

rounds, the improvement was about 8.5 percent. The introduction of ASPIRE undoubtedly led to some of these improvements as the ratings for all seven products that were reviewed in Round 1 improved in the current round. A significant influence was the development of Quality Declarations consistent with one of the strong recommendations of the evaluators.

- In both rounds, measurement error had the highest average inherent risk of any error source. It also ranked near the bottom in percent mitigated risk, defined as the total points earned divided by the maximum points achievable for an error source expressed as a percentage.
- By contrast, sampling error ranked the highest in percent mitigated risk, earning roughly 70% of the maximum points achievable in both rounds. Revision error is also highly ranked although it only applies to three products in Table 5 and the two National Accounts products.
- "Available expertise" and "compliance with standards and best practices" are generally rated higher than "knowledge of risks," "communication (of these risks) with users," and "achievement towards risk mitigation or improvement plans." The latter three criteria appear more challenging to most products.

ASPIRE identified many areas where improvements to data quality are needed with the highest priorities assigned to areas having high risks and low ratings. In addition, a number of "cross-cutting" recommendations were made. These are recommendations that affect multiple products such as: better documentation of quality and use of quality profiles, more evaluations of measurement errors, improved IT-client relationships, better succession planning in some areas, and so on. Costs varied considerably among the recommendations and limited resources constrained the scope of the improvements that Statistics Sweden could pursue. Because some improvement projects, particularly those that cut across product areas, required substantial allocations or reallocations of funding, decisions regarding which projects and activities to pursue in the

future should be left to management. Nevertheless, product areas may have some capacity to implement the most important improvements and this has happened to some extent.

The results of all three rounds of ASPIRE can be found in Biemer and Trewin (2012, 2013, and 2014). These reports are available by request from the authors.

### *4.4. Assessing the User Dimensions*

As noted previously, the ASPIRE system was expanded in Round 2 to incorporate a process for evaluating the four user dimensions of quality. These are Accessibility & Clarity, Comparability & Coherence, Relevance/Contents, and Timeliness & Punctuality. The primary goal of this application was to develop a process for assessing the user quality dimensions. The system was tested on two products: the LFS and the CPI. The evaluation framework is completely consistent with the Accuracy framework; that is, each dimension was first decomposed into mutually exclusive components (analogous to the error sources defined for Accuracy) which, for the most part, are those described in the ESS Quality Assurance Framework (ESS 2011). Quality for each component was assessed according to five criteria that are similar to the five Accuracy criteria; viz., Knowledge of User Needs, Communication with Users, Available Expertise (to address user needs), Compliance with Standards and Best Practices, and Plans toward Addressing User Needs and were applied to each of the components under a dimension.

The components associated with each user dimension appear in Table 6. As was done for Accuracy, checklists were developed for each criterion and were generic across dimensions and components within dimensions.

The LFS was evaluated for Timeliness & Punctuality and Comparability & Coherence and the CPI was evaluated for Relevance/Contents and Accessibility & Clarity. The assessment process, which proceeded much like the process for Accuracy, seemed to work well for their initial application. However, some needed improvements were identified. For example, the checklists and criteria could be enhanced to better capture the risks of poor quality associated with each dimension. Also, direct communication with the users of these statistics is recommended to provide information on quality from the broader user community. In this trial evaluation, we largely relied on the advice of product staff on their interaction with users.

## 5. Discussion

Although this article has focused on the application of ASPIRE to ten Statistics Sweden products, it can be applied much more generally. As we have demonstrated, it can be used for survey products, administrative data products, registers and 'compilation' products such as the National Accounts. It can also be applied in other government statistical offices as well as in private sector or university statistical products. By design, it performs best for products that recur regularly and that are reviewed repeatedly so that improvements (or deteriorations) in quality can be assessed across time. While one-time ASPIRE reviews could provide useful insights regarding a product's current quality-level, multiple reviews would be more effective if the

objective is quality improvement. We believe that annual reviews are sufficiently frequent to track improvements for most programs. Less frequent (say biennial) reviews may be sufficient for lower risk programs or programs whose improvement efforts require more than one year to generate measureable results.

Any method for evaluating the quality of products as complex as those considered in this article will have its limitations. Estimating the total MSE (or even its key components) for a product such as the CPI or quarterly GDP is virtually impossible because the data required are largely unobtainable. Further, any data that can be collected on nonsampling errors are themselves subject to nonsampling errors. For example, a survey of nonrespondents to estimate the nonresponse bias in the LCS/ULF is also subject to nonresponse. The ASPIRE approach does not provide direct measures of the total MSE of a product. However, ASPIRE's ratings are negatively correlated with the risks of poor data quality; specifically, improved quality ratings reflect lower error risks. In addition, ASPIRE ratings are positively affected when MSE components have been estimated. For example, the rating for Knowledge of Risks is elevated when the bias from the error source has been estimated. Likewise, the rating for Communication with Users is elevated if those estimates have been documented and disseminated.

As noted in Section 3, the primary goals of ASPIRE are to identify the current, most important threats or risks to the quality of a product, apply a structured, comprehensive approach for rating the efforts aimed at reducing these risks, and identify areas where future efforts are needed to continually improve process and product quality focussing on those high risk error sources where ratings are relatively low. We believe that product quality will improve to the extent that ASPIRE achieves these three goals. A key requirement for this is that inputs to process – in particular, the information needed to accurately assess each criterion – are accurate, complete, timely, and accessible by the evaluators. Thus, continuing to update and improve the documentation of quality is an important activity to ensure ASPIRE's success.

Based upon this work, we believe ASPIRE succeeds in four areas. First, the approach is comprehensive in that it (a) covers all the important sources of error for a product and (b) uses criteria that span all the important risks to product quality. Second, the checklists used to assign the ratings under each criterion seem quite effective at identifying and assessing both manifest and hidden risks to data quality. To the extent that the documentation and other information shared during the ASPIRE process is both accurate and complete, the current approach assigns reliable ratings that reflect true data quality risks. Third, ASPIRE successfully identifies areas where, from an organizational perspective, improvements are needed and have very high priority. It further prioritizes these needs when it is not possible or sensible to undertake all quality improvements. For example, areas having highest risk and lowest ratings, assuming other factors are equal, should be assigned highest priority for improvement. Of course, the overall importance of the product relative to other products also should be taken into account as well as the resource requirements and the likely success of the improvement effort.

Finally, if implemented appropriately, the ASPIRE framework should generally increase organizational transparency and accountability both internally and externally.

Within the organization, this will enhance communication across products and quality improvement projects thus fostering greater collaboration and sharing of quality improvement ideas and results. Externally, this transparency will lead to greater organizational credibility and product confidence. In addition, providing this detailed information on data quality issues to external users can generate external pressure on the organization to make swifter and greater progress on quality improvements.

One weakness of the model is that it is, at best, a proxy measure for product quality because it makes no attempt to estimate the TSE and its components. However, quantitative assessments of TSE are reflected in the ratings and can also be used to supplement the information obtained from our approach. Another potential weakness of the approach is that it can be somewhat subjective in that it relies heavily on the knowledge, skill, and impartiality of the evaluators. However, we believe it would be undesirable to remove all the subjectivity from the process because that would be akin to automating the review process. A purely objective process may not optimally utilize the expertise of the evaluators nor allow for more complex judgments to be applied to the process. It is important, however, that any subjectivity in the ratings does not lead to inequities and inconsistencies across reviews. A number of safeguards have been put in place to prevent these potential adverse effects including the quality guidelines, checklists, the rating revision process, and the ratings appeal process.

With respect to possible future research, there are several thrusts. First, further testing and evaluation of the ASPIRE approach should focus on its long-term effects on product quality. For example, there could be some assessment of value of improvements projects that have been launched following recommendations from the ASPIRE process. Key users should be informed of the improvements completed and still underway and consulted to obtain their views on whether quality has been improved. Thus, the evaluation could determine whether quality improvements have increased under ASPIRE and whether ASPIRE is worth the investment of resources. The evaluation might also assess whether actual improvements correlate well with the changes in ratings for individual products and the quantitative information on error components that might be available for some products. Finally, staff within the organization should be consulted in the evaluation to elicit their opinions regarding the benefits and issues associated with ASPIRE.

Second, research could be conducted to further reduce inter-rater variation as well as intra-rater bias. Cognitive laboratories might be used for this purpose. Third, further work could extend the ASPIRE approach to the user dimensions. Whilst external evaluators are preferred, a satisfactory evaluation of the user dimensions could rely primarily on internal evaluators by using the structured approach we propose for obtaining feedback from both internal and external users across the range of quality dimensions.

Finally, we hope to see ASPIRE or a similar approach be implemented in other NSOs to see if similar quality improvements can be realized in other countries and organizations. For the sake of cross-country comparisons, settling on a unified approach that is applicable across diverse NSOs and cultures would offer clear advantages.

## Appendix A – Evaluation Criteria and Guidelines for Accuracy

**Exhibit 1.1a. Knowledge of Risks**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ◎ |
|---|---|---|---|---|
| Program documentation does not acknowledge the source of error as a potential factor for product accuracy. | Program documentation acknowledges error source as a potential factor in data quality. **But:** No or very little work has been done to assess these risks. | Some work has been done to assess the potential impact of the error source on data quality. **But:** Evaluations have only considered proxy measures (example, error rates) of the impact with no evaluations of MSE (bias and variance) components. | Studies have estimated relevant MSE components associated with the error source and are well-documented. **But:** Studies have not explored the implications of the errors on various types of data analysis including subgroup, trend, and multivariate analyses. | There is an ongoing program of research to evaluate all the relevant MSE components associated with the error source and their implications for data analysis. The program is well-designed and appropriately focused, and provides the information required to address the risks from this error source. |

**Exhibit 1.1b. Communication with Users**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ◎ |
|---|---|---|---|---|
| Reports, websites, and other communications with data users and customers are devoid of any mention of the error source. | There is acknowledgement of the risks of error from this source. **But:** Communications have been largely inadequate considering the importance of these potential risks to data quality. | Communications with users and customers have adequately described the risk to many users. **But**: Information conveyed has largely been sampling errors and/or proxy measures with little communications regarding MSE components or the risks have been downplayed leading to a false sense of security. | Communications have shared some of the available information on the relevant MSE components that have been evaluated and the true risks to users have been appropriately conveyed. **But:** The information conveyed in could be improved in one or more of these areas: (a) more clarity so that complex ideas are comprehensible to less sophisticated users, (b) improved presentation so data analysts can apply the knowledge more directly in their analyses, or (c) a fuller discussion of the implications of the findings for various types of data analysis so that users can make informed decisions regarding the results. | Communications regarding the error source have been thorough, cogent, and clear. An appropriate level of detail has been included in the communications so that users should be fully aware of any risks of the error source to data quality and are provided with all the information they need to deal with the risks appropriately in their analyses. |

**Exhibit 1.1c. Available Expertise**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ◎ |
|---|---|---|---|---|
| Among the staff assigned to work on the product, either (a) there are no staff that are familiar with techniques that will be required to deal with the potential risks to accuracy for the product or (b) the expertise of staff that are assigned is sorely inadequate. | The available expertise required to study this error source and communicate the findings of such studies to data users is adequate in at least one important area. **But:** For most important areas expertise is still lacking. | The available expertise required to study this error source and communicate the findings of such studies to data users is adequate in most of the important areas. **But:** Either (a) there is at least one area that may be critical to accuracy where a higher level of expertise is needed or (b) there are one or more minor areas that could become important in the future that are not well staffed. | The available expertise required to study this error source and communicate the findings of such studies to data users is adequate in all important areas. There is a good working relationship with the key groups involved in activities associated with this error source. Staff are keeping up to date with developments in their areas of expertise **But:** There are one or more minor areas that could become important in the future which are not well covered. Current expertise is not adequate to achieve the highest ratings for all evaluation criteria for this error source or the expertise would not be readily available to work on these error sources. | The available expertise required to study this error source and communicate the findings of such studies to data users is more than adequate to achieve the high ratings across all evaluation criteria. The relevant experts are actively addressing errors from the source. There is an excellent working relationship with the key groups involved in activities associated with this error source. Staff are keeping up to date with and contributing to developments in their areas of expertise. |

**Exhibit 1.1d. Compliance with Standards and Best Practices**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ◎ |
|---|---|---|---|---|
| Staff are mainly unaware of standards and best practices that are relevant for this error source. If some awareness exists, there is no evidence that standards and best practices, as they related to this error source, have been applied to the product. Moreover, serious deficiencies exist that violate standards and best practices as they relate to this error source. | Staff are aware of standards and best practices and there is evidence that these have been applied to the product for this error source. **But:** There are still important areas of noncompliance that need to be addressed. These gaps are not currently being addressed or actions to address them have been inadequate. | Staff are well aware of relevant standards and best practices and have clearly applied them to the product. Important violations or gaps are being actively addressed. **But:** Either (a) compliance is not routinely monitored or (b) gaps in compliance exist for some minor areas that are not being addressed. | Staff are well aware of the relevant standards and best practices and have clearly applied them to the product. There are no serious violations of standards and best practices as they relate to this error source **But:** Some staff may not keep up to date with latest standards and developments in best practices that are relevant to their work. Compliance may not be routinely monitored. | The product is fully compliant with agreed standards and best practice. The relevant staff are fully aware of the standards and best practices and continually monitor the work to ensure that compliance is maintained. They are actively keeping up to date with and contributing to latest standards and developments in best practices. |

**Exhibit 1.1e. Achievement Towards Mitigation and/or Improvement Plans**

| Poor [1,2] ● | Fair [3,4] ◓ | Good [5,6] ○ | Very Good [7,8] ◒ | Excellent [9,10] ◎ |
|---|---|---|---|---|
| There is no evidence that any planning has been done for studying or mitigating the risks for this error source. | An overall plan for error reduction with measurable objectives exists for mitigating the risks for this error source.<br>**But:** The plan is not approved by the appropriate level of management. | A management-approved plan with measurable objectives exists. The plan adequately addresses the work required for mitigating the risks of poor data quality relative to this error source. . .<br>**But:** One of the following deficiencies with the plan exists:<br>a. The overall plan has not been updated in at least one year.<br>b. There is no accountability in place to ensure compliance with the plan. c. No mechanism is specified for gauging progress toward each objective.<br>d. No resources have been allocated to implement the plan. | Resources have been allocated to undertake this work. Considerable progress has been made on the plan for mitigating the risks to data. None of the deficiencies noted under the "Good" criteria are present.<br>**But:** Efforts have not yet produced the desired control over the error source that is stipulated in the plan. | Mitigation plans have been fully implemented or well underway. Progress toward all goals and objectives has been excellent. As a result, the level of error in the final estimates due to this error source is being maintained at an acceptable level for the primary purposes of the data. As a result of these efforts, the error source is under control and poses no or very little risk to data quality. Results of the mitigation activities have been fully documented. Accountability measures are in place to ensure compliance with the plans. The mitigation plans are reviewed and updated periodically. |

## Appendix B – Example of a Criterion Checklist (Knowledge of Risks)

*For each applicable error source, indicate either compliance or noncompliance with an item in the checklist by marking "Yes" or "No," respectively. In order to achieve a higher rating for a criterion, all items for that higher rating must be checked. You may use the "Comments" field to provide comments you deem necessary to explain your response to an item.*

| Knowledge of Risks | Check Box | Comments |
|---|---|---|
| 1. Documentation exists that acknowledges this error source as a potential risk. | ☐ Yes<br>☐ No<br>**Fair** | |
| 2. The documentation indicates that some work has been carried out to evaluate the effects of the error source on the key estimates from the survey. | ☐ Yes<br>☐ No<br>**Good** | |
| 3. Reports exist that gauge the impact of the source of error on data quality using proxy measures (e.g., error rates, missing data rates, qualitative measures of error, etc.) | ☐ Yes<br>☐ No<br>**Good** | |
| 4. At least one component of the total MSE (bias and variance) of key estimates that is most relevant for the error source has been estimated and is documented. | ☐ Yes<br>☐ No<br>**Very Good** | |
| 5. Existing documentation on the error source is of high quality and explores the implications of errors on data analysis. | ☐ Yes<br>☐ No<br>**Excellent** | |
| 6. There is an ongoing program of research to evaluate the components of the MSE that are relevant for this error source. | ☐ Yes<br>☐ No<br>**Excellent** | |

## 6. References

Andersen, R., J. Kaspar, and M. Frankel. 1979. *Total Survey Error*. San Francisco: Jossey-Bass Publishers.

Baldrige Performance Excellence Program 2013. *The 2013–2014 Criteria for Performance Excellence*. Available at: http://www.nist.gov/baldrige/ (accessed August 3, 2013).

Barkley, B.T. 2004. *Project Risk Management*. New York: McGraw Hill Professional.

Biemer, P. 2011. *Latent Class Analysis of Survey Error*. Hoboken, NJ: John Wiley & Sons.

Biemer, P. 2014. "Comment on 'On Information Quality' by Kenett and Shmueli." *Journal of the Royal Statistical Society,* Series A. Vol. 177, Part 1: 27–29.

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley & Sons.

Biemer, P. and D. Trewin. 2012. *Development of Quality Indicators at Statistics Sweden*. Report to Statistics Sweden, January 2012.

Biemer, P. and D. Trewin. 2013. *A Second Application of the ASPIRE Quality Evaluation System for Statistics Sweden*. Report to Statistics Sweden, January 2013.

Biemer, P. and D. Trewin. 2014. *A Third Application of ASPIRE for Statistics Sweden*. Report to Statistics Sweden, January 2014.

Brackstone, G. 1999. "Managing Data Quality in a Statistical Agency." *Survey Methodology* 25: 139–149.

Breyfogle, F. 2003. *Implementing Six Sigma*, 2nd edition. New York: John Wiley & Sons.

Conley-Tyler, M. 2005. "A Fundamental Choice: Internal or External Evaluation?" *Evaluation Journal of Australasia* 4: 3–11.

COSO, 2004. *Enterprise Risk Management – Integrated Framework*. Available at: http://www.coso.org/documents/coso_erm_executivesummary.pdf (accessed August 3, 2013).

COSO, 2013. *Internal Control – Integrated Framework, 2013*. Available at: http://www.coso.org/documents/coso%202013%20icfr%20executive_summary.pdf (accessed August 3, 2013).

Couper, M. and L. Lyberg. 2005. "The Use of Paradata in Survey Research." In Proceedings of the 55th Session of the International Statistical Institute, Sydney, Australia, April 7, 2005. Available at: http://isi.cbs.nl/iamamember/CD6-Sydney2005/ISI_Final_Proceedings.htm (accessed June 26, 2014).

Curtin, R., S. Presser, and E. Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64: 413–428.

Dalenius, T. 1967. *Nonsampling Errors in Census and Sample Surveys*. Report no. 5 in the research project Errors in Surveys, Stockholm University.

Deming, E. 1944. "On Errors in Surveys." *American Sociological Review* 9: 359–369.

Deming, E. 1986. *Out of the Crisis*. Cambridge, MA: MIT Press.

EFQM, 2013. "An Overview of the Excellence Model." Available at: https://www.google.com/url?q=http://www2.efqm.org/en/PdfResources/EFQM%2520Excellence%2520Model%25202013%2520EN%2520extract.pdf&sa=U&ei=9BasU4nkHsqTqAbUhIGQCg&ved=0CAUQFjAA&client=internal-uds-cse&usg=AFQjCNHthJnhRPIS1t6cfa4Ka9ePXOLRxg (accessed June 26, 2014).

Eltinge, J., P. Biemer, and A. Holmberg. 2013. "A Potential Framework for Integration of Architecture and Methodology to Improve Statistical Production Systems." *Journal of Official Statistics* 29: 125–145. DOI: http://dx.doi.org/10.2478/jos-2013-0007.

European Statistical System (ESS) 2011. "Quality Assurance Framework of the European Statistical System, Version 1.1." Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF (accessed August 9, 2013).

Eurostat 2005. "European Statistics Code of Practice, Revised Edition." Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF (accessed June 26, 2014).

Eurostat 2009. *Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009, Eurostat General/Standard report*, Luxembourg, April 4–5. Available at: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009R0223 (accessed June 18, 2014).

Gonzales, M.E., J.L. Ogus, G. Shapiro, and B.J. Tepping. 1975. "Standards for Discussion and Presentation of Errors in Surveys and Census Data." *Journal of American Statistical Association* 70: 5–23.

Groves, R.M. and L.E. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74: 849–879. DOI:http://dx.doi.org/10.1093/poq/nfq065.

Hansen, M., W. Hurwitz, and W. Madow. 1953. *Sample Survey Methods and Theory, Volumes I and II*. New York: John Wiley & Sons.

Hansen, M., W. Hurwitz, and L. Pritzker. 1967. *Standardization of Procedures for the Evaluation of Data: Measurement Errors and Statistical Standards in the Bureau of the Census*. Paper presented at the 36th session of the International Statistical Institute.

Imai, M. 1986. *Kaisen: the Key to Japan's Competitive Success*. New York: McGraw-Hill Education.

International Monetary Fund (IMF) 2003. *Data Quality Assessment Framework and Data Quality Program*. Available at: http://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm (accessed June 21, 2013).

International Standards Organization 2006. *Market, Opinion and Social Research ISO Standard No. 20252*. Available at: www.standards.org/standards/listing/iso_20252 (accessed August 8, 2014).

International Standards Organization 2009. *Risk Mangement: Principles and Guidelines for Implementation*, ISO/DIS 31000 Standard No. 31000. Available at: www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43170 (accessed August 8, 2014).

Journal of Official Statistics 2013. *Special Issue on Systems and Architectures for High-Quality Statistics Production*, edited by B. Lorenc, I. Jansson, P. Biemer, J. Eltinge, and A. Holmberg, Vol. 1, March, 2013.

Juran, J. and B. Godfrey. 1999. *Juran's Quality Handbook*. New York: McGraw-Hill.

Karsak, E.E. 2004. "Fuzzy Multiple Objective Decision Making Approach to Prioritize Design Requirements in Quality Function Deployment." *International Journal of Production Research* 42: 3957–3974.

Keeter, S., C. Miller, A. Kohut, R. Groves, and S. Presser. 2000. "Consequences of Reducing Nonresponse in a Large National Telephone Survey." *Public Opinion Quarterly* 64: 125–148. DOI: http://dx.doi.org/10.1086/317759.

Kenett, R.S. and G. Shmueli. 2014. "On Information Quality." *Journal of the Royal Statistical Society,* Series A 177: 3–38. DOI:http://dx.doi.org/10.1111/rssa.12007.

Kish, L. 1962. "Studies of Interviewer Variance for Attitudinal Variables." *Journal of the American Statistical Association* 57: 92–115.

Lequiller, F and D. Blades. 2006. *Understanding National Accounts*. Paris: OECD 2006. Available at: http://www.eastafritac.org/images/uploads/documents_storage/Understanding_National_Accounts_-_OECD.pdf (accessed June 21, 2013).

Lyberg, L. and P. Biemer. 2008. "Quality Assurance and Quality Control in Surveys." In *International Handbook on Survey Methodology*, edited by J. Hox, E. de Leeuw, and D. Dillman, 421–441. Mahwah, NJ: Lawrence Erlbaum Associates.

Lyberg, L., L. Japec, and P. Biemer. 1998. "Quality Improvement in Surveys – A Process Perspective." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 23–31.

Lyberg, L. 2012. "Survey Quality." *Survey Methodology* 38: 107–130.

McDavid, J., I. Huse, and L. Hawthorn. 2013. *Program Evaluation and Performance Measurement: An Introduction to Practice, Second Edition*. New York: Sage Publications.

Michalek, J.J., O. Ceryan, P.Y. Papalambros, and Y. Koren. 2006. "Balancing Marketing and Manufacturing Objectives in Product Line Design." *ASME Journal of Mechanical Design* 128: 1196–1204. DOI: http://dx.doi.org/10.1115/1.2336252.

Merkle, D. and M. Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, edited by R. Groves, D. Dillman, J. Eltinge, and R. Little, 243–257. New York: John Wiley and Sons.

Morganstein, D. and D. Marker. 1997. "Continuous Quality Improvement in Statistical Agencies." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 475–500. New York: Wiley and Sons.

Nealon, J. and E. Gleaton. 2013. "Consolidation and Standardization of Survey Operations at a Decentralized Federal Statistical Agency." *Journal of Official Statistic* 29: 5–28. DOI: http://dx.doi.org/10.2478/jos-2013-0002.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–606.

Neyman, J. 1938. *Lectures and Conferences on Mathematical Statistics and Probability*. Washington, DC: U.S. Department of Agriculture.

Organisation for Economic Cooperation and Development (OECD) 2011. *Quality Framework and Guidelines for OECD Statistical Activities*. Available at: http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs%282011%291&doclanguage=en (accessed June 21, 2013).

Office of National Statistics (ONS) 2007. *Guidelines for Measuring Statistical Quality, Version 3.1*. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html (accessed June 21, 2013).

Rossi, P.H., W.M. Lipsey, and H.E. Freeman. 2004. *Evaluation: A Systematic Approach*. 7th ed. Thousand Oaks, CA: Sage Publishers.

Seyb, A., R. McKenzie, and A. Skerrett. 2013. "Innovative Production Systems at New Zealand: Overcoming the Design and Build Bottleneck." *Journal of Official Statistics* 29: 73–97. DOI: http://dx.doi.org/10.2478/jos-2013-0005.

Statistics Canada 2009. *Statistics Canada Quality Guidelines, Fifth Edition*. Available at: http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-539-X&CHROPG=1&lang=eng (accessed March 10, 2014).

Statistiska centralbyrån 2001. *Quality Definition and Recommendations for Quality Declarations of Official Statistics*. Available at: http://www.scb.se/Grupp/Hitta_statistik/Forsta_Statistik/Metod/_Dokument/MIS2001_1.pdf (accessed June 18, 2014).

Stephan, F.F. 1948. "History of the Uses of Modern Sampling Procedures." *Journal of the American Statistical Association* 43: 12–39.

Struijs, P., A. Camstra, R. Renssen, and B. Braaksma. 2013. "Redesign of Statistics Production within an Architectural Framework: The Dutch Experience." *Journal of Official Statistics* 29: 49–71. DOI: http://dx.doi.org/10.2478/jos-2013-0004.

U.S. Bureau of the Census 1974. "Technical Paper 32: Standards for Discussion and Presentation of Errors in Data. U.S. Department of Commerce." U.S. Government Printing Office, Technical Paper 32, Department of Commerce.

U.S. Office of Management and Budget 2002. "Guidelines for Ensuring, and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies." *Federal Register*, 67, 36, February 22.

Received August 2013
Revised April 2014
Accepted June 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 417–419, http://dx.doi.org/10.2478/JOS-2014-0026

# Discussion

*Fritz Scheuren*[1]

## 1. Overall Comments

I love this article! Its wisdom and focus on action are refreshing, especially on a subject, like quality, that is often approached with big words and small-to-few deeds. So, first, as a reader, a big thank you!

I will emphasize some of the authors' points further in what follows; but as someone who also was a referee, let me provide a special thanks to the authors for their listening skills. Because of that I have nothing negative to say. However, while I cannot comment critically, let me still kibitz a bit, hopefully not to an annoying degree? I will use the Dickens' character (Dickens 1838), Oliver Twist, in particular, Oliver's polite words: " Please sir, I want some more?"

Quality is a complex, multi-faceted subject. That the authors chose to emphasize product quality and specifically, accuracy, conditioning on its other aspects was wise, given the audience of official statisticians they are writing too. As Juran has put it well, one needs to "fool the immune system" to get anything really new across. That is why starting with the quality attribute "accuracy," or mean square error (MSE), was so well chosen. It is a part of quality we all know well.

Incidentally, I was a student of Juran and worked beside him for a while in the 1980s, so I am quoting from memory. But I have provided one of many possible useful references (Juran and Godrey 1999) for those wishing to go deeper into the insights of this great quality guru.

But, if we are to escape "paradigm paralysis" and achieve full "systems thinking," as Deming has advocated, we must aspire (love that word too) for much more. Notice I have just used the word "system" in two senses. Sorry!

Again, with the phrase "systems thinking" I am quoting from memory. I have had the opportunity to be a student of Deming, as well as Juran. While famously nearly deaf in his old age, Deming retained his strong voice and Old Testament style to the end of his life – so unlike the gentle but equally insistent Juran. Deming and Juran were good friends, by the way. And, of course, I loved them both!

In the article the authors characterize the dimensions of product quality, besides accuracy, as *user dimensions* (their emphasis). Tukey might have said that this is a

[1] NORC, University of Chicago, 1155 East 60th Street, 3rd Floor, Chicago, IL 60637, U.S.A. Email: scheuren-fritz2@norc.org

"roughly right" formulation – as long as we include all stakeholders, not just end customers, as users. That means respondents and taxpayers and employees, among others.

For more discussion of the traditional and some new quality attributes or characteristics to be emphasized, see the second edition of the forthcoming book by Herzog et al. (2014).

But let me take a moment to mention the changing systems capabilities of many of those users. To this end, let me add a little more to the fine (albeit brief) discussion summarizing our professional literature on quality.

In the days when I was starting out in statistics (before most of you were born) end users at best were statistically literate enough to read, not a lot more than, a simple two-dimensional (R by C) paper-published table. They had no computing power beyond a calculator, if that.

Today, my/our customers take survey and administrative data sets in an electronic form, frequently inputting our data into their micro-simulation (*what if*) models. Customers, many times, are even co-creating with us their own information – perhaps, their own data products. My work on the March Income supplement to the US Current Population Survey would be one example (e.g., Turek et al. 2012)

How does what got said in the article change, even when seen from this wider perspective? Not much in my view. ASPIRE, as has been well explained in the article, has within its structure enough to accommodate this enlarged system: Where what we do for our end users is more of a service input for them and not just a traditional product, as in the old days when I was young.

But, by making this observation, maybe you will agree with me that there is a change in emphasis. That change can be important in some venues. So, this leads me to my first "more please."

1. We should all follow the advice that the authors give us to apply this toolkit not just elsewhere but close to home – in a cooperative effort with our (very sophisticated and perhaps even more resourced than we are) customers. Our joint systems are interlocked and that insight is actionable with the tools in this article. The simulations done in the US for tax policy, social security and welfare reform are instances where I have been involved personally.

Permit me briefly to discuss two further "more please" requests, not just for the authors but perhaps for all of us: (2) How we can listen better both top down and bottom up; and (3) How we can afford all this better and better quality? After all, it is unrealistic to expect more resources

2. The Japanese word, converted into English, Kaisen, as used in the article in passing is roughly translated as "continuous improvement." But it is a cultural construct, not a paradigm. It is an attitude towards our work and our lives that is, once imbedded in habit, a real asset to everything we do. ASPIRE must and can be partly like that too, not just periodic but empowering. A way to regularly check on the systems we work inside of.
3. Big changes are expensive and dangerous, especially for official statisticians, who are inherently risk adverse and live, like the earth, in a narrow band between fire and Ice. So, if ASPIRE is to become embedded, it must *not* be resource intensive. And not just in monetary terms but also in terms of opportunity costs? This is not

> easy for westerners, especially Americans like me, who want to fix everything quickly and then go in to do something new.

Again, let me emphasize the important contribution that this article or paradigm really represents. The ideas here deserve follow up. For example, how about establishing an international group to share knacks and lessons learned; maybe with short regular calls over Skype, say? Another paper in a few years to share disappointments (hard lessons) and triumphs should be attempted please too?

Consider one example here where we have acted collaboratively internationally as a community. In particular, official statisticians are now addressing, simultaneously in our micro data releases, the goals of greater transparency, greater data access, and enhanced data confidentiality. Consider the June 2014 issue of the sister Journal to JOS, the Statistical Journal of the International Association for Official Statistics (IAOS).

## 2. References

Association for official Statistics. 2014. Available at http://iospress.metapress.com/content/x18176857331/?genre=issue&issn=1874-7655&issue=current (accessed July 31, 2014).

Deming, W.E. 1986. *Out of the Crisis*. Cambridge: MIT Center for Advanced Engineering Study.

Deming, W.E. 1993. *The New Economics for Industry, Government, Education*. MIT, Center for Advanced Engineering Study.

Dickens, C. 1838. *Oliver Twist*, N.p.: Richard Bentley.

Herzog, T., F. Scheuren, and W. Winkler. 2014. *Data Quality and Record Linkage Techniques*, (2nd ed.). Springer.

Juran, J.M. and A.M. Godrey. 1999. *Juran's Quality Handbook*, (5th ed.). McGraw-Hill.

Turek, J. 2012. *Demystifying Microsimulation*. US: Department of Health and Human Services.

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 421–424, http://dx.doi.org/10.2478/JOS-2014-0023

# Discussion

*David Dolson*[1]

One of the major branding features for any producer of official statistics is the trust users can put in the quality of the statistics and information produced by the national statistical office (NSO). To that end NSOs as well as international coordinating bodies such as the European Statistical System, Eurostat, the International Monetary Fund and the United Nations Statistics Division have made management and statistical strategies to achieve high quality or fitness for intended use of their products a preoccupation for many years. A wide variety of useful reference documents have been produced by these organizations. In my references, I note a few that are particularly relevant to the current article. The United Nations Statistics Division Internet site is particularly useful since it in turn provides links to numerous other relevant sites and documents.

With their development of ASPIRE (A System for Product Improvement, Review and Evaluation) the authors have made a valuable contribution to the set of quality maintenance and improvement strategies available to producers of statistical information. The approach is well thought out, thorough and can be applied to great benefit within any statistics producing organization. Congratulations!

In this discussion I will highlight some of the major characteristics of the ASPIRE methodology and follow that by briefly describing a comparable program of Quality Reviews conducted at Statistics Canada. I will conclude by contrasting the two strategies with respect to their emphases, advantages and disadvantages.

## 1. ASPIRE

The ASPIRE framework and process are well described in the article; I include a very brief summary here for easy reference by readers of this discussion. ASPIRE emphasizes the accuracy dimension of quality and provides a systematic framework for addressing quality improvement in statistical programs and their products. Its main objectives are to identify important risks to product quality and areas where investment is needed to reduce risk and improve quality. This is done by application of a very structured and comprehensive rating of program efforts to reduce or manage risks. It leverages on total survey error principles and decomposes total error into its major components or sources and for each assesses risk to data quality using five evaluation criteria. After an extensive review of background material and meetings with the product team an evaluation team of independent external expert reviewers assign a rating on each evaluation criterion for each error source. The inherent risk for each error source is also assessed. A product's error source scores as

[1] Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6. Email: david.dolson@statcan.gc.ca

well as an overall score are derived. This then provides the basis for managers both to take decisions on where it is most important to invest effort into risk reduction (and, if successful, thus improving data quality) and, with repeat evaluations, to assess progress over time.

## 2. Quality Reviews

At Statistics Canada an organization unit called the Quality Secretariat was created in 2000 with a mandate to promote and support the use of sound quality management practices across the Agency. Starting in 2007 one of its major initiatives has been a program of Quality Reviews whose goals are broadly similar to those of the ASPIRE framework. However, the manner in which it is undertaken is somewhat different.

Each year a set of statistical programs is subjected to an independent internal assessment in which their practices to prevent erroneous data from being released are reviewed. In a first objective, risks are identified and assessed in terms of their likelihood of occurring and of their impact for the program and Statistics Canada if they materialize. While these risks and impacts are rated in a typical risk management framework there is no ASPIRE-like use of formal evaluation criteria and product error scores. Secondly, best practices that should be shared with other program areas are identified and recommendations are developed to address important residual risks to quality.

Programs for review are proposed by members of Statistics Canada's senior-most management committee. While some attempt is put to selecting programs across a range of areas, programs are also selected when it is strategically useful to do so. Good candidates for review include programs: about to undergo redesign; that have experienced quality issues or which have known vulnerabilities. Each year three to six reviews are conducted concurrently, all being coordinated by the Quality Secretariat. A separate review team, usually two people, is put together for each program to be reviewed. Reviewers are Statistics Canada employees at the middle management level and are assigned to review programs outside their current area of responsibility.

Reviewers conduct their review in a fashion much like that of the ASPIRE reviews. A summary of their findings is presented to the senior management committee and a more detailed report is delivered to the managers of the reviewed program. Copies are retained by the Quality Secretariat and are made available to other managers upon request. As well, information on the identified quality assurance risks and practices has been assembled together and made available to all employees.

In addition to obvious benefits to the reviewed programs, there are valuable benefits to the organization as a whole arising from the notion of sharing. The expertise of the various middle managers involved is shared to other programs and to the other participants in the reviews. In selection of programs and in initial kickoff meetings a strong emphasis is put on the positive nature of the undertaking and on improving quality by identifying and sharing of best practices, whether it be those of the reviewed program that may help in other areas or those of other areas that may help the reviewed program.

## 3. ASPIRE and Quality Reviews

These two strategies share similar goals – quality improvement in the products of statistical organizations. Either can constitute an additional element in an integrated

enterprise wide quality management program. Both achieve this via the integration of risk management and quality management concepts and strategies using small independent (more on this in a moment) review teams which consider the program/products under review, identifying strengths and areas of possible concern where action could or should be taken. Both primarily consider the accuracy dimension of quality. ASPIRE does this within a framework considering all quality dimensions while the Quality Reviews have the flexibility to be applied for other aspects of quality.

There are also some important distinctions.

ASPIRE provides a degree of rigour through its structure and comprehensiveness including formal evaluation criteria. This rigour helps ensure its robustness for use and consistent interpretation of findings across different products and in differing statistical organizations. Independent reviewers would be motivated to do so anyway but the ASPIRE rigour further helps ensure that reviewers are thorough and forthright in their evaluations. Although rigorous and clear in their governance and deliverables, the Quality Reviews proceed more from a best practices perspective and do not have the same extent of formal structure. The superior rigour and independent expertise of the reviewers in ASPIRE provides benefits externally for accountability and credibility that the Quality Review process cannot.

A very important element in these frameworks is the independence of the reviewers, both actual and perceived. Associated with this is the stature and expertise of the reviewers. ASPIRE achieves this by hiring external reviewers who are highly regarded experts in the domains of total survey error and quality management for statistical organizations. This conveys significant benefits. Their independence cannot be disputed and their authoritative standing can readily be influential and add value to the organization through the influence of high level expertise not currently available at the statistical office. However, such experts are not common and may not be readily available as needed by the statistical organization. Statistics Sweden has had the same reviewers for its first few ASPIRE rounds; this has helped ensure consistency in application of the process and in scoring. Now, ASPIRE has designed into it a robustness for inter-rater reliability but still I wonder about the challenges that may arise in the future when the review team changes or for an organization that cannot achieve the same degree of constancy in the reviewers.

The Quality Reviews differ. Reviewers are selected internally and different review teams are put together for each program. Clearly they cannot be as explicitly independent as the ASPIRE reviewers. Independence of these reviewers is addressed by ensuring they come from different organizational areas than the programs under review. Also very important in this regard is the Quality Secretariat's coordination and initial communications to reviewers concerning their role, their independence and expectations for forthright, honest and constructive evaluation. Over several years of Quality Reviews the Quality Secretariat has been very pleased with the degree to which these expectations have been fulfilled.

In selecting external experts, ASPIRE is potentially able to bring to bear new expertise and a degree thereof not available within the statistical organization. The internal reviewers used in the Quality Reviews provide knowledge and skills that are perhaps more fine tuned to the culture and business practices of the office. An important ancillary benefit of using internal reviewers is the training opportunity for the reviewers and the potential indirect improvements for the programs for which the reviewers are responsible.

Although ASPIRE can be applied more generally, it will perform to greatest advantage for recurring products that can be reviewed on repeat occasions. When done this way, as was the case for several products at Statistics Sweden, it will perform very well to assess progress against past findings and recommendations as well as to identify further opportunities for quality improvement. To date, Quality Reviews have not been used in this way but it would not be complicated to do so by implementing either repeat reviews or a process for reporting on progress on past review recommendations.

To conclude I would like to again congratulate the authors on their development and implementation of a great framework and process for quality improvement in the products of statistical organizations. Like the authors I also look forward to the experiences of other NSOs who implement ASPIRE or some other similar approach.

## 4. References

European Statistical System 2011. *Quality Assurance Framework of the European Statistical System*. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice (accessed June 16, 2014).

Eurostat 2011. *European Statistics Code of Practice – revised edition 2011*. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice.

International Monetary Fund 2003. *Data Quality Assessment Framework and Data Quality Program*. Available at: http://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm (accessed June 16, 2014).

Reedman, L. and C. Julien. 2013. "The Quality Assurance Reviews at Statistics Canada." In *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. November 4–6, 2013. Washington, D.C.

United Nations Statistics Division 2012. *Guidelines for the Template for a Generic National Quality Assurance Framework*. Available at: http://unstats.un.org/unsd/dnss/QualityNQAF/nqaf.aspx (accessed June 16, 2014).

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 425–429, http://dx.doi.org/10.2478/JOS-2014-0025

# Discussion

*Eva Elvers*[1]

## 1. Introductory Notes

First I would like to thank the Editors for inviting me to contribute to this discussion in the *Journal of Official Statistics*. Their motivation is that I am a person with long experience in Statistics Sweden but without direct involvement in the studied statistical products. Hence, my understanding of the ASPIRE model contains aspects that are both exterior (to the statistical products) and interior (to the organization). The comments are mine and not an official view from Statistics Sweden. They are a selection from my personal thoughts and experiences of the model as described in the article and of the related work at Statistics Sweden. Some of them are a bit provocative to stimulate the discussion.

The work has focused on accuracy and a key set of ten statistical products. These products are indeed diverse with registers and with surveys, including compilations, which use direct data collection, administrative data, other statistics, or a combination. Two of the results so far are: (i) the ASPIRE model, which has been presented at several international meetings and now in the *Journal of Official Statistics*, and (ii) three successive reports to Statistics Sweden with measures, comments, and suggestions. Moreover, and most important, the ten statistical products have made quality improvements. There have been further effects in the organization.

I will largely use the same terms as Biemer et al. and often without explanation. I will, however, use the term estimator (rather than estimate) in case of a random variable. Moreover, I will be a bit Swedish-oriented where I find the distinction important.

## 2. The Model, Its Ingredients, and Aims

The ASPIRE model involves a lot of information in an often condensed way. It is elegant in several respects. Fully used, all quality dimensions are included. The work for Statistics Sweden has focused on accuracy and so does the article.

### *2.1 A Common Understanding – Concepts and Terms*

Communication between external evaluators and statistical products being evaluated may not be easy. Concepts and terms that are used by the experts in their model should preferably agree with those already used in the organization being evaluated. In this case, with Statistics Sweden, quality concepts and terms of the European and the Swedish

[1] Statistics Sweden, Process Department, P.O. Box 24300, SE-10451 Stockholm, Sweden. Email: eva.elvers@scb.se

statistical systems could have been used. For example, a mapping between ASPIRE and Europe/Sweden could have been made to show differences clearly. The first evaluation period was intensive at a short notice, but further efforts towards a mutual understanding of concepts and terms could have been made successively. Some unnecessary confusion at Statistics Sweden – particularly around specification error and the Swedish template for quality declarations – could have been avoided in such a way.

### 2.2 *Quality Dimensions*

It was natural to start the evaluation at Statistics Sweden with accuracy. However, in my view Biemer et al. down-weight the other quality dimensions too much when they say in general: "To most statisticians and data analysts, good quality is synonymous with estimates having small mean squared errors (MSEs)". Coherence and comparability, for instance, are also important. Nor do I agree with the cited view on accuracy as "the dimension to be optimized in a survey while the other dimensions (the so-called *user dimensions*) can be treated as constraints during the design and implementation phases of production". Accuracy needs to be balanced with other quality dimensions, for example timeliness. So, further quality dimensions could have been considered together with accuracy to give a more complete picture.

### 2.3 *Error Risks Are Important and Deserve More Motivation*

Biemer et al. emphasize error risks, which are an important part of their model, and they use two types of risks. They emphasize the difference between *inherent* risk and *residual* risk, which refer to situations without and with the current efforts, respectively. They discuss risks, in principle and as part of the model. They say that risk involves both likelihood and impact. I have two major questions.

Residual risk is described in the text and it is assessed, but it is not really visible in the presentation through tables. Why is the residual risk not given a column in the matrix with quality ratings – does it not deserve a more explicit role in the model than now seen?

The Australian Bureau of Statistics, ABS, has a statistical risk assessment framework, where the building blocks are clearer to me. There are five levels for each of Likelihood and Consequences. These levels are combined, resulting in four levels of risk from low to extreme, as ABS (2010) describes. Would it be possible to clarify or expand ASPIRE in a similar, more explicit, way? Were there specific features of the Statistics Sweden management or statistical production process that would not have fit with the obviously different ABS quality approach (probably well known to the second author)? Does this possibly provide an indication of ways in which ASPIRE may need to be "tuned" to features of other statistical offices?

### 2.4 *An Informative Matrix – But it is Dangerous to Use Numerical Scores?*

A lot of information is condensed into the ASPIRE matrices. For instance, the matrix in Table 2 shows error sources and evaluation criteria together with risks to data quality and changes from the previous round. Symbols and fonts provide information in a compact way. This table gives a good overview of one statistical product.

However, I find the numerical scores over-simplified and a bit dangerous, for instance for evaluations and priorities for further work. Three simple examples (where the last one includes weighting):

- Biemer et al. give somewhat double messages stating both that the numerical scores "can be used for comparisons across time and products" and that "The interpretation of such comparisons may not be straightforward for several reasons".
- Biemer et al. suggest putting priority on the areas having highest risk and lowest ratings, if other factors are equal. That reasonable guidance is not obtained from the numerical scores alone.
- I heard about a case where the rating of one inherent risk was changed from middle to high. As a result of this higher risk, the overall quality score increased. This was found counter-intuitive for the quality level as such. – It is possible, of course, due to the use of a weighted average of scores.

I would like the authors to clarify the advantages of using numerical scores in addition to categories – and the disadvantages.

### *2.5 The Evaluation Criteria – Some Questions*

As an illustrative example, if the highest ratings for the two criteria Knowledge of Risks and Achievement Towards Mitigation or Improvement Plans are achieved, the MSE is known and under control for the primary purpose of the statistical product, according to Appendix A in the article. This is desirable, and with this achieved the other three criteria seem unnecessary. I would have liked to see more motivations behind the selection of the five evaluation criteria, which have all been included in the model as important factors for product quality.

In particular, I wonder about the criterion Communication with Users, which seems to refer to a one-way communication *to* users. The users will, of course, be able to use the statistics in a better way. Still, that communication does not influence the accuracy per se. Moreover, documentation, quality declarations/reports etc. belong to Accessibility & Clarity. However, a *bi*directional communication is likely to improve priorities and balances between quality dimensions and perhaps also between Accuracy components. Such balances are related to Relevance, too. I would have liked to see Accuracy together with other quality dimensions and components, as already indicated; and also to see process quality more clearly.

A further question relates to the situation where the statistical product is a register: how are the evaluation criteria formulated when it comes to MSE?

### *2.6 Accuracy and Total Survey Error With Decompositions*

The Swedish quality concept and quality declaration – with its hitherto descriptive listing of quality dimensions/components (Statistiska centralbyrån 2001, especially pp. 33–34) – uses the following sources of inaccuracy: sampling, frame coverage, measurement, nonresponse, data processing, and model assumptions. These components are fairly standard, except that "model assumptions" has a pronounced position. It is used for inaccuracy in addition to that from the other sources. The word *error* is deliberately

avoided, whereas Biemer et al. naturally use it in their decompositions of the total survey error in their formula (1) and the surrounding text.

When accuracy is measured by the MSE, both the estimator and the population parameter to be estimated (the target parameter with a Swedish term) play vital roles – and so does the random mechanism, which is not discussed here.

Concepts (specifications) are indeed essential. With a notation similar to that in the article formula (1), a simple survey with a collected/observed variable $y$ uses an estimator $\hat{Y}$ of the target parameter $Y$. There are other, more complex, cases where the collected variables and the variables of the target parameters do not have simple correspondences. Also, users may desire to interpret or use statistics with the target parameter $Y$ as if they were statistics with a somewhat different target parameter $X$. In my notation $X$ is not necessarily unobservable. The producer of the statistics then has to be clear about the ingredients of the presented statistics, including the target parameter: whether it is $Y$ or $X$. The accuracy of one and the same estimator differs, of course, between these two situations with different targets. The producer might prefer to use different estimators, though, in the two cases.

The statistical product Foreign Trade of Goods (FTG) is a bit complex with respect to variables, as Biemer et al. describe. Simply expressed, the collected invoice value, $y$, is converted into the target statistical value, $x$, with the aid of a specific sample survey that collects $x$ in addition to $y$. Hence, the target parameter $X$ is estimated by a direct estimator, which I would prefer to call $\hat{X}$ – not $\hat{Y}$ as Biemer et al. do. The inaccuracy caused by this FTG procedure (which includes the observed $y$ and an estimated conversion model) is put under the heading Model assumptions in the Swedish quality declaration. Biemer et al. instead use two error sources, Specification error and Model/estimation.

As for the article sentence "Some would argue that specification error should be part of the Relevance/Contents dimension", I would say – which is quite different – that the choice of target parameters has its place there, including that this choice influences the relevance of the statistics for a user with a particular interest. As for Accuracy, I already described my view, encompassing the MSE. Dissimilarities between what is observed, targeted, and desired – whether variables or parameters – will, of course, come into play somehow, depending on relationships, modeling etc. The statistical product FTG provides just one example.

Revision error is an unfortunate term, since the revision activity normally means an improvement, where one or more preliminary estimates are successively modified, arriving at the final estimate. Would it not be better to talk about revision size?

## 3 Some Concluding Remarks

### *3.1 The Words from Experts Are Heard*

Biemer et al. state some advantages with external evaluators in comparison with internal ones. I find it interesting and important to observe also that a suggestion made by an external expert automatically gets attention – more attention than the same suggestion made by somebody internally. It is more likely to be taken as crucial and to lead to activities. As an example, a development project on methods to assess measurement errors

started last year, partly because Biemer and Trewin emphasized this as an error source with high risk in many of the evaluated statistical products.

### *3.2 Avoid a Strong Person-Dependency*

In the short run, it is convenient to have the same evaluators in order to save time and to measure changes in a reliable way. There may still be difficulties, though, as Table 3 indicates: the evaluators have changed some of their own assessments from the previous year. Biemer et al. discuss, similarly, inter-rater variation and ways to reduce such variation. How strong is the current, remaining, person-dependency of the assessments in the ASPIRE model?

In the long run, there are advantages to have further expertise views. When the Scientific Advisory Board of Statistics Sweden discussed the ASPIRE model, there were warnings about measuring the same thing and using the same evaluators over time. The scope may then be narrowed to what is measured and to certain aspects.

### *3.3 Is the Model Essential?*

There are many benefits to Statistics Sweden from the work by Biemer and Trewin. I would say that some major reasons are their (i) expertise, structured discussions about quality, and ways to encourage and note improvements; in combination with (ii) high priority of this work at Statistics Sweden together with internal knowledge – knowledge that has become more visible and also expanded.

I cannot help wondering how essential the model is for the results achieved. Would the same conclusions and improvements have been made if Biemer and Trewin had chosen a different model or route for their work? My guess is "largely yes". This reflection should not be interpreted as a criticism of ASPIRE. It is rather a suggestion to reconsider some ASPIRE ingredients and priorities. Some examples are the numerical scores, the evaluation criteria, and clearer connections to process quality. Some guidance to quality in relation to costs would be interesting but is quite demanding. Such a reconsideration of the model might decrease the person-dependency and broaden the perspective.

In line with this, I would even say that it is useful for an organization to consider its model(s) for quality evaluations with some regularity. There may be good reasons to modify the focus and emphasize, or even add, new priorities. Also, statistical offices may learn from each other.

## 4. Reference

ABS 2010. Quality Management of Statistical Processes Using Quality Gates, Dec 2010, Appendix ABS Statistical Risk Assessment Framework. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1540.0Appendix1Dec%202010 (accessed June 23, 2014)

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 431–435, http://dx.doi.org/10.2478/JOS-2014-0024

# Discussion

*John L. Eltinge*[1]

## 1. Introduction

The authors have produced a very interesting contribution to the literature on evaluation and management of the quality of official statistics. This discussion presents some complements to, and possible extensions of, the Biemer et al. work. Section 2 highlights some contextual factors that may be useful in considering quality reviews. Section 3 outlines some possible extensions through a deeper assessment of some components of a total survey error model, through additional exploration of "user dimensions" of data quality, and through more direct linkage with stakeholder utility.

## 2. Context

Quality evaluation is an important part of the complex process of managing a National Statistical Office (NSO) to meet the information needs of a wide range of governmental and nongovernmental stakeholders. This management process must be grounded firmly in well-developed overarching principles, for example, as articulated in National Research Council (2013) and in related references cited by Biemer et al. However, the best implementation of those principles, including best practice in a quality review, is somewhat context-specific. Biemer et al. provide a useful description of the Statistics Sweden context in which they developed and applied their ASPIRE approach. In considering applications of similar approaches in other NSOs, it is worthwhile to consider a wider range of contextual factors. Here, we highlight three: origins, prospective changes and budgetary issues.

### *2.1. Origins of a Quality Evaluation*

The origins of a quality evaluation can have substantial influence on its scope and processes, on the preferred qualifications and external stature of its reviewers, on transparent and public communication of its results, and on the likely impact of its primary findings and recommendations. In some cases, quality reviews are well-established components of general management practice within a statistical organization, with relatively clear norms and expectations regarding the role, authority and responsibility of each participant; definitions and standards for specified quality components; and likely

[1] U.S. Bureau of Labor Statistics, Office of Survey Methods Research, 2 Massachusetts Avenue NE, Washington DC 20212, U.S.A. Email: Eltinge.john@bls.gov

remedial actions to be taken for components that are identified as problematic. In other cases, quality reviews may be introduced as part of a relatively new effort to improve institutional effectiveness. In still other cases, the decision to institute quality reviews may arise in whole or in part from a specific high-profile problem with quality. Under the latter two scenarios, norms and expectations may be quite uncertain at the start of the review process, and may require careful negotiation.

Also, in some forms of the latter two scenarios, quality problems can arise from deeper management issues. Examples include organizational structure; information flow; control mechanisms; and alignment of group or individual incentives with organizational goals. In addition, these problems can arise in part from perceived constraints associated with general market conditions; or associated with the larger governmental organization(s) that contain, or contract with, the statistical organization. It would be useful to explore the extent to which the ASPIRE system supports summary identification of these broader management problems, while avoiding excessive scope creep.

Finally, the origins of a given quality evaluation may have an important role in determining the specific standards that will be used in the full evaluation process. Of special interest is the extent to which the standards arise from internal consensus, or align with relevant external benchmarks.

### 2.2. *Prospective Methodological or Managerial Changes, and Related Incentives*

The authors' narrative and examples demonstrate a degree of linkage between the stated quality evaluation criteria and prospective methodological or managerial changes. Additional exploration of that linkage can provide valuable information about the subareas in which in-depth evaluation would be most productively focused. For example, consider a hypothetical survey for which preliminary evaluation indicated substantial problems with two error components, labeled B and C. In addition, suppose that management had a substantial degree of control over design factors that could lead to changes in component B, but little or no control over factors that could change component C. An evaluation report should highlight the problems identified for both components B and C; direct further analytic attention to feasible design or operational changes that could improve B; and note the need for future study of both the problems with C and the lack of control over its underlying factors. However, until reviewers have a clearer reading on feasible control mechanisms for component C within a specific NSO, more detailed examination of C may not be an effective use of limited evaluation resources.

Also, in considering prospective methodological and managerial changes to improve quality, it is worthwhile to distinguish between quality problems that lend themselves reasonably well to interventions that are sharply defined in scope and time; and other quality problems that can be attributed to a broader set of factors, may involve cumulative effects that have developed over several years or even decades, and thus may warrant a broader set of interventions. It would be of interest to study the ways in which one might "tune" the ASPIRE approach for these distinct sets of problems and prospective improvement efforts.

Finally, the Biemer et al. narrative reinforces the common observation that successful reviews of complex technical organizations depend on a substantial degree of cooperation

and engagement by the organization being reviewed. Similar comments apply to implementation of recommended changes. Consequently, it is valuable for all review participants to have clear positive incentives for free flow of relevant information, and for constructive engagement in sound implementation of recommended changes. Stated in decision-theoretic terms, structuring a quality review as a positive-sum exercise will increase the likelihood of long-term quality improvements. Conversely, misperceptions that quality reviews are punitive audits, or other forms of zero-sum or negative-sum exercises, will reduce the likelihood of strong and efficient quality improvements.

### *2.3. Budget Limitations and Cost Issues*

The authors identify another important contextual factor in their comment (Section 1, third paragraph), "NSOs world-wide are struggling to maintain high quality products as operating budgets continue to decline". It would be useful to study the ways in which budgetary and cost issues may inform the use of the ASPIRE system or similar approaches to quality evaluation. Three points may be of special interest. First, it can be important to explore the extent to which NSO managers have a clear understanding of cost-quality trade-offs, at a sufficiently fine level of granularity, that will help them respond to financial problems in ways that ameliorate the impact on quality. In addition, good information on cost-quality trade-offs may be important for decisions on implementation of quality improvements, per the comments near the ends of Sections 1 and 4.3 of Biemer et al.

Second, it is useful to distinguish among two related problems with budgetary and cost constraints:

(a) Constraints on the aggregate level of resources available to the NSO on a long-term sustainable basis; and related year-to-year or month-to-month uncertainties about availability of these resources.
(b) Legislative, regulatory or managerial restrictions that lead to increased aggregate costs, non-optimal allocation of available resources, or reduction in discretionary resources available for allocation to quality improvement efforts. One prominent example is the "stovepiping" phenomenon observed in many NSOs. This can restrict the flexible assignment of personnel with highly specialized skills, and may reduce the efficient flow of information. Other examples include restrictions on training of personnel in high-priority technical areas; and restrictions on the use of certain types of computing hardware and software. Also, in some cases, NSOs may receive external mandates that require re-allocation of scarce resources to activities that are not directly relevant to the central NSO mission, nor to improvement of the balance of quality, cost and risk of statistical products.

Areas (a) and (b) may require substantially different approaches to management of cost-quality trade-offs, and that in turn may affect institutional priorities on the specific types of analyses requested for quality evaluations.

Third, it would be useful to explore the prospective use of quality reviews in communicating with senior management and external decision-makers regarding the impact of both (a) and (b) on quality and on stakeholder utility.

## 3. Possible Extensions: Deeper or Broader Exploration of Some Dimensions of Quality

### 3.1. Deeper Quantitative Exploration of Error Risks and Comparison of Design Options

For many social and managerial processes, one may seek to measure important characteristics, and to develop related mathematical models, but one must also recognize practical limitations of these measurement and modeling efforts. In parallel with this general observation, Section 5 of Biemer et al. notes that although TSE models center on evaluation of quantitative error measures like MSE, "it makes no attempt to estimate the TSE and its components." Instead the ASPIRE model currently focuses on qualitative assessment of the "accuracy" component of quality. Also, as noted by the authors, comprehensive mathematical modeling of the properties of some complex statistical products can be very problematic. Nonetheless, it would be worthwhile to explore the extent to which additional mathematical structure may help to clarify some related issues and decision processes for the ASPIRE system.

For example, the authors' comments on "residual risk" and "inherent risk" in their Subsection 3.1 could lead to several avenues of additional research. Extending their qualitative conceptual development, one may define a vector $D$ that characterizes all relevant features of the methodological and managerial design; and then focus attention on a specific vector $D_C$ associated with a "current" design. In addition, one may consider a vector $D_B$ that represents a "baseline" design that complies with minimal normative standards, but is weaker than the "current" design. One could then extend the authors' ideas to define an aggregate total survey error $A = \hat{Y} - Y$; error cumulative distribution functions $F_C(a) = F(a|D = D_C)$ and $F_B(a) = F(a|D = D_B)$ for the current and baseline cases, respectively; and the corresponding mean squared errors $M_C = E(A^2|D = D_C)$ and $M_B = E(A^2|D = D_B)$. Under this expanded framework, one could use $\{F_C(\cdot), M_C\}$ and $\{F_B(\cdot), M_B\}$ to characterize in additional detail the authors' notions of "residual risk" and "inherent risk," respectively; and to link these risks with features of the two designs. For example, models for some components of $M_C$ and $M_B$ may provide additional information for some parts of the ASPIRE evaluation. Also, examination of $F_C(\cdot)$ and $F_{\boldsymbol{B}}(\cdot)$ may help in exploration of certain "error risks" that involve relatively extreme tail events that are not fully captured by mean squared error terms.

### 3.2. Broader Exploration of Constrained Optimization, Satisficing and Other Multi-Objective Approaches

As noted by Biemer et al. and many other authors, NSOs need practical approaches to address several dimensions of quality. Subsection 2.2 of Biemer et al. highlights the approach of Biemer and Lyberg (2003), which treats "accuracy as the dimension to be optimized in a survey while . . . user dimensions . . . can be treated as constraints. . . ." That can be a reasonable conceptual approach to balancing multiple quality objectives, but there are other conceptual approaches that also warrant consideration.

For example, in many cases there is no single established standard for a given "user dimension" of quality. In cases where it may be feasible to vary the standard for a given user dimension on an experimental basis, the resulting sensitivity analyses may help the

NSO to obtain a more nuanced understanding of the trade-offs between the "accuracy" and "user" dimensions of quality. It also may be useful to combine this type of sensitivity analysis with the direct assessment of user dimensions discussed in Subsection 4.4 of Biemer et al.

In addition, it appears that some NSO decision processes regarding quality may not approximate constrained optimization as such. Instead, NSO decision processes that involve multiple objective functions may share characteristics of minimax decisions or satisficing (in the original sense of Simon (1956), with emphasis on meeting certain minimum thresholds in each of several components of quality, cost and risk). It would be of interest to identify the practical circumstances, if any, under which these differing approaches lead to substantially different conclusions regarding quality assessments and recommendations for improvement efforts. Also, in resource-constrained settings, minimum quality thresholds can become de facto maximum quality standards, unless the responsible individuals and groups have clear incentives to improve beyond the previously established minimum levels. For such cases, it would be of interest to study ways in which ASPIRE or other quality-evaluation systems can address these issues to ensure further progress toward the stated goal of continual improvement.

### *3.3. Linkage With Value Conveyed to Stakeholders*

In closing, the literature on official statistics often emphasizes "fitness for use" as an overall quality criterion. However, that literature has placed less emphasis on detailed linkage between standard quality criteria and the value delivered to key stakeholders, or to the general public, through specific high-priority uses of particular published series or microdata files. Further exploration of that linkage may be very challenging and complex in some cases, but may be important in helping decision-makers understand the practical benefits conveyed by meeting a specific set of quality criteria.

## 4. References

National Research Council 2013. *Principles and Practices for a Federal Statistical Agency, Fifth Edition*. Committee on National Statistics, edited by Constance F. Citro and Miron L. Straf. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Simon, H.A. 1956. "Rational Choice and the Structure of the Environment". *Psychological Review* 63: 129–138.

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 437–442, http://dx.doi.org/10.2478/JOS-2014-0027

# Rejoinder

*Paul Biemer, Dennis Trewin, Heather Bergdahl, and Lilli Japec*

Our sincere thanks go to the discussants for their thoughtful, positive and, in some cases, critical comments. Collectively the comments provide many fruitful ideas for strengthening and improving ASPIRE as the system continues to evolve at Statistics Sweden and at other NSOs who may adopt ASPIRE wholly or partly. We are optimistic that ASPIRE can only improve if it is applied and the results, both positive and negative, are shared frankly and openly. Therefore, we fully endorse Fritz Scheuren's suggestion of establishing an international group to "share knacks and lessons learned" on ASPIRE and other similar approaches. As Statistics Sweden continues to apply ASPIRE, our intent is to continue to report our experiences at conferences, presentations, and in publications.

Fritz Scheuren is absolutely correct in referencing the great quality gurus Juran and Deming regarding ASPIRE. The authors took much inspiration from the work of these pioneering innovators in quality management. As Scheuren suggests, the application of ASPIRE at Statistics Sweden is already having a Kaisen effect in that incremental, continual quality improvement, as promoted by ASPIRE, is becoming engrained in the culture of the organization. Kaisen is definitely taking root there.

David Dolson's remarks clearly illustrate that ASPIRE is just one approach for possibly achieving similar objectives in an NSO. Statistics Canada's Quality Secretariat has been implementing a Quality Review program each year since 2007. Stats Canada's Quality Review program shares some similarities with ASPIRE. For example, like ASPIRE, their program attempts to identify the major risks to data quality and how to mitigate them across multiple programs. However, as Dolson notes, there are key differences. Reviewers are internal to the organization and there are no quality criteria nor are results reported in a numerical format. According to Dolson, the Canadian system does not possess the rigor and comprehensiveness of ASPIRE which, he believes, provides more robustness and greater consistency across products and greater comparability across organizations. However, one aspect of the Quality Review program that we may wish to adopt for ASPIRE is the emphasis on identifying and sharing of best practices across products, not only those under review but across all products in the organization. ASPIRE does this to some degree in its report to management of all products' ratings with their justifications. Also, the reports highlight a number of major "cross-cutting" issues which we know the Executive of Statistics Sweden has found most useful. However, ASPIRE tends to focus on the poorer practices. Drawing out best practices more emphatically and formally could be an important improvement for ASPIRE.

Dolson makes a number of excellent points in discussing the benefits and challenges of using internal and external reviewers. Unlike ASPIRE which uses the same two reviewers

for all products, Stats Canada assigns different, internal review teams for each program. Independence of these reviewers is addressed by ensuring they come from different organizational areas than the programs under review. Although the Stats Canada Quality Secretariat is pleased with the impartiality of the reviewers, we think review objectivity is a difficult attribute to assess and are skeptical that internal reviews are always objective in critical and sensitive situations. Statistics Sweden internal evaluators tended to report no concerns regarding product quality and few areas needing improvement. Quite a contrast to the ASPIRE findings. We agree with Dolson that the use of external reviewers would address any suspicion of partiality of reviewers and would enhance the credibility of the evaluation process.

At this point, we should note that ASPIRE is just one component of Statistics Sweden's quality management system. Because Statistics Sweden is ISO 20252 certified, all statistical products must meet these minimum standards. ISO 20252 provides a quality framework with requirements for numerous processes such as interviewer monitoring, keying, coding, and disclosure control. It requires an ongoing program of internal compliance monitoring which is performed by internal quality auditors who, like at Statistics Canada, are selected from outside of the department being audited. The purpose of these audits is simply to determine whether the ISO 20252 standards and guidelines are being appropriately followed. For this purpose, internal auditors can perform well with a good measure of objectivity. On the other hand, for ten of Statistics Sweden's most important products, ASPIRE strives to achieve a much higher level of quality than is ensured by ISO 20252 alone. As previously noted, attempts at Statistics Sweden to use internal evaluators for this higher purpose have not succeeded and, thus, the external evaluators were called in.

Eva Elvers provides a whole host of comments from someone who has experienced the ASPIRE process first hand at Statistics Sweden and can, therefore, draw upon her experiences with ASPIRE from within the organization. Many of her comments are largely about semantics; for example, the use of the word "error"; whether TSE includes specification error; when to use "estimate" versus "estimator" and issues with other terms we use that may differ slightly from the way some at Statistics Sweden would define them. However, our terminology is consistent with the TSE literature; for example, the term "error" has been used in this literature for more than 70 years.

The entire ASPIRE process including definitions, terms, criteria, and so on has been and continues to be thoroughly vetted at Statistics Sweden. For example, the ASPIRE evaluators have met with the survey methodologists at Statistics Sweden many times (both in Stockholm and in Örebro) over the course of three years to solicit their comments and suggestions on all aspects of the ASPIRE approach. There still remain some issues, particularly regarding terminology, where unanimity was not possible and it was necessary to form a consensus in order for the process to move forward. That there remain lingering questions in this area is neither surprising nor problematic, in our view. For the next round of ASPIRE (Round 4, which will commence in December 2014), we will continue these discussions that we are sure will be both fruitful and enlightening for all involved. Realistically, in any organization of highly intelligent and independent minds, there will always remain areas of disagreement and, thus, consensus must substitute for unanimity to make progress.

Thus, we will not attempt to address semantics here, opting instead to address three of Eva Elvers more substantive questions or comments as follows:

1. Does it make sense to treat user dimensions as constraints when optimizing Accuracy?
2. What are the advantages of using numerical scores?
3. What motivated the five quality criteria used?

Point (1) was initially proposed in Biemer and Lyberg (2003) and further expounded and illustrated in Biemer (2010). Essentially, maximizing accuracy while being constrained by the other quality dimensions simply means that the resources for maximizing accuracy are somewhat constrained by the survey budget once the budget necessary to meet the specifications for the other dimensions has been allocated.

For example, regarding Timeliness and Accessibility/Clarity, the survey design may specify that data collection for the survey should be completed within nine months, and that data files will be released to the public within 15 months. The design may further specify that data files will be provided for download online with full documentation at the time of release. For Comparability, methodologies used in previous implementations of the survey should be continued in the new implementation. Ideally, the survey budget should take into account these objectives in the allocation of resources for the survey.

Now let $C_T$ be the total budget for the survey and $C_U$ denote the combined, estimated costs for achieving the specified objectives for the user dimensions. The remaining budget (i.e., $C_A = C_T - C_U$) is the budget available to maximizing Accuracy. The task for the survey designer is to implement the data collection, data processing, weighting, and estimation phases of the survey to maximize Accuracy, while ensuring that survey costs do not exceed $C_A$ and the time from the start of data collection to the release of data files does not exceed 15 months. Thus, the design specifications for data collection, data processing, weighting, and estimation should ideally minimize TSE subject to these constraints on the total budget. This approach attempts to maximize the total survey quality once the design objectives and specifications under each dimension are set in accordance with both user and producer requirements.

As Biemer (2010) notes, the optimization strategy is likely an iterative process because the designer may realize that the budget, $C_A$, is inadequate to achieve an acceptable level of Accuracy. If additional resources are not available, then the user dimensions should be respecified in collaboration with users and the budget $C_U$ reduced as necessary to free up resources for Accuracy. Of course, the impact of the budget reallocation on the most important user quality dimensions should be minimized.

John Eltinge provides an excellent point related to (1) in Subsection 3.2 of his comments. He notes that there are no established standards for the user dimensions and the NSO may wish to experiment with alternative specifications of the user objectives to better understand the trade-offs among the quality dimensions as well as between $C_A$ and $C_U$.

With regard to (2), the Swedish Ministry of Finance directed that the presentation of the results of the quality reviews be concise, transparent, and accessible to administrators and stakeholders who are not familiar with the many, complex details of the statistical production process. The Ministry also placed priority on indicators that reflect quality improvements. Experience with ASPIRE has clearly demonstrated that the numerical

ratings and the graphical displays (particularly, the Harvey balls) satisfy these directives quite effectively. An obvious disadvantage of this simple approach is the risk of oversimplification. For example, a product's ratings may have improved from one round to the next for two high risk error sources, say A and B. However, the improvements for A may have much greater influence on overall data quality than the improvements for B. Of course, this information is not contained in the rating symbols. Digging into the details behind the improvements will reveal the true story, but that will require reading the report rather than simply relying on the ratings matrix.

Another risk of using numerical ratings is that staff may believe the goal is to improve the product's rating rather than to improve the product's quality. This is not necessarily a bad thing as long as improving the product ratings will result in real improvements in product quality. So far, ASPIRE has shown that improving quality will improve ratings and vice versa.

It is worth noting that ASPIRE can be easily customized for to suit the requirements of an NSO. It does not need to be applied precisely as described in the article. Indeed, at Statistics Sweden, there have been some important modifications through the first three rounds in light of experience and a few of these are described in the article. Elvers suggested a different structure for assessing risk. We are not convinced that the additional detail is needed but ASPIRE could easily incorporate this more complex risk assessment structure if it were deemed desirable.

Regarding (3), the five quality criteria were developed after numerous discussions among staff at Statistics Sweden and the evaluators. Together, we believe they span the scope of quality improvement attributes for most products. Knowledge of Risks seems an obvious starting point for quality improvement and its inclusion is well-supported in the literature. As an example, this criterion appears in the evaluation criteria for analytic reports published by the U.S. Office of Management and Budget OMB (2001, pp. 2–6). Further, as Deming famously said "Lack of knowledge . . . .that is the problem" (Deming, n.d.,). Communication with Users (two-way communication implied) is believed to also be essential for improving quality for two reasons: (a) users provide important knowledge about quality that can only be obtained through using the data and (b) users often will ramp up the pressure on an organization to improve quality for a specific product. Such pressure is often needed in organizations where there are few resources for quality improvements and many quality improvement needs. In Round 2, we added "Communication with Providers" (again two-way) to this criterion after realizing that providers of data for a product have a profound influence on product quality and need to be "kept in the loop" regarding how poor quality of the data elements they supply might affect overall product quality.

For quality improvement efforts to be effective, the necessary expertise should be available and applied to the product. Thus, Available Expertise is an important aspect of ASPIRE and may explain why progress on real quality improvement is lacking despite the substantial efforts and resource investments. At a minimum, product design and activities should comply with whatever standards are applicable including national or EU standards as well as the NSOs own standards. However, in ASPIRE, such compliance only rates "Good" on the five-point scale. Compliance with Best Practices raises the bar and is included in order to guide products toward practices that equal or exceed the state of the art

with regard to a particular error source. Finally, no improvements can take place without planning to improve and realize those plans. Therefore, the inclusion of Achievement towards Mitigation and/or Improvement Plans is an obvious and essential criterion that reflects real progress toward error risk reduction.

Elvers raises the question of whether all these criteria are needed. She asks: if a product rates a perfect score on criteria 1 and 5, are criteria 2 to 4 then superfluous? We think not. We believe a product would not be able to attain perfect scores on criteria 1 and 5 much less sustain them, without attending to the other three criteria. Communication with both providers and users, adequate expertise to address quality issues, and attention to standards and best practices are critical and essential attributes for achieving high quality.

She raises a good point regarding the evaluation of registers where the estimation of MSE components, which is ASPIRE's primary metric for estimators, does not apply. Registers, like data sets more generally, are comprised of rows and columns whose intersections create cells that contain values which may be either erroneous or missing. Rather than bias and variance, ASPIRE substitutes more appropriate metrics to describe the error in the register data; in particular, validity and reliability for gauging systematic and variable errors, respectively. These metrics can even be used to capture the error resulting from missing values if the missing values are imputed either using simple approaches such as mean imputation or more complex, model-based approaches, if available. Approaches for assessing the quality of register data are very much in a nascent stage and more work is needed in this area; nevertheless, we believe our classification of error sources for registers is a useful starting point.

We very much appreciate John Eltinge's further elaborations on some of the more challenging concepts in the article. Due to space, we limit our response to two important points that are particularly relevant and have not yet been touched on in this response. First, we agree with his comment in Subsection 2.1 that "quality problems can arise from deeper management issues." This is true for any organization and Statistics Sweden is no exception. Many of these problems relate to communication issues, collaboration barriers, questions regarding responsibility and authority and other problems brought about by organizational "stove piping" (as commonly observed in large-scale statistical organizations), complex management structures, and the ever-changing external environment. Naturally, in the course of conducting in-depth interviews with each product team, ASPIRE identifies such problems and it is completely in the scope of the review to report them to management. For example, in the Round 3 report we noted "a lack of co-operation between the National Accounts staff and data providers," also for "some statistical areas the need to improve the relationship between the IT department and their client areas", and "the lack of succession planning in some statistical areas." Issues of a more sensitive nature were conveyed orally to top management rather than in the written report and there were several of these as well.

Second, in Subsections 2.2 and 2.3, Eltinge rightly notes that it can be quite difficult for an NSO to determine the high risk and high priority areas to address when the budget is inadequate to address them all. An example of the hypothetical situation he posits is measurement error (error source B in his notation) versus household nonresponse (error source C in his notation). Particularly for the LFS, considerable resources have been directed to understanding the causes of nonresponse and reducing its effects on the

estimates. However, in terms of the "quality improvement per monetary unit," the return on investment (ROI) may be quite low relative to the ROI for measurement error for the same expenditure. Possibly redirecting even a fraction of nonresponse reduction resources towards understanding the causes and reducing the effects on measurement error on the estimates might result in a much greater ROI. Unfortunately, the data necessary to compare these two ROIs are often not available but could be obtained through appropriately designed evaluation studies. ASPIRE seeks to promote this view to counter the sentiment that response rates must remain high to ensure confidence in, and credibility of, the survey. Often, the latter view drives the decision to expend more and more resources to incrementally increase response rates, with little or no improvement in TSE.

Decisions on resource allocation for quality improvement are rightly the responsibility of management. We believe that ASPIRE assists them greatly in this important task by identifying those error sources with high risk with relatively low ratings.

We very much appreciate and value the comments of the four discussants and will continue to consider them as we move forward with ASPIRE. They contain many excellent suggestions and ideas for improving ASPIRE and, more generally, for developing better processes for statistical production. Thanks also to JOS for providing this forum and the journal space to fully discuss this important topic for NSOs world-wide.

## References

Biemer, P.P. 2010. "Total survey error design, implementation, and evaluation." *Public Opinion Quarterly* 74: 817–848. DOI: http://dx.doi.org/10.1093/poq/nfq058.

Deming W. Edwards. (n.d.). BrainyQuote.com. Available at: http://www.brainyquote.com/quotes/quotes/w/wedwardsd380788.html (accessed July 28, 2014).

Office of Management and Budget 2001. "Measuring and Reporting Sources of Error in Surveys". Statistical Policy Office, Working Paper 31. Available at: https://fcsm.sites.usa.gov/reports/policy-wp/ (accessed August 7, 2014)

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 443–457, http://dx.doi.org/10.2478/JOS-2014-0028

# Panel Attrition: How Important is Interviewer Continuity?

*Peter Lynn*[1], *Olena Kaminska*[1], *and Harvey Goldstein*[2]

We assess whether the probability of a sample member cooperating at a particular wave of a panel survey is greater if the same interviewer is deployed as at the previous wave. Previous research on this topic mainly uses nonexperimental data. Consequently, a) interviewer change is generally nonrandom, and b) continuing interviewers are more experienced by the time of the next wave. Our study is based on a balanced experiment in which both interviewer continuity and experience are controlled. Multilevel multiple membership models are used to explore the effects of interviewer continuity on refusal rate as well as interactions of interviewer continuity with other variables. We find that continuity reduces refusal propensity for younger respondents but not for older respondents, and that this effect depends on the age of the interviewer. This supports the notion that interviewer continuity may be beneficial in some situations, but not necessarily in others.

*Key words:* Longitudinal survey; multiple membership multilevel model; nonresponse; refusal.

## 1. Introduction: Interviewer Continuity

For longitudinal surveys, the perceived benefit of having the same interviewer assigned to sample members at each wave is a factor that can drive important aspects of survey planning and design. Many survey researchers believe that interviewer continuity – particularly for face-to-face surveys – brings benefits, primarily in terms of continued cooperation, though possibly also in terms of improved measurement. Consequently, they may sometimes be willing to prioritise the assignment of the same interviewer as at the previous wave, even when alternative strategies may be less costly or more convenient. For example, when a respondent moves home between waves the researcher may prefer to deploy the same interviewer even if he or she now has to travel 30 km to the address, rather than a different interviewer who lives only 5 km away. Considerations of interviewer continuity can also influence decisions about whether to award a survey data collection contract to the existing contractor or to an alternative bidder, as the latter scenario will

[1] Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK. Email: plynn@essex.ac.uk and olena@essex.ac.uk

[2] Centre for Multilevel Modelling, University of Bristol. Tyndall Avenue, Bristol, BS8 1TH, UK. Email: h.goldstein@bristol.ac.uk

**Acknowledgments:** The contribution of the first author was funded by UK Economic and Social Research Council (ESRC) Award no. ES/H029745/1, "Understanding Society and the UK Longitudinal Studies Centre" (Principal Investigator: Prof. Nick Buck). The contributions of the second and third authors were funded by the ESRC Survey Design and Measurement Initiative via a research grant to Prof. John Bynner for the project "Solving the problem of attrition". The third author is partly funded by the LEMMA node of the ESRC National Centre for Research Methods at the University of Bristol. We are grateful to NatCen Social Research for implementing the survey field work and providing the data.

typically result in considerably less, if any, interviewer continuity at the next wave. Therefore it is important for survey managers and survey commissioners to understand the value of interviewer continuity in order to make cost-effective decisions.

There are plausible theoretical reasons why interviewer continuity may reduce refusal propensity. These reasons relate to trust, tailoring and consistency.

Trust in the survey interviewer on the part of the sample member is an important influence on whether or not the sample member chooses to cooperate (Beerten and McConaghy 2003; Hox and de Leeuw 2002; Morton-Williams 1993). It is plausible that a sample member will, on average, trust a continuing interviewer more than a replacement one. This should occur if the sample member has experienced no negative consequences (such as crime or unwanted sales calls) of having previously invited this person into their home to interview them. Heightened trust, and therefore reduced refusal propensity, would thus be associated with interviewer continuity.

Tailoring of communication and tactics by interviewers reduces the chances of a refusal (Groves et al. 1992). A continuing interviewer is potentially able to draw upon prior knowledge of relevant characteristics of the sample member and his or her household that would not be available to a replacement interviewer. This additional knowledge could make the continuing interviewer better at tailoring both his or her calling patterns and the arguments that he or she uses to persuade the sample member to take part. This additional ability to tailor could therefore lead to continuing interviewers achieving both greater contact propensity and reduced refusal propensity (though the additional ability to tailor will be reduced if the survey organisation makes effective efforts to feed forward to the interviewer relevant information about the contact and persuasion attempts from previous waves).

Consistency is generally seen as a desirable personal trait (Cialdini 2008, chap. 3). After committing oneself to a position one should be more willing to comply with requests for behaviours that are consistent with that position. This is a likely explanation for the foot-in-the-door effect in surveys (Freedman and Fraser 1966; Groves and Couper 1998). A sample member who has previously agreed to an interview may be more likely to agree to a similar request in order to appear consistent if it is the same interviewer making the request. Thus a greater influence of the norm of consistency could result in reduced refusal propensity being associated with continuing interviewers.

However, although it is plausible that interviewer continuity might have the effect of reducing refusal rates, other things being equal, there is very little empirical evidence on this point. A number of longitudinal surveys observe that reinterview rates are higher amongst cases where the same interviewer makes the approach at a subsequent wave (e.g., Rendtel 1990; Schräpler 2001; Waterton and Lievesley 1987). But such an association does not imply causality. In particular, in face-to-face surveys where interviewers tend to work in specific geographic areas, it is quite possible that interviewer continuity and respondent cooperation rates have some common causes. For example, these may be associated with geographical mobility or employment mobility in the local area. A study which used more sophisticated analysis techniques found no effect of interviewer continuity on refusal rate (Pickery et al. 2001). To our knowledge, only one previous study has used a randomised design to attempt to assess the effect of interviewer continuity on reinterview rate on a face-to-face survey. This study involved an interpenetrated design at

Wave 2 of the British Household Panel Survey in 1992. No effect of interviewer continuity on reinterview rate was found either at Wave 2 (Campanelli and O'Muircheartaigh 1999) or at Waves 3 and 4 (Campanelli and O'Muircheartaigh 2002).

Aside from confounding effects of interviewer continuity with area effects, we note two additional limitations of previous studies of interviewer continuity. As far as we are aware, neither have been noted in the literature:

- Interviewer continuity is, by definition, associated with increasing interview experience. For example, those interviewers who interview the same respondents over three waves of an annual panel survey all have two years more interviewing experience at the time of Wave 3 than they had at the time of Wave 1. In cases where there is no interviewer continuity, replacement interviewers are therefore likely to be less experienced, on average, than continuing interviewers. Experience is known to be associated with reinterview propensity and should therefore be controlled in any study of the effect of interviewer continuity;
- The effect of interviewer continuity on reinterview propensity could be positive for some respondents (those who have a good rapport with their interviewer, perhaps) and negative for others (those with a poor rapport). Thus, regardless of whether or not there is a main effect of interviewer continuity, there may be an interaction of interviewer continuity with variables associated with rapport or 'liking' the interviewer. Identification of such interactions could be helpful for survey organisations faced with practical decisions about allocation of panel survey cases to interviewers.

In this article we examine the effect of interviewer continuity on refusal propensity using new experimental data. Our experimental design simultaneously controls continuity and interviewer experience. Additionally, our analysis considers interactions of respondent characteristics with interviewer continuity. We believe that these are two original contributions to the literature.

## 2. Study Design

The March–April 2008 round of the NatCen Social Research Omnibus Survey involved interviewing a random sample of the population aged 16 and over living in the United Kingdom. We shall refer to this survey as "Wave 1". Respondents who agreed to be recontacted for further research ($n = 1{,}188$) formed the sample for the study reported here. (Response rate was 55% to the Wave 1 survey and 78% of respondents agreed to be recontacted. However, we would note that inference in our study relies on random allocation within the sample who agreed to be recontacted, so we are not reliant on sampling-based inference.) Ample respondents were allocated to one of four treatment groups for a follow-up interview in March–May 2009 ("Wave 2"). The four treatment groups were:

- Same interviewer
- Different interviewer of the same grade
- Different interviewer of each of two different grades (grade was defined as a 3-category variable)

Thus the two control variables are interviewer continuity (whether or not the same interviewer is assigned to the sample case at both waves) and interviewer grade (in three categories). Grade indicates the position of an interviewer on the NatCen pay scale and therefore, as with any pay scale, tends to reflect a combination of competence and experience. We believe that interviewer grade is a good measure of the relevant characteristics that can differ between continuing and different interviewers in non-experimental studies, namely those aspects of ability that are associated with length of time working as an interviewer. This is because NatCen interviewers are promoted to higher grades based on a number of criteria, some of which are related to experience *per se* and others of which are related to performance. Thus grade would seem to capture the aspects of interviewer experience that are relevant to refusal propensity (organisational skills, ability to perceive the concerns and circumstances of respondents, ability to persuade). A low-grade interviewer is likely to have little experience, or could alternatively have more experience but not have performed very well. Of course, any association between interviewer experience and refusal rates could be due to either a selection effect (less successful interviewers quit interviewing) or a learning effect (interviewers become more successful over time as they gain new skills). Carton and Pickery (2010) find support for dominance of the selection effect. We do not address the cause of any association. Our intention is simply to control differences between continuing and different interviewers in characteristics that influence refusal propensity, regardless of the cause of those differences.

Allocation to treatment began by allocating each continuing interviewer to one quarter of his or her Wave 1 respondents. This was done at random except for three primary sampling units (very rural areas) where assignment to random subsets of respondents would have been prohibitively expensive. In these cases, respondents were chosen to be allocated to the same interviewer based on geographical location. Remaining respondents were then allocated to other interviewers of different grades, producing the distribution in Table 1. The effect of interviewer promotion between waves is shown in Table 2 and illustrates the importance of controlling interviewer grade. In total, 181 interviewers worked on Wave 1 of the survey, of whom 69 also worked on Wave 2. A further 136 interviewers worked only on Wave 2, meaning that overall 317 worked on one or both waves. Of these, 51% were female, 43% were aged over 60, 29% had no more than two years of experience as a NatCen interviewer, 52% had between two and ten years' experience, and 18% had more than ten years' experience.

Our key analysis variable is an indicator of interviewer change. We use two forms of this variable, a nine-category version and a three-category version (see results sections

*Table 1. Balanced sample design: interviewer continuity and interviewer grade*

| | Different Interviewer | | | Same Interviewer | |
|---|---|---|---|---|---|
| Number of assigned Wave 2 cases | Lowest grade 2009 | Middle grade 2009 | Highest grade 2009 | All grades 2009 | Total |
| Lowest grade 2008 | 97 | 117 | 131 | 115 | 460 |
| Middle grade 2008 | 114 | 100 | 105 | 115 | 434 |
| Highest grade 2008 | 73 | 75 | 69 | 77 | 294 |

*Table 2. Grades at each wave amongst continuing interviewers*

| Number of assigned Wave 2 cases | Same Interviewer: Lowest grade 2009 | Same Interviewer: Middle grade 2009 | Same Interviewer: Highest grade 2009 | Total |
|---|---|---|---|---|
| Lowest grade 2008 | 57 | 58 | 0 | 115 |
| Middle grade 2008 | 0 | 98 | 17 | 115 |
| Highest grade 2008 | 0 | 0 | 77 | 77 |

below for details of how these are used). The nine-category version is based on the twelve categories in Table 1, but a) combining to single categories all cases with a different interviewer of higher grade and all cases with a different interviewer of lower grade, and b) creating an additional category for cases with the same interviewer but of a higher grade (i.e., an interviewer who had received a promotion in the interim). The nine categories are listed in Table 4.

In the three-category version, the first category consists of all cases involving a different, lower grade, interviewer at Wave 2. The second category consists of cases involving a different interviewer of the same or higher grade. The third category consists of all cases allocated to the same interviewer at Wave 2. Comparison of the second and third categories will allow us to identify the effect of interviewer change, controlling for change in grade.

The Wave 2 interview was introduced as a survey about safety on public transport, consisting primarily of a module of questions on this topic that had been asked also at Wave 1. Sociodemographic and classificatory questions were also asked. Mean interview length was 21 minutes. Of the 1,188 issued sample cases, eleven were found to be ineligible for reinterview (deceased or moved out of the UK). Of the remainder, 844 were successfully interviewed, 119 were not contacted and 179 refused the Wave 2 interview. Other reasons for nonresponse accounted for the remaining 35 cases. Thus, amongst eligible cases, Wave 2 contact rate was 90% and cooperation rate was 80%, giving an overall conditional wave response rate of 72%.

## 3. Analysis Methods

Our analysis of refusal propensity is restricted to the 1,058 sample members who were successfully contacted at Wave 2, amongst whom the refusal rate was 17%. We use multiple membership multilevel logistic models of propensity to refuse conditional on contact. The dependent variable is coded 1 if the sample member refused the interview at Wave 2 and 0 otherwise. Thus, positive coefficients indicate an increased propensity for the undesirable outcome.

A formal statement of the basic model is as follows:

$$\text{logit}\left(\pi_{i(j_1,j_2)}\right) = X_{i(j_1,j_2)}\beta + w_{j_1}u_{j_1} + w_{j_2}u_{j_2}; \quad w_1 + w_2 = 1 \tag{1}$$

where $\pi_{i(j_1,j_2)}$ is the probability of a refusal for sample member $i$ interviewed by interviewers $j_1, j_2$ respectively at Waves 1 and 2 and the random effects are assumed

normal. Further details for such models are given by Goldstein (2011, chap. 13). In this model, conditional on the fixed effects in the model denoted by $X_{i(j_1,j_2)}\beta$, there are two random interviewer effects contributing to the response from Waves 1 and 2 respectively, namely $u_{j_1}, u_{j_2}$. The corresponding weights reflect the relative importance of the Wave 1 and Wave 2 interviewers. The overall interviewer effect is thus a weighted average of the two interviewers, or where there is no change in interviewer, simply the effect of that interviewer. We have chosen to assign the same Wave 1 weights to each Wave 1 interviewer and likewise for Wave 2. One of the aims of our analysis is to determine the relative weights which result in the best-fitting model (see below).

The multiple membership structure of the data arises from treating the interview occasions as Level 1 units and the interviewers as Level 2 units. This is not a standard two-level model since the Level 1 units, rather than being fully nested within each Level 2 unit (interviewer) with an associated effect from that interviewer, are influenced by a weighted average of the effects associated with both (if they are different) of the interviewers assigned to them. This is reflected in Model (1). The multiple membership model also differs from a cross-classified model where there are two sets of unrelated units (at occasion one and occasion two): treating our data that way would provide no way to differentiate cases where it is actually the same interviewer and where it is a different one at each occasion.

For model estimation we use Markov chain Monte Carlo (MCMC) estimation with orthogonal parameterisation and hierarchical centering with a burn-in length of 5,000 and 20,000 iterations implemented in MLwin 2.19 (Browne 2009; Rasbash et al. 2009).

Multilevel multiple membership models allow us to assign different relative weights to interviewers at Wave 1 and Wave 2. However, we are unable to determine the weights on *a priori* grounds. We are only aware of one previous study that considered the relative influence on Wave 2 participation of the Wave 1 interviewer and the Wave 2 interviewer. Pickery et al. (2001) found that the Wave 1 interviewer had a stronger influence on Wave 2 refusal propensity than the Wave 2 interviewer, though this conclusion was based solely on a comparison of coefficients from separate models, without any formal test. We therefore use empirical methods to select appropriate importance weights by selecting the model with best fit among the models with different weights. Our best fit criterion is to select the model with the smallest Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002).

As the random effect of interviewers turns out not to be significant (see Section 4 below), we do not test for fixed effects of interviewer change between waves or of any other interviewer characteristics. Instead, using the initial weights, we proceed to test random effects of twelve characteristics of respondents in order to establish whether interviewers vary in their relative success with different sample subgroups. These twelve characteristics represent all the sociodemographic variables available in the Wave 1 data for the full sample.

We test all categorical predictor variables (other than interviewer change) as dichotomies, as the model otherwise becomes overparameterised when we include interactions with interviewer change. Few of the variables are naturally dichotomous so combination of categories is necessary. This is done by fitting simple logistic regression models of refusal with the variable in question (full version) as the sole predictor variable, first combining categories with estimated coefficients that are not significantly different

from one another ($P > 0.10$) and subsequently, if necessary, combining categories with the smallest absolute difference in estimated $\beta$-coefficients until only two categories remain. In addition to the dichotomous predictors, we have one continuous predictor, age. The twelve resultant predictor variables are listed in Table 3.

For each predictor variable listed in Table 3, we first tested whether the variable had a random coefficient at interviewer level. Significance was judged in terms of whether the 95% interval estimate for a single parameter included zero. More generally, the DIC statistic was used to compare models where models differed in terms of two or more parameters. Retaining each significant variable, our intention was then to develop a full random effects model through backwards elimination, retaining only those predictors and their random coefficients which remain significant. However, as it turned out (see below) this step was not necessary as only one predictor variable showed significance.

When testing the significance of random slopes we use initial Level 2 weights of 0.5 for each wave, until we have identified the final model. We then fit that model with alternative combinations of weights and select the combination that results in the smallest DIC. Finally, we test interactions with the three-category interviewer change variable of each variable for which there is a significant random effect. We use the three-category version in order to retain sufficient statistical power to detect effects. Each of the interactions that is significant in these one-interaction models is then included in a combined model.

*Table 3. Predictor variables tested for interaction with interviewer change*

| Variable | Description | Coding (Ref = 0) | Number of respondents in category 1 |
|---|---|---|---|
| Sex | Sex | 1 = Female | 599 |
| Age | Age | Continuous | |
| Edu | Education level | 1 = Lower than first degree | 164 |
| Rdwell | Dwelling type | 1 = Flat (0 = house) | 168 |
| Rarea | Interviewer assessment of condition of houses in the area | 1 = Mainly good (0 = mixed or mainly poor) | 530 |
| Rhouse | Interviewer assessment of condition of house relative to other houses in the area | 1 = Same as or worse than other houses in the area (0 = Better than others) | 942 |
| Rmarried | Marital status | 1 = Single | 209 |
| Rnumadl | Number of adults in the household | 1 = 4 or more | 52 |
| Kids | Number of children in the household | 1 = 1 or more | 250 |
| Work | Whether respondent currently in employment | 1 = not working | 494 |
| Rent | Housing tenure | 1 = renting (0 = own outright or buying on a mortgage) | 294 |
| Disab | Whether respondent has a disability | 1 = no | 770 |

Note: Total number of respondents in the analysis is 1,058. Predictor variables were all collected at Wave 1 of the survey (and are therefore available for both respondents and nonrespondents at Wave 2).

## 4. Results: Interviewer Effect

We first fit a null model to test for a random intercept for interviewer combinations. The fit of this model is almost identical whether we specify the weights to be 1.0 for Wave 1 and 0.0 for Wave 2 (DIC = 873.0), 0.5 for each wave (DIC = 873.7), or 0.0 for Wave 1 and 1.0 for Wave 2 (873.3). By comparing the above models to a base model containing only a fixed-effect intercept (Model 1 in Table 5, DIC = 872.8), we note that adding a random interviewer combination effect does not improve the model fit. Also, the random effect (in each of the three above weight specifications) is not significant.

We therefore find no evidence of variation between interviewer combinations in propensity for a sample member to refuse. There is therefore no variation that can be explained by fixed characteristics of interviewers. To confirm this we fit a model in which the sole fixed effect predictor is the nine-category interviewer change variable. The fit of the model is slightly worse (DIC = 879.5) than the null model with only a fixed intercept (DIC = 872.8), and none of the coefficients for interviewer change reach significance (we tested all pairwise combinations of interviewer change and none was associated with a significantly different refusal propensity). The unweighted refusal rates for each interviewer combination are presented in Table 4.

## 5. Results: Random Effects of Respondent Characteristics

Though we found no evidence that interviewer combinations vary in their propensity to elicit a refusal, on average, it is possible that they may differ in the extent to which this propensity varies between sample members with different characteristics. We therefore test whether there is random slope variance associated with each of the twelve respondent characteristics listed in Table 3. We add each random slope in turn to the model which otherwise contains only the fixed intercept. For all respondents' characteristics other than age, the random slope variance is not significant (the mean of 20,000 MCMC parameter estimates is not significantly different from zero and the mode is zero to five decimal places). The only variable for which the random slope variance achieves significance is respondent age. DIC actually increases when the random effect of age is added to the model, but the covariance of age with the intercept is estimated to be 0.00, so we fix the covariance to zero, thereby reducing the number of parameters to be estimated. With the covariance removed, the random effect of age remains significant and DIC reduces.

*Table 4. Refusal rates by interviewer combination*

| | Refusal rate | *n* |
|---|---|---|
| Same interviewer: low grade | 19.2 | 52 |
| Same interviewer: medium grade | 7.8 | 90 |
| Same interviewer: high grade | 11.3 | 71 |
| Same interviewer: higher grade | 14.7 | 68 |
| Different interviewer, same grade: low | 15.2 | 79 |
| Different interviewer, same grade: medium | 10.5 | 86 |
| Different interviewer, same grade: high | 16.1 | 56 |
| Different interviewer, lower grade | 18.2 | 236 |
| Different interviewer, higher grade | 13.8 | 320 |

This suggests that interviewer combinations may differ in the extent to which they are relatively more (or less) likely to elicit a refusal from older (or younger) respondents. It is therefore of interest to know whether this variation can be explained by fixed characteristics of interviewers, notably interviewer change.

For the model containing a fixed intercept and a random slope of respondent age, we compare alternative assignment of weights to the two waves. We find that minimum DIC is achieved with weights of 0.25 for Wave 1 and 0.75 for Wave 2, suggesting that the Wave 2 interviewer has approximately three times as much influence on the Wave 2 outcome as the Wave 1 interviewer (Table 5). We use these weights in subsequent modelling.

## 6. Results: Interactions Between Interviewer and Respondent Characteristics

We next explore whether the variation between interviewers in the effect of respondent age on refusal propensity (significant random slope for respondent age) can be explained by known characteristics of interviewers, notably interviewer change. We therefore explore fixed-effect interactions between respondent age and interviewer characteristics. The three-category version of the interviewer change variable is used: a different interviewer of a lower grade, a different interviewer of the same or higher grade, and the same interviewer.

The interaction between respondent age and interviewer change does not reach statistical significance, though the model with this term added (including the respective main effects as fixed effects) is a better fit (DIC = 870.4) than the model with only a fixed intercept and a random effect of respondent age (DIC = 893.5). However, we can also explore the possible effects of other known characteristics of interviewers, namely age and sex. Specifically, we hypothesise that the random effect of respondent age may be related to interviewer age. Such an interaction could be driven by liking, whereby respondents are more likely to comply with a survey request from someone they like (Groves et al. 1992) and are more likely to like someone who is similar to themselves (Stotland and Patchen 1961), in this case in terms of age. Alternatively, the effect could be driven by

*Table 5. Comparison of models*

| Model no. | Fixed part | Random part | Weights (Wave 1 : Wave 2) | DIC |
|---|---|---|---|---|
| 1 | Intercept | None | 0.5 : 0.5 | 872.8 |
| 2a | Intercept | Respage | 0.5 : 0.5 | 867.7 |
| 2b | Intercept | Respage | 0.25 : 0.75 | 867.5 |
| 3 | Intercept<br>Intchg<br>Agedum<br>Intchg*Agedum | Respage | 0.25 : 0.75 | 856.9 |

Notes: Respage is respondent age in years; Agedum is a binary indicator of whether or not the respondent is aged over 60 (at Wave 2); Intchg is a five-category variable indicating whether the Wave 2 interviewer is a) same as Wave 1, up to 60, b) same as Wave 1, over 60, c) different, same or higher grade, up to 60, d) different, same or higher grade, over 60, e) different, lower grade. All models based on $n = 1{,}058$.

a tendency to show greater respect towards elders, which would suggest that younger respondents should be less likely to refuse to older interviewers.

We create a new five-category variable defined by interviewer change and interviewer age. This variable is created by subdividing both the cases with the same interviewer at Wave 2 and the cases with a different interviewer of same or higher grade into those where the Wave 2 interviewer is aged over 60 and those with a younger interviewer. The cases with a different interviewer of lower grade are not subdivided by interviewer age as this distinction is not of substantive interest (as there is no comparison group of same interviewers of lower grade). We also recode respondent age as a binary variable indicating whether or not the respondent is aged over 60. This is done to gain statistical power, and the cut point is chosen based on previous research that shows people of retirement age to be distinctive in terms of the determinants of survey participation (Lynn 2012 showed that people aged over 60 were more likely to agree to take part in an interviewer-administered survey, more likely to continue participating in a panel, and that their decision to take part was more likely to be sensitive to incentives to do so.) The sample contained 324 respondents aged over 60 and 734 aged 60 or under.

The interaction between respondent age and this five-category measure of interviewer change and age combinations includes significant differences (details in Section 7 below) and the model fit is significantly improved (DIC $=$ 856.9, compared to 867.5 in the model with only a random effect of age). We therefore retain this term in the model and proceed to test the interaction of interviewer sex with respondent age. This interaction is not significant and does not improve model fit. We also test the effects of interactions of respondent age with sex of Wave 1 interviewer and with age of Wave 1 interviewer, both instead of or as well as the interaction with age of Wave 2 interviewer. None of these interactions improve the model. Thus, we retain as our final model the model containing, in the fixed part, the interaction between respondent age (two categories) and the combination of interviewer age and interviewer change (five categories), plus a random effect of respondent age (continuous variable). This model is denoted Model 3 in Table 5.

## 7. Final Model

The final model is summarised in Table 6. To aid interpretation, Figure 1 displays the model-predicted propensities to refuse for each combination of interviewer continuity and respondent age (different interviewer of a lower grade is not shown, as this is not of relevance to the central theme of this article, as explained earlier). The model suggests that for sample members aged up to 60, interviewer continuity reduces the propensity for refusal if the interviewer is aged over 60 (left-hand panel in Figure 1). For sample members aged over 60, assigning an older interviewer reduces the propensity to refuse, regardless of whether or not it is the same interviewer who carried out the Wave 1 interview (right-hand panel in Figure 1). Specifically, for sample members aged up to 60, assignment of the same interviewer, aged over 60, results in a significantly lower probability of refusing than assignment of a different interviewer aged 60 or under ($p = 0.04$) or assignment of a different interviewer over 60 ($p = 0.03$). For sample members aged over 60, assignment of a different interviewer, aged over 60, results in a significantly lower probability of refusing than assignment of the same interviewer, aged

*Table 6. Final model of propensity to refuse*

| | Coefficient | Standard Error | |
|---|---|---|---|
| Fixed Part | | | |
| Intercept | −1.59 | 0.29 | ** |
| respondent age 60+ | −0.49 | 0.60 | |
| same interviewer 61+ | −0.83 | 0.46 | |
| different interviewer <61 | 0.00 | 0.34 | |
| different interviewer 61+ | 0.04 | 0.37 | |
| different interviewer lower grade in w2 | 0.23 | 0.35 | |
| same int 61+* resp age 60+ | −1.52 | 1.50 | |
| different interviewer <61* respondent age 60+ | −0.69 | 0.75 | |
| different interviewer 61+* respondent age 60+ | −2.07 | 1.02 | ** |
| different interviewer, lower grade in w2* resp age 60+ | −0.50 | 0.74 | |
| Random Part | | | |
| Level: combination of 2008 interviewers (35%) and 2009 interviewers (65%) | | | |
| var (intercept) | 0.147 | 0.172 | |
| var (age-gm) | 0.00119 | 0.00068 | |
| Model Fit | | | |
| DIC: | 856.9 | | |
| Units: interviewers (2009) | 227 | | |
| Units: respondents | 1058 | | |

Notes: Dependent variable is an indicator of whether the sample member refused to cooperate at Wave 2. Base is all sample members contacted at Wave 2. Reference category for respondent age is 60 or under. Reference category for interviewer change is the same interviewer, aged 60 or under.

up to 60 ($p = 0.03$). There is also a suggestion that continuity with an interviewer aged over 60 results in a lower probability of refusing than continuity with an interviewer aged 60 or under, though the difference is only of marginal significance ($p = 0.10$ for respondents over 60 and $p = 0.07$ for respondents 60 or under).

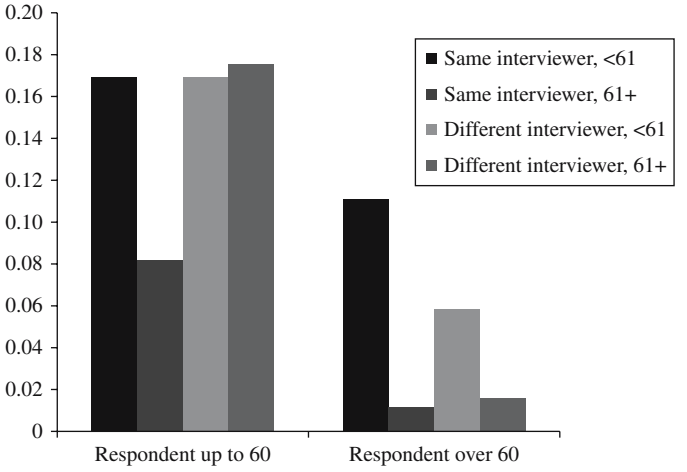



*Fig. 1. Predicted propensity to refuse, by interviewer continuity, interviewer age and respondent age*

It is interesting to note that the effect of interviewer continuity for younger sample members would have appeared larger if we had not controlled for interviewer experience. The difference in predicted probability of refusal between the same interviewer over 60 and a different interviewer of lower grade is even greater ($p = 0.01$) than the differences reported in the previous paragraph between the same interviewer and a different interviewer of the same or higher grade (of either age group).

## 8. Discussion

This experimental study has provided evidence of heterogeneous effects of interviewer continuity on cooperation by panel survey members. We believe it is the first study to find such evidence. Specifically, we find that continuity reduces refusal propensity for one sample subgroup (respondents aged 60 or under) but not for another (respondents aged over 60) and that this effect depends on a characteristic (age) of the interviewer. This supports the notion that interviewer continuity may be beneficial in some situations, but not necessarily in others. Whether interviewer continuity is beneficial may depend on the characteristics of the previous interviewer, the available alternative interviewers, and the respondent. What we conclude from this is that interviewer continuity should neither be blindly pursued in all cases nor completely ignored. Rather, survey organisations would be well advised to attempt to restrict the pursuit of interviewer continuity to situations where it is likely to matter. This can be thought of as an example of targeting of survey design features (Lynn, forthcoming).

We find that for younger respondents, interviewer continuity may only be beneficial if the interviewer is aged over 60. And in the case of older sample members, changing the interviewer may be beneficial if this involves switching from a younger to an older interviewer. The effect for younger respondents is intriguing, though the explanation is unclear. Maybe the trust of younger respondents is more likely to be engendered by older interviewers. Maybe older interviewers are generally better at tailoring but this only matters when the respondent is younger. Maybe younger respondents feel more strongly the need to appear consistent when the interviewer is older. Or maybe a greater positive age difference between interviewer and respondent engenders greater respect. The explanation of this finding requires further research.

Furthermore, we have demonstrated the importance of controlling for interviewer experience in studying interviewer continuity. We would have overestimated the benefits of continuity had we ignored experience, as changing to a less experienced (lower-grade) interviewer tends to increase the probability of a refusal.

It should be remembered that observed main effects of interviewer continuity are likely to mask a range of respondent-specific effects. Thus even if, for example, a switch to a different, lower-grade, interviewer reduces cooperation propensity on average, there may be some respondents for whom such a switch is neutral, or even positive. In other words, the effect may not be uniform across respondents. Our finding that the effect of interviewer continuity on refusal propensity differed between younger and older sample members is an example of such a nonuniform effect.

Our study is somewhat exploratory and some of the decisions we made in the course of the analysis were data driven rather than theory driven. For this reason, the specific

substantive findings should be treated with caution. Furthermore, our complex models require large sample sizes for good estimation. Other interactions between respondent characteristics and interviewer continuity may have become apparent with greater statistical power. Good measures of other relevant characteristics could also reveal other interactions. In particular, we would expect that the effect of interviewer continuity should depend on the rapport between respondent and interviewer and the extent to which the respondent likes the interviewer. Rapport and liking should depend on the combination of characteristics of respondent and interviewer, not merely the characteristics of the respondent. But in this study we had available only very limited characteristics of the interviewer. Furthermore, the available respondent characteristics may not be the most relevant ones. We suggest that future studies should consider measuring respondent personality and behavioural traits and preferences or, ideally, aspects of the respondent-interviewer interaction. Direct questions to the respondent regarding how they perceived the interviewer may provide the most powerful indicators of the likely effect of interviewer continuity. There are, of course, issues to be addressed in asking such questions. If they are administered by the interviewer who is the subject of the questions, there will be a risk of social desirability bias affecting the answers given (DeMaio 1984). Thus a confidential self-completion mode may be preferred for the administration of these questions. Aside from the mode in which the questions are asked, there is also work to be done to develop questions that effectively capture the extent to which the respondent is likely to be willing to be reinterviewed by the same interviewer.

We recognise that interviewer grade is not a perfect measure of the relevant concepts of experience or performance capability. There is an opportunity for future studies to benefit from attempting to measure more directly the qualities of an interviewer that determine success at making contact and gaining cooperation. Measures of experience might include numbers of cases worked, the period of time over which these cases were worked, and the variability in characteristics of those cases. Measures of competence might include input-adjusted outcome measures, such as response rates conditional on sample characteristics. Separate identification of experience and competence in future studies might provide insights into the mechanisms by which interviewer grade effects operate. This could assist sample allocation decisions.

This study was designed to identify the effects of interviewer continuity, not to explain the causes of such effects. We posited three possible causes: trust, tailoring and consistency. There is no particular reason why any of these causes should not apply more strongly to younger respondents than older respondents, or to older interviewers rather than younger ones. Thus, the identification of heterogeneous effects cannot assist us to identify the cause of the effects.

We cannot rule out the possibility that interviewer continuity effects are sensitive to the survey context. Our study is based on a request to take part in a relatively short interview (21 minutes) on a particular topic (safety on public transport). Results for a different type of survey request could be different. This issue could warrant investigation.

In conclusion, we have demonstrated that the effect of interviewer continuity on subsequent survey response may be rather more complex than has been implied by previous literature. The effect may depend on the interaction between characteristics of the previous interviewer, of the available alternative interviewers, and of the respondent. We

have found examples of such interaction. We have also demonstrated the importance of controlling for the effect of interviewer experience, of appropriate analysis methods, and of capturing interviewer characteristics. We believe there is considerable potential to learn more about the nature of interviewer continuity effects. This knowledge could help to reduce panel survey refusal rates in the future. But to gain this knowledge, further research would benefit from better measures of both respondent and interviewer characteristics, including interviewer experience and ability, and direct measures of the respondent's perception of his or her interviewer. In addition, randomised designs and appropriate analysis methods are needed.

## 9. References

Beerten, R. and M. McConaghy. 2003. "Respondents' Confidence in Survey Taking and Their Co-Operation With Government Surveys: Some Evidence From the UK." Paper presented at the *Annual meeting of the American Association for Public Opinion Research*, Sheraton Music City, Nashville, TN, May 2003. Available at: http://www.allacademic.com/meta/p116455_index.html (accessed December 13, 2013).

Browne, W.J. 2009. *MCMC estimation in MLwiN*. Version 2.10. Bristol: Centre for Multilevel Modelling, University of Bristol.

Campanelli, P. and C. O'Muircheartaigh. 1999. "Interviewers, Interviewer Continuity, and Panel Survey Nonresponse." *Quality & Quantity* 33: 59–76. DOI:http://dx.doi.org/10.1023/A:1004357711258.

Campanelli, P. and C. O'Muircheartaigh. 2002. "The Importance of Experimental Control in Testing the Impact of Interviewer Continuity on Panel Survey Nonresponse." *Quality & Quantity* 36: 129–144. DOI:http://dx.doi.org/10.1023/A:1014928107205.

Carton, A. and J. Pickery. 2010. "Interviewer (Non-)Response Performance Over Time." Paper presented at the *21st International Workshop on Household Survey Nonresponse*, Nuremberg, August 30.

Cialdini, R.B. 2008. *Influence: Science and Practice*. 5th ed. Boston BA: Prentice Hall.

DeMaio, T.J. 1984. "Social Desirability and Survey Measurement: A Review." In *Surveying Subjective Phenomena (Vol. 2)*, edited by C.G. Turner and E. Martia, 257–281. New York: Russell Sage Foundation.

Freedman, J.L. and S.C. Fraser. 1966. "Compliance Without Pressure: The Foot-in-the-Door Technique." *Journal of Personality and Social Psychology* 4: 196–202.

Goldstein, H. 2011. *Multilevel Statistical Models*. 4th ed. Chichester: Wiley.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R.M., R.B. Cialdini, and M.P. Couper. 1992. "Understanding the Decision to Participate in a Survey." *Public Opinion Quarterly* 56: 475–495. DOI:http://dx.doi.org/10.1086/269338.

Hox, J. and E. de Leeuw. 2002. "The Influence of Interviewers Attitudes and Behaviour in Household Non-Response." In *Survey Nonresponse*, edited by R. Groves, D. Dillman, J. Eltinge, and R. Little. London: Wiley.

Lynn, P. 2012. "The propensity of older respondents to participate in a general purpose survey." *Understanding Society Working Paper* 2012-03. Colchester: University of

Essex. Available at: www.iser.essex.ac.uk/publications/working-papers/understanding-society/2012-03 (accessed December 13, 2013).

Lynn, P. Forthcoming. "Targeted Response Inducement Strategies on Longitudinal Surveys." In *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. Abingdon, UK: Psychology Press.

Morton-Williams, J. 1993. *Interviewer Approaches*. Aldershot, UK: Dartmouth.

Pickery, J., G. Loosveldt, and A. Carton. 2001. "The Effects of Interviewer and Respondent Characteristics on Response Behaviour in Panel Surveys: a Multilevel Approach." *Sociological Methods and Research* 29: 509–523. DOI:http://dx.doi.org/10.1177/0049124101029004004.

Rasbash, J., F. Steele, W. Browne, and H. Goldstein. 2009. *A User's Guide to MLwiN Version 2.10*. Bristol: Centre for Multilevel Modelling, University of Bristol.

Rendtel, U. 1990. "Teilnahmebereitschaft in Panelstudien: Zwischen Beeinflussung, Vertrauen und Sozialer Selektion." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 42: 280–299.

Schräpler, J.-P. 2001. "Respondent Behavior in Panel Studies. A Case Study of the German Socio-Economic Panel (GSOEP)." *DIW Discussion Paper* 244. Berlin: DIW.

Spiegelhalter, D., N. Best, B.P. Carlin, and A. van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit (with discussion)." *Journal of the Royal Statistical Society, Series B* 64: 583–640. DOI:http://dx.doi.org/10.1111/1467-9868.00353.

Stotland, E. and M. Patchen. 1961. "Identification and Change in Prejudice and authoritarianism." *Journal of Abnormal and Social Psychology* 62: 250–256. DOI:http://dx.doi.org/10.1037/h0043040.

Waterton, J. and D. Lievesley. 1987. "Attrition in a Panel Study of Attitudes." *Journal of Official Statistics* 3: 267–282.

Received December 2012
Revised December 2013
Accepted December 2013

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 459–477, http://dx.doi.org/10.2478/JOS-2014-0029

# Item Nonresponse in Face-to-Face Interviews with Children

*Sigrid Haunberger*[1]

This study examined item nonresponse and its respondent and interviewer correlates by means of a population-based, panel survey of children aged 8 to 11 who were surveyed using standardised, face-to-face interviews. Using multilevel, logistic analyses with cross-level interactions, this article aims to examine which effects of item nonresponse are subject to children as respondents or to the interviewers and the interview setting. Depending on the type of question, we found different effects for respondent and interviewer variables, as well as interaction effects between child age/interviewer age as well as child gender/interviewer gender. However, interviewer variance is for the most part not significant.

*Key words:* Panel survey; interviewer effects; interviewing children; item nonresponse; multilevel logistic analysis.

## 1. Introduction

### *1.1. Focus on Children As Respondents in Social Research*

Today, children are seen as independent individuals in social survey research and no longer an ignored minority. Survey researchers interested in the growing-up, perspectives, attitudes, beliefs, and behaviour of children increasingly collect data from children themselves. Proxy reporting by parents or other caregivers is no longer seen as a suitable and satisfactory mode of data collection (Scott 1997). This is exemplified by the many child surveys where children's opinions and attitudes are collected using different modes of data collection and over a different period of time: for example, the Child Longitudinal Study (Germany), the Child Survey (Austria), the British Household Panel Study (Great Britain), the Young People's Social Attitudes Survey (Great Britain), the National Longitudinal Study of Children and Youth (Canada), the Child Development Supplement to the Panel Study of Income Dynamics (United States), and the European Longitudinal Study of Pregnancy and Childhood, to mention just a few. Large-scale assessments like PISA (Programme for International Student Assessment), PIRLS (Progress in International Reading Literacy Study) or TIMSS (Trends in International Mathematics and Science Study) are also worth noting.

Although survey methodology has been developed mainly for studies in adult populations, research has been done on adapting it for use with children and evaluating the influence of their cognitive growth on data quality (see, for example, Borgers et al. 2000,

[1] University of Applied Sciences Northwestern Switzerland, School for Social Work, Riggenbachstr. 16, CH-4600 Olten, Switzerland. Email: sigrid.haunberger@fhnw.ch

**Acknowledgments:** The author thanks the referees of JOS and the Associate Editor, for their stimulating and constructive comments.

Borgers et al. 2003). Item nonresponse in child surveys in general, and specifically in standardised face-to-face child interviews, however, has received only limited attention to date (see, for example, Borgers and Hox 2001, Fuchs 2008). Compared with self-completion questionnaires that target children as respondents, surveying children by means of interviews is of particular interest, because the interviewer and the interview situation may affect the young respondent's behaviour. We relate the frequency of item nonresponse on particular types of questions to the characteristics of the respondents (children), the interviewers, and the interview setting. The purpose of our main research is to discover whether and how child and interviewer characteristics as well as the interview setting affect item nonresponse in standardised, face-to-face interviews with children.

## 2. Past Research and Theoretical Framework

### *2.1. Past Research on Children As Survey Participants in General*

This section briefly reviews past research on children as survey participants in general. An analysis by Borgers et al. (2003, p. 91) examined the correlation of child characteristics and offering vague quantifiers and labelled response options with stability over time. They found, contrary to their expectations, that younger children did not have more difficulty with cognitively challenging questions. The older the child, however, the greater the stability was over time. Compared with younger children, older children can take greater advantage of fully labelled response options.

A methodological survey experiment on the effect of several question characteristics on the reliability of the responses conducted by Borgers et al. (2004) revealed no effects of negatively formulated questions on the reliability measures; the authors advised offering about four response options when children are respondents.

De Leeuw and Otter (1995) showed that a clear interaction existed between the age of children and the effect of ambiguous questions. Older children handled ambiguity much more easily than younger children.

Fuchs (2005, p. 701) examined several experiments on response order, question order, scale effects, and the effects of the numeric values associated with the response categories with children. His results indicated that younger and less educated children answered survey questions from a cognitively less ambitious perspective than adults did. In a later study by Fuchs (2008), the interviewer respondent interaction was videotaped, and all children underwent extensive cognitive tests. The results showed that younger children (ages 8–9) show considerably more problematic behaviours, suggesting problems in understanding and answering survey questions, than older respondents (ages 13–14).

### *2.2. Past Research on Item Nonresponse in Child Surveys*

This section briefly reviews the state of knowledge in the field of item nonresponse in child surveys. Borgers and colleagues (1999) investigated the influence of child characteristics and cognitive growth on data quality when surveying children by means of meta-analytic techniques. They found that gender and year of education influenced item nonresponse and internal consistency in a large number of different, multi-item scales. The hypothesis that data quality increases with cognitive growth was supported.

Furthermore, Borgers and Hox (2001) investigated the effect of item and personal characteristics on item nonresponse in written questionnaires used with schoolchildren. They found that item nonresponse is relatively rare, and the predicted response differences are relatively small. They concluded that young children do not perform as well as children who have been in education longer (they produce more item nonresponse), but their response behaviour is still satisfactory.

With a more qualitative, semi-standardised approach, Vogl (2011) recently explored the question-answer process in child interviews (ages 5–11). Focussing on 'don't know' responses, the results indicated fewer 'don't know' responses as children grow older due to their cognitive state; problems with the research instrument did not result in differences in the number of 'don't know' responses.

Another analysis of an adult survey by Shoemaker et al. (2002) used question sensitivity and cognitive effort to distinguish between 'don't knows' and refusals. They found that more sensitive questions attracted more refusals, whereas questions that require more cognitive effort received more 'don't knows'. Note that cognitive effort also correlated significantly with refusals. There is also evidence of item nonresponse in the event that adult respondents do not have adequately precise answers (Juster and Smith 1997), or as Fuchs (2008) reasoned, children might answer survey questions even if they have problems processing them.

To summarise past research, we can state that younger children are able to answer survey questions in an appropriate way if survey instruments are tailored to them. Nevertheless, as children grow older, their ability to answer survey questions and to handle ambiguity increases. This is also evident from the fact that item nonresponse declines with increasing age and/or year of education in all of the studies mentioned above.

### *2.3. The Influence of Interviewers on Item Nonresponse in Child Surveys*

In the special case of face-to-face interviews, the interviewer plays an important role in the question-answer process, even with children as respondents. Regarding item nonresponse in standardised, face-to-face interviews with adults, there is empirical evidence that interviewers are not neutral collectors of data but can influence the answers obtained (Pickery and Loosveldt 1998; Pickery and Loosveldt 2001). The interviewer can have a positive influence in reducing item nonresponse but may also induce item nonresponse (De Leeuw et al. 2003, p. 165). The results of comparisons of interviewer effects on factual and attitudinal questions in several studies are heterogeneous, with some of them finding that attitudinal measures are subject to higher interviewer variance. Greater effects for attitudinal questions have been found especially for questions with open-ended responses, emotionally charged questions, questions with difficult items (such as income or occupation), and questions that lack specification of an interviewing procedure (Groves 2004, p. 374). Findings on interviewer effects in adult surveys show that younger and less-educated interviewers have a higher level of item nonresponse (Huddy et al. 1997).

Item nonresponse is often the result of interaction between two sources of survey errors (Groves 2004), for instance the interaction between an interviewer and a respondent. Not much is known about how children react and behave face-to-face with a strange interviewer. There could be a huge social distance between young children and adult

interviewers. Therefore we assume that young children adapt their responses to the suspected expectations of adult interviewers and might have a tendency towards social desirability (De Leeuw et al. 2004).

### 2.4. *The Interview Setting: Presence and Intervention of Third Parties During the Interview*

The influence of third parties during the interview, especially parents, may bias responses from children in a positive or negative way: positively, if children are trying to answer the questions honestly and truthfully in the presence of their parents; negatively, if – especially as regards sensitive issues – the presence of parents or other persons leads to untruthful statements (Scott et al. 1995, p. 261; Reuband 1987). In general, the presence of third parties during standardised, face-to-face interviews is often undesirable, since researchers suspect there may be negative consequences for the question-answer process. Reuband (1984) reported a proportion of third parties during an average of about one third of adult interviews; similarly, Haunberger (2005) reported a high number of third parties present during standardised, face-to-face interviews with children, for the most part the children's parents (see Table 2 for details). Nevertheless, third parties may not necessarily act as a disrupting factor, but rather can exert a social control function and therefore contribute, especially in the case of factual questions, to more truthful answers (Reuband 1984).

### 2.5. *Asking and Answering Survey Questions: Cognitive and Communicative Processes*

The respondents' answers comprise a cognitive and communicative process (Schwarz and Sudman 1995). When answering survey questions, respondents must perform several tasks. First, they must interpret the question in order to understand what is meant, and second, they must retrieve relevant information. Third, they must integrate that information into a private opinion to finally formulate and edit a response (see Tourangeau et al. 2000 for details). This cognitive approach to the answering process shows that it is necessary to distinguish between different types of item nonresponse, which can have different causes and different meanings: Item nonresponse can easily occur when questions about events in the past are asked, or sensitive topics are probed, or when the questions are too difficult, uninteresting, too embarrassing, or too threatening.

Middle childhood (ages 8–11) has been referred to as a pathway to future (cognitive) development. In the early middle years of childhood, children gradually increase their logical thinking, memory, and learning strategies, and consolidate important academic skills such as reading and writing. In the later middle years of childhood, children gradually expand their ability to apply learned concepts to new tasks and are increasingly interested in learning life skills from adults (Kail 2011). Therefore, answering a survey question in middle childhood might be a particular challenge, because children's cognitive, memory, communicative, and social faculties are still developing.

### 2.6. *Towards a Theory of Item Nonresponse*

The model of the response process posited by Beatty and Hermann (2002; also see Groves et al. 2009) distinguishes between four levels of cognitive states regarding information

required by survey questions: *available*, *accessible*, *generatable*, and *inestimable*. The four states are ordered by the level of retrieved knowledge suitable for a question response. If the required information can be retrieved with minimal effort, the substantive response is *available* or *accessible*. If the required information is not known exactly, the substantive response is barely *generatable* or completely *inestimable*, resulting in item nonresponse. Therefore a hypothetical question should be inestimable as it is based on assumptions rather than facts.

### *2.7. Research Question and Hypotheses*

Given the background of the relevant research and theoretical assumptions, we want to investigate whether, and if so, how child and interviewer characteristics and the interview setting affect item nonresponse. For this purpose, several hypotheses have been developed.

Empirical evidence points to the fact that younger and less educated children produce more item nonresponses, leading to our first hypothesis.

*Hypothesis 1*: *With increasing cognitive functioning (measured by age and educational achievement), item nonresponse in standardised, face-to-face interviews with children will be reduced.*

In our next hypothesis, we specify a nondirectional premise, as research on interviewer characteristics influencing item nonresponse in standardised, face-to-face interviews with children is still lacking.

*Hypothesis 2*: *Interviewer characteristics will influence the impact of item nonresponse in standardised, face-to-face interviews with children in different ways.*

As we pointed out, we assume that it is primarily young children who adapt their responses to the supposed expectations of adult interviewers because of the huge social distance between them.

*Hypothesis 3*: *Cross-level effects between child and interviewer characteristics (especially age) will influence the impact of item nonresponse in standardised face-to-face interviews with children in different ways.*

Furthermore, we suppose that third parties during the interview act as mediators, especially for children in the presence of their parents trying to answer the questions honestly and truthfully, leading to our last hypothesis.

*Hypothesis 4*: *Third parties during the interview will influence item nonresponses in standardised, face-to-face interviews with children.*

## 3. Method

In this next section, the data, variables, and multilevel logistic analysis are introduced.

### *3.1. Data Set*

The data used in the analyses come from the Child Longitudinal Study conducted by the German Youth Institute. They are based upon a prospective longitudinal survey with two national, representative group samples in the following age groups: children in the last year of kindergarten (five-year-olds) and second-year primary school children. Children in the older cohort (and their parents) were interviewed in three survey stages at intervals of approximately 18 months.

*Table 1. Child Longitudinal Study, sample size*

| | 1st wave (2002) | 2nd wave (2004) | 3rd wave (2005) |
|---|---|---|---|
| Age group children | 8–9 | 9–10 | 11–13 |
| Interviewer (*n*) | 96 | 54 | 51 |
| Sample size (*n*) | 1,042 | 722 | 620 |
| Response rate | 50.58% | 35.05% | 30.09% |

Note: Gross sample $N = 2{,}060$

The first wave was conducted in the autumn of 2002, the second wave in the spring of 2004, and the third wave in the spring of 2005. As the study was not conducted for methodological purposes, a major drawback is that it was not possible to obtain measures of the interviewers' beliefs, expectations, and psychological characteristics or even to arrange an experimental setting. The sample size, the response rates and the number of interviewers for each wave are presented in Table 1.

### *3.2. Variables*

#### 3.2.1. Selection of Variables

In a first step, we calculated descriptive analyses for the whole data set to obtain a first impression of the distribution of item nonresponse, and subsequently made a preselection of variables. The following topic areas are addressed in the questionnaires: personality traits: 2 scales, the child's interests and activities: 3 scales, behaviour in conflict situations (with mother): 2 scales, school: well-being in school: 1 scale, parents' interest in school: 1 scale, achievement motivation: 1 scale, victims of violence: 1 scale, bullying: 1 scale, friends: child's network of friends: 1 scale, happiness with friends: 1 scale, behaviour in conflict situations (with friends): 1 scale, family climate: 1 scale, satisfaction with neighbourhood: 1 scale. All scales were asked over the three panel waves.

We found that the percentage of item nonresponse in this child survey is generally low, which creates two limitations for the selection of our dependent variables. On the one hand, we had to exclude questions with item nonresponse equal or less than 2 percent from the outset, on the other hand we were unable to follow the suggestion in the literature and distinguish between 'don't know' answers and refusals (Shoemaker et al. 2002). A separation between 'don't know' answers and refusals would have left too few cases for the analysis. Nevertheless, it was possible to select one scale from almost every topic area. This corresponds to a share of 40 percent of all scales in the questionnaire, which were used in the item nonresponse analyses.

#### 3.2.2. Linking Variables to the Model of the Response Process

In a second step, we linked the remaining variables with the Beatty-Hermann (2002) model of the response process. For self-description and leisure activities, we assumed that children were able to retrieve information with minimal effort (information available). The children's own achievement motivation, family climate, and behaviour in conflicts with friends could be retrieved with effort (information accessible) and represents a sensitive topic. Children might not have much knowledge of their parents' interest in school,

so information had to be estimated, resulting in a higher rate of item nonresponse (information generatable). We classified the question about the children's behaviour in hypothetical situations as inestimable.

In summary, our selected, dependent variables included questions about different topics. (See Appendix, Table A1 for question wording, response scale and percentage of item nonresponse per wave.)

#### 3.2.3. Recoding the Dependent Variables

In a third step, all dependent response variables were dichotomised, resulting in scales with the categories adequate responses (0) and item nonresponse (1). Remember that our category for item nonresponse includes 'don't know' answers as well as refusals.

The dependent variables vary considerably in question length, sensitivity, and response scales. Obviously, the highest item nonresponse rate was found for questions offering an explicit 'don't know' category (child's rating of parental interest in school, child's achievement motivation). In any case, the main purpose of this article is to clarify whether and how child and interviewer characteristics and the interview setting affect item nonresponse in standardised, face-to-face interviews with children, and not to explain the amount of item nonresponse due to different response scales.

### *3.3. Independent Variables*

The selection of the independent variables on the respondent and interviewer level was restricted due to the variables available in the existing data file and is largely based on the empirical evidence reported in Section 2.

#### 3.3.1. Respondent Variables (Children)

On the respondent level, we included the following variables in the multilevel logistic analysis (see Table 2 for details):

*Gender* (girl: 0, boy: 1), *age* (metrical, centred around the grand mean), *educational achievement* (mean: marks in mathematics, language, and reading, running from very good: 4 to fail: 1, centred around the grand mean), social and cognitive open-mindedness, self-efficacy (strongly disagree: 1 to strongly agree: 4) (both mothers' estimations).

*Interviewer rating: children's willingness to respond* (low: 0, high: 1), *open-mindedness* (low: 0, high: 1), *concentration skills* (low: 0, high: 1) and *language skills* (poor: 0, good: 1). Interviewers rated children's abilities after completion of the interview on a 6-point scale (very good: 1 to very poor: 6), which was dummy coded by the author afterwards.

#### 3.3.2. Interviewer Variables

On the interviewer level, we included the following variables in the multilevel logistic analyses, which were divided into two main categories (see Table 2 for details):

*Interviewer characteristics:* We applied a code indicating more than just one interviewer throughout the three waves to each response in each of the waves: *same or different interviewer* (different interviewer: 0, same interviewer in at least two panel waves: 1), *gender* (female: 0, male: 1), *age* (metrical; centred around the grand mean).

*Table 2. Child Longitudinal Study, independent variables*

| | 1st wave (2002) | 2nd wave (2004) | 3rd wave (2005) |
|---|---|---|---|
| **Respondent variables** | | | |
| Age (mean/sd)[1] | 8.5 (0.51) | 9.5 (0.51) | 10.5 (0.51) |
| Gender (boys) | 51% | 50% | 48% |
| Educational achievement (mean/sd)[1] | 1.73 (0.51) | 1.87 (0.54) | 1.95 (0.60) |
| *Personality traits* | | | |
| Self-efficacy (mean/sd) | 1.86 (0.47) | 1.86 (0.47) | 1.86 (0.47) |
| Cognitive + social open-mindedness (mean/sd) | 2.34 (0.45) | 2.34 (0.45) | 2.34 (0.45) |
| *Interviewer rating* | | | |
| Willingness to respond (good) | 81% | 87% | 90% |
| Open-mindedness (good) | 64% | 70% | 76% |
| Concentration (high) | 50% | 56% | 65% |
| Language skills (good) | 83% | 86% | 91% |
| **Interviewer variables** | | | |
| Age (mean/sd)[1] | 41.5 (11.5) | 48.7 (8.9) | 50.5 (8.3) |
| Gender (male) | 57% | 51% | 52% |
| Same interviewer (at least in 2 waves) | – | 50% | 50% |
| Presence of third: yes | 84% | 69% | 49% |
| Intervention of third: yes | 28% | 12% | 7% |
| Difficulties: yes | 8% | 5% | 4% |
| Interviewer length (mean/sd) | 42.7 (13.3) | 39.6 (15.0) | 45.8 (10.1) |
| Interviewer (n) | 96 | 54 | 51 |
| Sample size (n) | 1042 | 722 | 620 |

Note: Educational achievement in the original version (4 = fail, 1 = very good), [1]for multilevel logistic analyses centred around the grand mean

*Interview setting: presence of third parties during the interview* (no: 0, yes: 1), *intervention of third party during the interview* (no: 0, yes: 1), *difficulties during the interview due to a third party being present* (no: 0, yes: 1), *interview length* (metrical, in minutes).

### *3.4. Multilevel Logistic Analysis*

Multilevel analyses offer the best prospects to inspect interviewer effects on survey data because of the clustering of respondents by interviewers (Hox 2010).

In our case the use of a standard, two-level model would be inapplicable, since our dependent variables have binary outcomes: $Y = 1$ for item nonresponse, $Y = 0$ for response. With the software HLM 7.0 we specified a nonlinear analysis for binary outcomes. Therefore the binary outcome model uses a binomial sampling model and a logit link function (see Bryk and Raudenbush 2004 for details).

Before performing the multilevel analysis with panel data in HLM we reshaped the wide data files into long form, resulting in a pooled data set with 2,384 cases on each level. Level 1 missing data was automatically deleted when running the analyses. We controlled whether a correlation existed between the amount of item nonresponse in one panel wave per case and unit nonresponse in the following panel wave. We found no correlation,

which is not surprising given that the participation of the child is highly dependent on the participation of the parents.

After running the analyses, HLM offers different outputs (unit-specific model versus population-average model with robust standard errors). We present estimates of the population-average model with robust standard errors, since it is more appropriate for estimating the predicted population proportion and it is much less susceptible to misspecifications and distributional assumptions since it is based on generalised least squares estimation with robust standard errors (Zeger et al. 1988).

We present an example of model specification using the binary dependent variable 'family climate'. All respondent characteristics are included in Level 1 (see Equation 1). Subscripts $i$ belong to the respondents and subscripts $j$ to the interviewers.

Level 1 Model

$$\begin{aligned}
&\text{Prob}(\textit{family climate}_{ij} = 1|\beta_j) = \phi_{ij} \quad \log\left[\phi_{ij}/(1-\phi_{ij})\right] = \eta_{ij} \\
&= \beta_{0j} + \beta_{1j}^{*}(GENDER_{ij}) + \beta_{2j}^{*}(AGE_{ij}) + \beta_{3j}^{*}(EDUACHIEVEMENT_{ij}) \\
&+ \beta_{4j}^{*}(OPENMIND_{ij}) + \beta_{5j}^{*}(SELFEFFICACY_{ij}) + \beta_{6j}^{*}(WILLINGNESS_{ij}) \\
&+ \beta_{7j}^{*}(OPENMIND_{ij}) + \beta_{8j}^{*}(CONCENTR_{ij}) + \beta_{9j}^{*}(LANGUAGE_{ij})
\end{aligned} \tag{1}$$

Interviewer characteristics and characteristics of the interview setting are included in Level 2. We specified a random intercept model, since only the parameters associated with the constant vary across interviewers. The residual at the interviewer level can be denoted as $u$.

In order to better disentangle the effect of the child's gender and the effect of the interviewer's gender due to item nonresponse, we included cross-level interactions on Level 2 (for example: $\beta_{1j}$ represents the interaction between Level 1 variable 'gender of the child' and Level 2 variable 'gender of the interviewer') (see Equation 2).

Note that $\beta_{3j}$ to $\beta_{9j}$ represents the coefficients from Equation 1, without specifying an interaction effect.

Level 2 Model

$$\begin{aligned}
\beta_{0j} =& Y_{00} + Y_{01}^{*}(INT_SAME_j) + Y_{02}^{*}(INTGENDER) + Y_{03}^{*}(INTAGE_j) \\
&+ Y_{04}^{*}(INTLENGTH_j) + Y_{05}^{*}(THIRD\,PARTIES_j) + Y_{06}^{*}(INTERVENTION_j) \\
&+ Y_{07}^{*}(INTDIFFICULTIES_j) + u_{0j} \\
\beta_{1j} =& Y_{10} + Y_{11}^{*}(INTGENDER_j) \\
\beta_{2j} =& Y_{20} + Y_{21}^{*}(INTAGE_j); \\
\beta_{3j} =& Y_{30}
\end{aligned} \tag{2}$$

$$\beta_{4j} = Y_{40}$$

$$\beta_{5j} = Y_{50}$$

$$\beta_{6j} = Y_{60}$$

$$\beta_{7j} = Y_{70}$$

$$\beta_{8j} = Y_{80}$$

$$\beta_{9j} = Y_{90}$$

## 4. Results

Table 3 reports the results of the multilevel logistic analyses for item nonresponse in standardised, face-to-face interviews with children. For the random part, we included values for the interviewer variance ($u0j$) in the table, which corresponds to the intercept-only model. To increase interpretability of interactions, the value zero must be meaningful and actually occur in the data. For age and educational achievement we accomplished this by centring both variables on their grand mean. For gender, females were zero-coded (Hox 2010, pp. 63–68). In each column, we reported the coefficients, *t*-ratio and asterisks as indicators of the level of significance. We explained results for all analyses separately, referring only to results reaching the $p < 0.05$ level.

We first look at the interviewer level. Item nonresponse in the question about *self-description* is only explained by the variable indicating the same interviewer in at least two waves. If the interview was conducted by the same interviewer, this increased item nonresponse in the question about self-description.

On the respondent level the child's age, concentration and language skills affected the amount of item nonresponse. With increasing age, concentration and language skills, item nonresponse decreases. In addition, a significant interaction effect appeared between child gender/interviewer gender; meaning girls and female interviewers produced less item nonresponse in the question about self-description. However, the variance at the interviewer level is not significant.

On the interviewer level, only interview length affected the number of item nonresponses to the question about *leisure activities*. Increasing length of the interview correlates positively with more item nonresponse. We are not able to specify a cause and effect relationship, since we cannot clearly determine whether increased interview length led to more item nonresponses or whether more item nonresponses led to an increased interview length. On the respondent level, we found two significant effects: With increasing age, item nonresponse decreases. Children with good concentration skills produced more item nonresponses if they were asked about their hobbies. Taking a look at the interaction effect, the coefficient of child age/interviewer age is positive and statistically significant; meaning that with an increase in the age of the child and the interviewer, more item nonresponse occurs for this question. Again, the interviewer variance is not significant.

For *achievement motivation* we found only two significant effects at the interviewer level. Item nonresponses to the question about *achievement motivation* are due to the

*Table 3. Results for HLM non-linear models with the logit link function for item nonresponse in standardised, face-to-face interviews with children*

| | Self description | | | Leisure activities | | | Achievement motivation | | | Parents interest school | | | Family climate | | | Behaviour:conflict with friends | | | Behaviour: hypothetical situations | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Coeff. | T-ratio | Sig. | Coeff. | T-ratio | Sig. | Coeff. | T-ratio | Sig. | Coeff. | T-ratio | Sig. | Coeff. | T-ratio | Sig. | Coeff. | T-ratio | Sig. | Coeff. | T-ratio | Sig. |
| **Interviewer level** | | | | | | | | | | | | | | | | | | | | | |
| Gender: male | 0.05 | 0.18 | | 0.07 | 0.27 | | 0.26 | 1.57 | | 0.15 | 0.95 | | −0.08 | −0.21 | | −0.01 | 1.48 | | 0.09 | 0.38 | |
| Age in years | 0.02 | 1.86 | † | 0.01 | 1.01 | | **0.01** | **2.30** | * | 0.00 | −0.13 | | **0.07** | **4.63** | *** | 0.01 | 1.48 | | 0.00 | 0.59 | |
| Interviewer: same | **0.61** | **2.88** | ** | 0.22 | 0.98 | | 0.08 | 0.65 | | 0.00 | 0.04 | | **1.71** | **4.54** | *** | 0.40 | 1.83 | † | 0.23 | 1.07 | |
| *Interview situation* | | | | | | | | | | | | | | | | | | | | | |
| Presence of third Parties: yes | −0.18 | −0.83 | | −0.19 | −0.80 | | −0.14 | −1.08 | | −0.01 | −0.05 | † | 0.25 | 0.78 | | 0.00 | 0.01 | | −0.30 | −1.60 | |
| Intervention of third Parties: yes | −0.21 | −0.99 | | 0.22 | 0.99 | | 0.01 | 0.11 | | 0.05 | 0.38 | | 0.12 | 0.40 | | **0.67** | **3.27** | ** | −0.33 | −1.76 | † |
| Difficulties: yes | −0.31 | −0.83 | | 0.04 | 0.11 | | **0.59** | **2.75** | ** | 0.22 | 1.00 | | 0.33 | 0.77 | | **0.63** | **1.91.** | * | −0.26 | −0.64 | |
| Interview length In minutes | −0.01 | −1.24 | | **0.02** | **3.47** | *** | 0.00 | 0.13 | | 0.01 | 1.70 | | **−0.02** | **−2.02** | * | −0.02 | −1.75 | † | 0.01 | 1.25 | |
| **Respondent level** | | | | | | | | | | | | | | | | | | | | | |
| Gender: boy | 0.15 | 0.54 | | −0.56 | −1.83 | † | −0.01 | −0.03 | | −0.14 | −0.84 | | 0.30 | 0.79 | | 0.09 | 0.35 | | 0.03 | 0.11 | |
| Age in years | −0.45 | −3.69 | *** | **−0.44** | **−3.35** | *** | **−0.13** | **−1.96** | * | −0.11 | −1.72 | † | **−0.40** | **−2.37** | ** | 0.09 | 0.29 | | 0.06 | 0.59 | |
| Educational Achievement (mean) | −0.30 | −1.48 | | −0.19 | −1.02 | | **−0.34** | **−2.99** | ** | **−0.37** | **−3.24** | ** | −0.30 | −1.29 | | −0.09 | −0.50 | | 0.10 | 0.58 | |
| *Personality traits* | | | | | | | | | | | | | | | | | | | | | |
| Social open-mindendness (mean) | 0.21 | 0.84 | | −0.25 | −1.05 | | −0.03 | −0.21 | | 0.17 | 1.22 | | −0.43 | −1.48 | | **0.41** | **2.06** | * | −0.12 | −0.64 | |
| Self-efficacy (mean) | −0.06 | −0.26 | | 0.22 | 1.00 | | −0.21 | −1.59 | | **−0.44** | **−3.36** | *** | 0.07 | 0.24 | | −0.35 | −1.83 | † | −0.09 | −0.47 | |
| *Interviewer rating* | | | | | | | | | | | | | | | | | | | | | |
| Willingness to respond (good) | −0.39 | −1.59 | | −0.04 | −0.15 | | −0.08 | −0.42 | | **−0.36** | **−2.12** | * | **−0.82** | **−2.42** | ** | −0.25 | −0.86 | | **−0.56** | **−2.18** | * |
| Open-mindedness (good) | −0.36 | −1.60 | | −0.42 | −1.53 | | −0.02 | −0.11 | | −0.04 | −0.28 | | −0.42 | −1.27 | | 0.11 | 0.47 | | −0.29 | −1.16 | |
| Concentration (high) | **−0.46** | **−2.07** | * | **0.55** | **2.32** | ** | **−0.49** | **−3.62** | *** | **−0.28** | **−2.03** | * | -0.01 | -0.02 | | 0.03 | 0.15 | | −0.22 | −0.96 | |
| Language skills (good) | **−0.45** | **−2.03** | * | −0.30 | −1.03 | | −0.19 | −1.15 | | −0.22 | −1.38 | | **−0.87** | **−2.89** | ** | −0.27 | −1.06 | | 0.19 | 0.70 | |
| **Interaction effects** | | | | | | | | | | | | | | | | | | | | | |
| Child age/interviewer age | −0.01 | −0.78 | | **0.02** | **1.89** | * | −0.01 | −1.81 | † | 0.00 | −0.06 | | 0.01 | 0.83 | | −0.01 | −0.91 | | −0.01 | −1.01 | |
| Child gender/ Interviewer gender | **−0.81** | **−2.16** | * | 0.42 | 1.08 | | −0.21 | −0.93 | | −0.03 | −0.14 | | −0.57 | −1.14 | | 0.09 | 0.29 | | −0.08 | −0.23 | |
| *Random effects* | | | | | | | | | | | | | | | | | | | | | |
| Intercept $\gamma_{00}$ | −2.93 | | *** | −3.19 | | *** | −1.49 | | *** | −1.33 | | *** | −3.41 | | *** | −2.04 | | *** | −2.30 | | *** |
| Level 2 variance uOj | 0.30 | | | 0.43 | | | 0.15 | | | 0.37 | | | 0.49 | | | 0.22 | | | 0.26 | | |

Note: $*p \leq 0.05$, $**p \leq 0.01$, $***p \leq 0.001$, $\dagger p \leq 0.10$

interviewer's age. The older the interviewers, the more item nonresponses occurred. Difficulties during the interview led to increased item nonresponse.

Furthermore, we found three significant effects on the respondent level. Similar to the interviewer's age, older children produce more item nonresponse.

Poor academic performance produced more item nonresponse regarding the question of achievement motivation. Last but not least, with decreasing concentration skills, item nonresponse increases.

In regard to the question about parents' interest in school, we found no significant effect at the interviewer level and four significant effects at the respondent level. Poor academic performance produced more item nonresponse regarding the question of parents' interest in school. The more self-efficacy children have, the more meaningful responses will be produced. With decreasing willingness to respond and decreasing concentration skills, item nonresponse increases.

Turning to our next model, analysing *family climate*, we found three significant effects on the interviewer level: Item nonresponse increases with increasing age of the interviewer. If the interview was conducted by the same interviewer, item nonresponse increases. The shorter the interview, the more item nonresponse will be produced. We identified three significant effects at the respondent level: Item nonresponse increases with decreasing age of the children. The smaller the willingness to respond and the poorer the language skills, the more item nonresponse will be produced.

At the interviewer level, item nonresponse to the question about the *behaviour in conflicts with friends* is explained by two variables: interventions of third parties and difficulties during the interview leading to more item nonresponse.

At the respondent level we found one significant variable: The greater the social and cognitive open-mindedness, the more item nonresponse will be produced.

For the question about *behaviour in hypothetical situations*, item nonresponse was only affected by one significant variable on the respondent level: The greater the willingness to respond, the more meaningful responses were produced.

## 5. Conclusions and Discussion

The main aim of this article was to answer the question whether and if so, how child and interviewer characteristics and the interview setting affect item nonresponse in standardised, face-to-face interviews with children.

For this purpose, we used data from the Child Longitudinal Study conducted by the German Youth Institute, where children (ages 8-11) were interviewed using standardized interviews in three survey waves. To analyse item nonresponse, we selected questions that met two requirements: They had to cover substantial item nonresponse and should be compatible with theoretical guidelines. We computed multilevel logistic models with the software HLM 7.0 to better disentangle interviewer from respondent effects.

In Hypothesis 1 we tested whether item nonresponse in standardized, face-to-face interview with children would be reduced with increased cognitive functioning (measured by age and educational achievement). Our results support this hypothesis for the majority of the questions analysed. This is in line with other empirical evidence (Borgers et al. 1999; Borgers and Hox 2001; Vogl 2011).

In Hypothesis 2, we tested whether interviewer characteristics would influence the impact of item nonresponse in standardized, face-to-face interviews with children in different ways. We found that interviewers in child interviews are not neutral collectors of data, but detected no systematic pattern for item nonresponse due to interviewer characteristics. A closer look at the values of the interviewer variance turns out to be somewhat disillusioning: Not a single value showed significance. This means that in all models the between-interviewer variance is acceptably mild, so it could have been ignored and we could have used simpler, single-level statistical models (Hox 2010). However, for reasons of consistency we present hierarchical models. The nonsignificant variance could indicate that there might be other, more meaningful interviewer variables which have not been taken into account.

Concerning Hypothesis 3, we found two cross-level interactions between child and interviewer characteristics. Depending on the different types of questions, it seems that the effect of child age on item nonresponse was moderated by interviewer age in one question; the effect of child gender on item nonresponse was moderated by interviewer gender in another question. Given few interaction effects, there is only little support for this hypothesis.

Our fourth and last hypothesis tested whether third parties during the interview would influence the number of item nonresponses in standardised, face-to-face interviews with children. Contrary to our assumption that third parties would act as mediators and reduce item nonresponse, we found that the intervention of third parties increases item nonresponse in one of the questions. This is good news for surveys with children, because a third presence did not influence the question-answer process concerning item nonresponse for the majority of the questions analysed.

### *5.1. What Do These Results Mean for Surveys With Children?*

Overall, the amount of item nonresponse in the child survey was considerably low.

The highest item-nonresponse rate was found for questions offering an explicit 'don't know' category, though not necessarily for sensitive questions. This might mean that children aged 8–11 by and large perform well in face-to-face surveys.

Respondents' characteristics that correlate with item nonresponse are age and education. This may be an indication that interviewer training should focus more on how to deal with young and less-educated children.

The interviewers' rating of the child's ability to manage the interview points to concentration skills as an important factor. Item nonresponse increases with decreasing concentration skills, independent of age and education.

To improve the children's concentration, the survey researcher could vary the structure of the questionnaire by using a range of different question forms.

In the future, third parties will continue to be present during interviews with children. But this is good news for data quality, as their presence does not influence item nonresponse.

### *5.2. What Do These Results Mean for Survey Research on Item Nonresponse?*

In order to explain and predict item nonresponse, it is important to completely understand what happens during the question-answer process. Although a number of approaches exist (Krosnick, 1991; Tourangeau et al. 2000), we still lack a comprehensive theory explaining item nonresponse in surveys. Even Borgers and Hox (2001) conclude that they were not able to unequivocally confirm or reject Krosnick's satisficing theory. Furthermore, it is not

clear whether these approaches can be adapted to child surveys (first attempts by Vogl 2011). This also applies to the Beatty-Hermann (2002) model of the response process. We regard it as more of a heuristic than a verifiable theory. Therefore the present study did not aim to test the model in a strict empirical sense, but uses it as a helpful framework to rearrange our dependent variables.

The analysis of secondary data material has considerable disadvantages. First, there were only a limited number of interviewer characteristics available. Second, because of the small proportion of item nonresponse in general, we were unable to separate 'don't know' responses from refusals. Against the background of empirical evidence (see Section 2), we assume differences in the influence of interviewer and respondent characteristics on item nonresponse, broken down by these two categories.

More elegant ways to shed light on the question-and-answer process in standardised, face-to-face interviews with children would be experimental designs (first attempts by Fuchs 2008) or collecting reasons for item nonresponse and viable interviewer characteristics from the outset (De Leeuw et al. 2003).

## Appendix

*Table A1. Child Longitudinal Study, dependent variables*

| Question topic<br><br>Beatty-Hermann model of response process for item nonresponse | Percentage of item nonresponse per wave | | | Question wording | Response scale |
|---|---|---|---|---|---|
| Self-description available | (1)<br>7.0 | (2)<br>4.5 | (3)<br>2.4 | . . .15 items with which one can describe oneself; for example:<br>love to laugh.<br>I'm sometimes sad.<br>I like to scuffle.<br>et cetera | Four-point scale without 'don't know' option:<br>yes, rather yes, rather no, no |
| Leisure activities available | 6.3 | 2.6 | 3.2 | 6 items about things one can do alone or with others, for example:<br>playing game consoles,<br>make music,<br>go to the cinema<br>et cetera | Dichotomous scale without 'don't know' option:<br>yes, no |
| Parents' interest in school generatable | 25.6 | 17.8 | 18.3 | 6 items, for example:<br>Do your parents note school certificates and ratings?<br>Are your parents satisfied with your academic performance in general?<br>et cetera | Dichotomous scale offering 'don't know' option:<br>yes, no, don't know |
| Achievement motivation accessible | 17.9 | 20.3 | 20.7 | 6 items, for example:<br>Do you often have problems getting along at school?<br>Do you enjoy learning?<br>et cetera | Dichotomous scale offering 'don't know' option:<br>yes, no, don't know |
| Family climate accessible | 4.2 | 2.8 | 3.1 | 5 items about how one feels about the family:<br>I'm happy when my family is together.<br>We have got many conflicts in our family.<br>et cetera | Four-point scale without 'don't know' option:<br>always, often, seldom, never |

*Table A1. Continued*

| Question topic | Percentage of item nonresponse per wave | | | Question wording | Response scale |
|---|---|---|---|---|---|
| Beatty-Hermann model of response process for item nonresponse | | | | | |
| Behaviour: conflict with friends accessible | 10.1 | 9.0 | 16.2 | 13 items on reaction, if child has conflict with friends:<br>I roar with anger at him/her.<br>I leave so as not to be annoyed anymore.<br>I hustle, kick or beat him/her.<br>et cetera | Four-point scale without 'don't know' option: always, often, seldom, never |
| Behaviour: hypothetical situations inestimable | – | – | 9.7 | 9 items about evaluation of different hypothetical situations:<br>How good or bad are you at telling a child that he/she has done something that has annoyed you?<br>How good or bad are you at calling a new child to make an appointment with him/her?<br>et cetera | Five-point scale offering 'don't know' option: very bad, rather bad, ok, rather good, very good, don't know |

Note: Detailed questionnaires are available (in German only) on: www.dji.de/kinderpanel

## 6. References

Beatty, P. and D. Hermann. 2002. "To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse." In *Survey Nonresponse*, edited by R. Groves, D. Dillman, R.J.A. Little, and T.L. Eltinge, 71–86. New York: Wiley.

Borgers, N., E. de Leeuw, and J. Hox. 1999. "Surveying Children: Cognitive Development and Response Quality in Questionnaire Research." In *Official Statistics in a Changing World*, edited by A. Christianson, 133–140. Stockholm: SCB.

Borgers, N., E. de Leeuw, and J. Hox. 2000. "Children as Respondents in Survey Research: Cognitive Development and Response Quality." *Bulletin de Méthodologie Sociologique* 66: 60–75.

Borgers, N. and J. Hox. 2001. "Item Nonresponse in Questionnaire Research with Children." *Journal of Official Statistics* 17: 321–335.

Borgers, N., J. Hox, and D. Sikkel. 2003. "Response Quality in Survey Research with Children and Adolescents: The Effect of Labeled Response Options and Vague Quantifiers." *International Journal of Public Opinion Research* 15: 83–94.

Borgers, N., D. Sikkel, and J. Hox. 2004. "Response Effects in Surveys on Children and Adolescents: The Number of Response Options, Negative Wording, and Neutral Mid-Point." *Quality & Quantity* 38: 17–33.

Bryk, A.S. and S.W. Raudenbush. 2004. *Hierarchical Linear Models for Social and Behavioural Research: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

De Leeuw, E., N. Borgers, and A. Smits. 2004. "Pretesting Questionnaires for Children and Adolescents." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer, 409–429. New York: Wiley.

De Leeuw, E., J. Hox, and M. Huisman. 2003. "Prevention and Treatment of Item Nonresponse." *Journal of Official Statistics* 19: 153–176.

De Leeuw, E. and M. Otter. 1995. "The Reliability of Children's Responses to Questionnaire Items: Question Effects in Children's Questionnaire Data." In *Advances in Family Research*, edited by J.P. Hox, B.F. van der Meulen, J.M. Kanssens, J.J. ter Laak, and L.W.C. Tavecchio, 251–257. Amsterdam: Thesis Publisher.

Fuchs, M. 2005. "Children and Adolescents as Respondents. Experiments on Question Order, Response Order, Scale Effects and the Effect of Numeric Values Associated with Response Options." *Journal of Official Statistics* 21: 701–725.

Fuchs, M. 2008. "The Reliability of Children's Survey Responses: The Impact of Cognitive Functioning in Respondent Behavior." In *Proceedings of Statistics Canada Symposium*, Accessed March 15, 2013, http://www.statcan.gc.ca/pub/11-522-x/2008000/article/10961-eng.pdf.

Groves, R. 2004. *Survey Errors and Survey Costs*. Hoboken, NJ: John Wiley & Sons.

Groves, R.M., F.J. Fowler Jr,, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Haunberger, S. 2005. "Interviewer und Befragte im Kinderpanel. Interviewdauer und Panelbereitschaft." In *Kinderleben – Aufwachsen Zwischen Institutionen, Familie und*

*Freunden. Band 2*, edited by C. Alt, 285–316. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hox, J.J. 2010. *Multilevel Analysis. Techniques and Applications*. Second Edition. New York and Hove: Routledge.

Huddy, L., J. Billig, J. Bracciodieta, L. Hoeffler, P. Moynihan, and P. Pugliani. 1997. "The Effects of Interviewer Gender on the Survey Response." *Political Behavior* 19: 197–220. DOI: http://dx.doi.org/10.1023/A:1024882714254.

Juster, F.T. and J.P. Smith. 1997. "Improving the Quality of Economic Data: Lessons from the HRS and AHEAD." *Journal of the American Statistical Association* 92: 1268–1278.

Kail, R.V. 2011. *Children and Their Development*. Cambridge: Pearson Education.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. DOI: http://dx.doi.org/10.1002/acp. 2350050305.

Pickery, J. and G. Loosveldt. 1998. "The Impact of Respondent and Interviewer Characteristics on the Number of 'No Opinion' Answers." *Quality & Quantity* 32: 31–45. DOI: http://dx.doi.org/10.1023/A:1004268427793.

Pickery, J. and G. Loosveldt. 2001. "An Exploration of Question Characteristics that Mediate Interviewer Effects on Item Nonresponse." *Journal of Official Statistics* 17: 337–350.

Reuband, K.-H. 1984. "Dritte Personen beim Interview – Zuhörer. Adressaten oder Katalysatoren der Kommunikation?" In *Soziale Realität im Interview. Empirische Analysen und Methodische Probleme*, edited by H. Meulemann and K.-H. Reuband, 117–156. Campus: Frankfurt am Main.

Reuband, K.-H. 1987. "Unerwünschte Dritte beim Interview. Erscheinungsformen und Folgen." *Zeitschrift für Soziologie* 16: 303–308.

Schwarz, N. and S. Sudman. 1995. *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey-Bass Publishers.

Scott, J. 1997. "Children as Respondents: Methods for Improving Data Quality." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 331–350. New York: Wiley.

Scott, J., M. Brynin, and R. Smith. 1995. "Interviewing Children in the British Household Panel Survey." In *Advances in Family Research*, edited by J. Hox, B.F. van der Meulen, M.A.M. Janssens, J.J.F. ter Laak, and L.W.C. Tavecchio, 259–266. Amsterdam: Thesis Publisher.

Shoemaker, P.J., M. Eichholz, and E.A. Skewes. 2002. "Item Nonresponse: Distinguishing Between Don't Know and Refuse." *International Journal of Public Opinion Research* 14: 193–201.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Vogl, S. 2011. "Children Between the Age of 5 and 11: What 'Don't Know' Answers Tell Us." *Quality & Quantity* 46: 993–1011. DOI: http://dx.doi.org/10.1007/s11135-011-9438-9.

Zeger, S.L., K.-Y. Liang, and P.S. Albert. 1988. "Models for Longitudinal Data: A Generalized Estimating Equation Approach." *Biometrics* 44: 1049–1060. DOI: http://dx.doi.org/10.2307/2531734.

Received March 2013
Revised May 2014
Accepted June 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 479–497, http://dx.doi.org/10.2478/JOS-2014-0030

# Optimizing Opt-Out Consent for Record Linkage

*Marcel Das*[1] *and Mick P. Couper*[2]

This article reports on a study testing the effects of different ways of administering an opt-out consent for record linkage in a probability-based Internet panel. First, we conducted cognitive interviews to explore reactions to a draft version of the opt-out consent text. Second, we conducted a two-factor experiment to test the effects of content manipulations and mode. The results indicate that the way in which respondents were informed did not have much effect on opting out. Results from a follow-up survey on attitudes regarding privacy, confidentiality, and trust, along with knowledge questions about the process of linking, showed no evidence that presenting the opt-out consent statement makes respondents more concerned about privacy. Knowledge about the aspects of record linkage is generally not high. When looking at long-term effects of sending an opt-out consent statement, we found no evidence that this leads to higher attrition or lower participation rates.

*Key words:* Informed consent; administrative data; probability-based Internet panel.

## 1. Introduction

There is growing interest in linking survey data to administrative records, whether to enhance the quality and quantity of data available on sample respondents, to reduce the response burden, to compensate for missing survey data, or for other reasons (see, e.g., Calderwood and Lessof 2009). A key question in such record linkage is whether consent must be obtained from respondents and, if so, how best to do so in order to minimize refusals and any consent bias that may result.

In the Netherlands, Statistics Netherlands (SN) makes microdata available for statistical research. Customized administrative datasets can be prepared by SN so that they can be linked to the large number of available survey datasets. Legally authorized institutes, including universities and policy-oriented institutes, can be given the relevant authorization to access these data for analysis under the Statistics Netherlands Act.

CentERdata, a research institute housed on the campus of Tilburg University (the Netherlands), administers the Longitudinal Internet Studies for the Social sciences (LISS) panel, an online panel consisting of about 5,000 households, comprising 8,000 individuals.

[1] CentERdata and Tilburg School of Economics and Management, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: das@uvt.nl

[2] Institute for Social Research, University of Michigan, P.O. Box 1248, Ann Arbor, MI 48109, U.S.A. Email: mcouper@umich.edu

**Acknowledgments:** We are grateful to the Associate Editor and three anonymous referees for helpful comments and suggestions for improvement. We also thank Eleanor Singer and Annette Scherpenzeel for detailed comments on the setup of the experiment and follow-up survey, and on the first draft of the article. The LISS panel data were collected by CentERdata (Tilburg University, The Netherlands) through its MESS project funded by the Netherlands Organization for Scientific Research (grant number 176.010.2005.017).

The panel is based on a probability sample of households drawn from the population register by SN. All household members aged 16 and over are asked to join the panel and participate in the monthly questionnaires. Households that could not otherwise participate are provided with a computer and Internet connection. Survey data collected in the LISS panel can easily be linked to the administrative data available at SN, since the original sample was drawn by SN.

When linking survey data to administrative records one should obey the law with respect to confidentiality and privacy issues. The Dutch Data Protection Authority (Dutch DPA; www.dutchdpa.nl) supervises the fair and lawful use and security of personal data in the Netherlands. Rules are strict, but several exceptions hold for scientific research. There is no explicit legal requirement or rule to ask respondents (again) for consent to link their survey data to administrative records, given the fact – when joining the panel – they had consented that their data can be used for (purely) scientific purposes, and linking and analysis takes place in the secure environment at SN.

However, there is also an ethical issue. Institutes such as SN and the Netherlands Institute for Social Research (SCP) commonly use an opt-out version informing the respondent about the linkage of survey data to administrative records, giving the respondent the opportunity to object. We also decided to use an opt-out version when plans were first made to link survey data collected in the LISS panel to administrative data available at SN. However, before presenting the opt-out statement to the entire panel, we conducted an experiment using only a small sample of the LISS panel to optimize the opt-out consent for record linkage.

This article adds to the limited knowledge on how best to present opt-out consent statements. The objective is to identify the optimal wording for persuading respondents to consent to survey data being linked to administrative records, to remove *unfounded* fears or distrust, but also to ensure that respondents understand what they are consenting to. The article is structured as follows. Section 2 gives some background on consent to record linkage. Section 3 describes the data and methods. The section starts with a brief introduction about the LISS panel and continues with a description of the cognitive interviews (to evaluate a draft version of the wording), the experiment fielded in the LISS panel (to test the effects of content manipulations and mode), and follow-up survey (to measure the knowledge about the linking process and test the effects on attitudes). Section 4 presents the results. Based on the results of the experiment and follow-up survey, an opt-out statement was then presented to the entire panel. Section 5 discusses the implementation and effects on attrition and responding behavior in the long(er) run. Finally, Section 6 concludes with a general discussion.

## 2. Background

There are a number of concerns driving the research on consent for record linkage. A primary concern is the issue of nonconsent bias, that is, the extent to which those who consent to record linkage may be different from those who do not, thereby biasing the estimates derived from the subset of the sample with linked data (see Sakshaug and Kreuter 2012). A second concern is that low consent rates may limit the sample size for analysis, increasing the variance of the estimates. A third concern is that of whether the

consent was informed, that is, the extent to which the respondent's decision to consent to linkage was based on a clear understanding of what was requested. Our article focuses on the latter two concerns.

Much of the research on consent to record linkage focuses on consent requests administered by an interviewer that require an explicit response from the respondent, whether in writing (e.g., a signature) or orally (see, e.g., Sakshaug et al. 2012; Sakshaug and Kreuter 2012; Sala et al. 2012). However, because of the renewed interest in self-administered survey methods such as mail and web, whether as stand-alone modes or as part of mixed-mode data collection (see Couper 2011), increased attention is being paid to ways of obtaining consent for record linkage in such modes.

There is evidence that the process of obtaining consent – and particularly the administrative requirement to document such consent – may affect respondents' willingness to consent. For example, Singer (1978, 2003) reported that some respondents who may have been willing to participate in a survey were not willing to sign a consent form. Sala et al. (2013) provide an estimate of the negative effect of having to sign a consent form: 3.9% of the sample who consented to record linkage verbally refused to sign a consent form. Hunt et al. (2013) found that requiring explicit opt-in consent (in the form of a reply card) prior to a mail survey significantly reduced participation. Requiring the collection of identifying data (such as social security number (SSN) or identification number) to facilitate data linkage may further lower consent rates (see, e.g., Dahlhamer and Cox 2007; Bates 2005). Given the administrative burden of explicit consent – especially that of documenting the consent – and the concerns about low consent rates and nonconsent bias, researchers are exploring opt-out alternatives to the more traditional forms of explicit consent.

Writing in the context of medical research, Junghans et al. (2005, p. 1) stated: "Opting in is deemed ethically more defensible, as it relies on active participation of individuals, and some evidence shows that this is what patients expect. The opt-out method has come under scrutiny as it relies on both inertia and the moral assumption that most people are willing to help researchers in principle." In an experiment by Junghans et al. (2005), 510 patients were randomized to an opt-in (in which patients had to return the reply card or call if they wished to participate) or opt-out (in which patients had to the return the reply card or call if they did not wish to participate) procedure. In both cases, patients who had consented or had not objected were subsequently contacted by phone to make an appointment. Of those in the opt-in condition, 48% returned the card or called and made appointments and 38% were seen in the clinic, while in the opt-out group, 59% made an appointment and 50% were seen in the clinic, while 20% actively opted out.

Writing similarly about medical research, Vellinga and colleagues (2011) note that active or opt-in consent limits participation and introduces consent bias. They argue that "if risks for the participants are very low, an opt-out arrangement or passive consent is generally the most efficient procedure without violating the option of providing choice." (Vellinga et al. 2011, p. 1).

In the field of education, research on school-based surveys has compared active consent procedures (where written parental consent is required for participation of minors) to passive consent (opt-out; where parental permission is assumed unless they explicitly object). In general, passive consent is found to result in higher participation rates and may

reduce selection bias in such research (see, e.g., Anderman et al. 1995; Ellikson and Hawes 1989; Eaton et al. 2004; Fendrich and Johnson 2001).

The studies reviewed above are about consent to participate in research, not consent for record linkage. There is relatively little research in the survey world on opt-out consent alternatives, specifically with regard to record linkage. Two exceptions are from the US Census Bureau. In a telephone-based experiment of alternative consent wording, Bates (2005) found that asking for the last four digits of the social security number (SSN) increased consent rates over asking for the full 9-digit SSN (50.6% versus 36.8%) in an explicit consent request, but that framing the request as opt-out ("Do you have any objections?") with no request for SSN further increased consent to 63.4%.

Pascale (2011) conducted a similar split-ballot experiment in the context of a US Census Bureau telephone survey. Where addresses were available an advance letter was sent, which included an explanation of linking plans, and instructed respondents to inform the interviewer during the interview if they did not want their data to be linked. During the telephone survey introduction, these respondents were asked if they had received the letter. If they said yes, and did not inform the interviewer that they objected to linking, this was considered implicit consent. If they did not say yes, an explicit request for consent was made. These cases were then randomized to three conditions, one mentioning accuracy as a reason for linkage, a second mentioning cost, and a third mentioning time. Pascale (2011) reports that 38% of household respondents who participated in the survey gave implicit consent to link and were not asked the explicit linking question. Among those who were asked the explicit linking question, she found no significant differences in the consent rates between reasons for linkage (83.0% for accuracy, 85.3% for cost, and 83.6% for time). Combining the implicit and explicit consent, 90% of the sample consented to record linkage.

The sparse research that exists suggests that opt-out (implicit) consent procedures increase consent rates over opt-in (explicit) procedures. No research as yet has documented whether they also reduce consent bias. And virtually no research exists on whether respondents who consent implicitly understand that they have consented, and what they have consented to. Two exceptions to this are the studies by Ellikson and Hawes (1989) and Pascale and Mayer (2004). The first-mentioned study shows that parents who did not return their consent form when asked for explicit consent often did not intend to withhold consent, whereas those who failed to return the form in the implicit condition did not object to their children's participation. Pascale and Mayer find that among respondents who declined a request for consent to share the respondent's data with other household members during later waves, more than 80% misunderstood the request, many believing it was a request for a subsequent interview with other household members.

In summary, very little experimental research exists on the effects of different ways of asking for consent to record linkage, and there is even less research on what respondents understand by the request and whether the request changes attitudes regarding confidentiality and data sharing. This article represents an attempt to begin to fill these gaps, by focusing on 1) methods of informing respondents about opt-out consent options in an Internet panel and 2) measuring knowledge and attitudes following exposure to an opt-out consent request.

## 3. Data and Methods

In this section we describe the methods we used to learn about how and whether respondents understand opt-out consent requests, whether opt-out rates vary by mode and question wording, and whether the opt-out consent request had any impact on attitudes to privacy and knowledge about record linkage. We first start with a detailed description of the LISS panel which plays a central role in the experiment and follow-up survey we carried out.

### *3.1. LISS Panel*

The LISS panel is a representative sample of Dutch households who participate in monthly Internet surveys. The panel is based on a probability sample of households drawn from the population register with the help of Statistics Netherlands. Recruitment was carried out in a traditional way: First, an announcement letter was sent together with a brochure explaining the nature of the panel study. Second, an interviewer contacted the selected respondents by telephone or face to face, asking them to participate in a ten-minute interview. At the end of the interview the request to participate in the panel was made. Households that could not otherwise participate are provided with a computer and Internet connection.

Respondents *with* Internet access who consented to participate in the LISS panel received a confirmation by e-mail, as well as a letter with login code, an information booklet, and a reply card. Respondents confirmed their willingness to participate in the panel either by returning the reply card or via the Internet by using the login code provided in the letter. Respondents *without* Internet access confirmed their willingness to participate by returning the signed reply card, after which equipment and a broadband connection were provided. The confirmation procedure ensured double consent. For 63% of the total gross sample the contact person expressed willingness to participate in the panel at the end of the recruitment interview (first consent), while 48% of the total gross sample finally registered and started to participate in the monthly interviews (second consent). Detailed information about the LISS panel can be found at www.lissdata.nl or in Scherpenzeel and Das (2011).

The LISS panel follows changes in life course and living conditions and monitors trends in household composition, covering a broader range of topics and approaches than surveys typically cover. Panel members are provided with an incentive for each completed questionnaire. One member in the household provides the household data and updates this information at regular intervals. Researchers from the Netherlands and abroad are invited to submit research proposals, which after review and acceptance are fielded in the LISS panel free of charge. Data from the longitudinal core study as well as data from the individual research proposals are freely available for academic researchers. This yields an enormous amount of multidisciplinary data, and linking with administrative data increases the research opportunities even further.

### *3.2. Cognitive Interviews*

To evaluate a draft version of the text including the opt-out question, we conducted cognitive interviews with eleven respondents in the fall of 2010. These cognitive

interviews were intended to provide insight into the comprehensibility and persuasiveness of the draft text, as well as into possible points for improvement. The interviews took around 40 minutes on average and were held at the premises of TNS NIPO, a market research agency located in Amsterdam, the Netherlands.

For the purpose of the cognitive interviews we did not select respondents from the LISS panel. Instead, TNS NIPO selected the respondents from their database of persons willing to participate in surveys. The selection of respondents was stratified by age, gender, and education. The youngest respondent in the sample was a 19-year-old female, and the oldest respondent was a 68-year-old female. The education level ranged from primary school to university level. The interviews were monitored by researchers from CentERdata from a separate room. Respondents were aware of that, and all gave consent to this.

At the start of each cognitive interview the respondent was informed about the LISS panel, and was asked to imagine he or she was one of the LISS panel members. The interviewer explained that members of the panel have agreed to participate in monthly interviews, complete the questionnaires voluntarily, and receive an incentive for their participation. The topics of the questionnaires vary substantially, and the collected data are only used for scientific research. All cognitive interview respondents confirmed they understood the setting before reading the draft consent text.

After the introduction the interviewer asked the respondent to read the following text:

> *We guarantee confidentiality in all our studies. Your answers to the questionnaires are used for scientific purposes only. We always comply with the Personal Data Protection Act. We never provide information to public agencies such as the Tax Administration or the UWV* [Employee Insurances Implementing Agency].
>
> *For some surveys we complement the answers with information obtained from Statistics Netherlands (SN). This allows us to (better) answer research questions. The data are processed using secured computer systems. If you object to having your answers combined with SN data, please contact the CentERdata helpdesk: 0800 – 023 14 15 (or by e-mail: lisspanel@uvt.nl).*

When read, the interviewer asked for a first impression, whether they would opt out, and a series of follow-up questions to get a clear picture of whether the respondent understood the text.

### *3.3. Experiment in the LISS Panel*

In addition to the effect of level of detail in (or length of) a text about consent for linking, there is also the issue of which mode is best to use for communicating the information. As mentioned earlier, the LISS panel is an online panel for which (most) respondents have an e-mail account. In the cognitive interview we presented the text on paper. Instead of sending an e-mail to the panel members, sending the opt-out consent for record linkage by regular mail could be a better alternative. A letter may be more likely to be seen and/or read and it may appear more legitimate to respondents. On the other hand, sending the opt-out consent by mail may help to draw attention to the request, may increase respondents' suspicions and thus increase the opt-out rate. Moreover, it is more expensive and requires more effort in terms of logistics than sending an e-mail message.

To test the effects of content manipulations and mode we conducted a two-factor experiment in a random subsample of the LISS panel. For this experiment we randomly selected 500 households from the LISS panel. The two factors are: length of consent text (with short and extended text as levels) and mode of communication (with letter and e-mail as levels). We based the short text on that used in the cognitive interviews.

The body of the (Dutch) text contained 184 words. In the extended version we gave more examples, as well as (more) details on how the process of linking is carried out, how the linked data will be used by researchers, who exactly has access to the linked data, and where the linked data are physically stored. The body of the extended text contained 369 words. Translated versions of both the short and extended text are provided in the Appendix.

We randomly assigned the selected households to one of the four conditions. Within each household the experimental condition was the same for all members. We sent every household member (aged 16+) who participates in the panel an e-mail or letter personally. We sent the letters on February 14, 2011 and e-mails one day later, to ensure as far as possible that all respondents received the text on the same day. Respondents could object to linking their records by sending an e-mail message or calling the (toll-free) number of the CentERdata helpdesk. The helpdesk staff was trained to answer questions and alleviate concerns about record linkage. If the respondent preferred to opt out, the helpdesk always confirmed by e-mail that respondents' individual records will never be used in any linking with administrative data.

### *3.4. Follow-Up Survey*

The experiment described in the previous subsection focused on the effects of content and mode manipulations on opt-out rates. However, the objective was also to ensure that respondents understood what they were consenting to. Therefore, we included a series of questions in the monthly rounds of fieldwork two weeks after the panel members were informed about the linkage of their records to administrative data. The questions were about general attitudes regarding privacy, confidentiality, and trust, as well as attitudes toward survey organizations (part I) and some knowledge questions about the process of linking (to see whether they understood the opt-out statement; part II).

We included all 745 panel members who received an e-mail or letter in the sample for the follow-up survey. In addition, we randomly selected a separate group of 500 households in the LISS panel (containing a total of 776 panel members aged 16 and older). This latter group acted as a control group, to see whether exposure to the opt-out statement changed attitudes.

Part I replicated the questions that were used in Singer and Couper (2011). Both the experimental group and the control group received this part of the follow-up survey. With this part we tested the effects of presenting the consent statement on attitudes. The first question referred to personal privacy:

*Overall, how concerned are you about your (personal) privacy?*

We asked similar questions on concerns about violations of privacy rights by banks, credit card companies, tax authorities, research on public opinion (either by government or in

general), computers (storing large amounts of information), and continued confidentiality of information possessed by private and public organizations. Finally, we asked questions on trust that information gathered about the respondent is treated confidentially by three types of organizations.

Part II was only presented to the experimental group. The order in which part I and part II were presented to the respondents was randomized. Part II started with an introduction referring to the e-mail (or letter) that had been sent to this group previously. We then asked these respondents whether they recalled reading the opt-out statement. We presented the opt-out text in the survey (over several screens) to all respondents in the follow-up survey who indicated they had not read the e-mail (or letter), according to their original experimental treatment with respect to length of consent text (either short or extended). This was done to make sure that all respondents were informed about the record linkage before answering the knowledge question in part II. This question consisted of seven statements and was introduced as follows:

> *To better understand whether the explanation is perceived as clear or unclear, we present a few statements to you about the content of the [e-mail / letter]. Please specify whether you think each of the statements is true or false.*
>
> *It is not an exam. You do not have to review the [e-mail / letter] again.*

Exact formulations of the seven statements are presented in the Results section (Table 3). Respondents could answer with: true, false or don't know. On the basis of these seven statements we calculated a 'total score value' for each respondent, as an indication of how well the aspects in the process of record linkage are known. We constructed this score as follows. For each correct answer we assigned a score of $+1$, while for each incorrect answer we assigned a score of $-1$. For "don't know" answers, we assigned a score of 0. In this way the possible values for the total score value could range from $-7$ to $+7$, with a value of $-7$ for 'all answers incorrect' and a value of $+7$ for 'all answers correct'.

An alternative to the total score value could be a measure of perceived risk, with a score of $+1$ assigned if the answer the respondent gives indicates a higher perceived risk, regardless of the truth. For example, when a respondent believes name, gender, and date of birth will be sent to Statistics Netherlands, the perceived risk of disclosure is high. We assigned a score of $-1$ if the opposite holds, and as before we assigned a score of 0 to a "don't know" response. We defined the resulting sum score as a measure of perceived risk of disclosure. Again, this measure potentially ranges from $-7$ to $+7$.

## 4. Results

### *4.1. Cognitive Interviews*

Based on the cognitive interviews, the following conclusions were drawn:

- The text does prompt respondents to consent;
- The text creates both trust and confusion;
- No one understands what they are consenting to.

Respondents generally took a trusting approach. They assumed that 'it should be fine', and the first reaction to the text was positive. Respondents perceived the message as providing information, rather than inquiring. The majority assumed that one could contact the helpdesk to ask questions, but did not grasp that panel members could also do so to register their objection to record linkage. While most respondents said they would immediately consent, a few expressed uncertainty.

The respondents perceived the first paragraph as clear and persuasive. Some fears or distrust only emerged after the text (and particularly the second paragraph) was read for a second time, and following further probing. Respondents took the most important message to be that anonymity is assured, but the intended message (linking the survey data to administrative data) did not come across clearly. Some thought that their data records would be linked to similar data records in the registers from other persons, "to get a more complete picture". A similar finding is described in Gray (2010), based on cognitive interviews with respondents in preparation for the seventh wave of the Family and Children study.

Virtually no one in the cognitive interviews understood exactly what they were consenting to. This was mainly due to the formulation used in the text and the incomplete information about the different steps in the process of linking data. However, as was argued by Singer et al. (1992), providing more details and lengthy explanations might arouse respondents' suspicions rather than alleviating their concerns. Furthermore, it does not necessarily increase the respondents' willingness to respond. In the context of the collection of paradata – data about the process of data collection – Singer and Couper (2011) found that providing more information about the paradata reduces the willingness to participate in the research and, more substantially, the willingness to permit use of the paradata collected.

On the basis of the results of the cognitive interviews we fine-tuned the text for the experiment, and added some more information about the linking process (including an example) in the second paragraph (see Appendix).

### *4.2. Experiment in the LISS Panel*

In total we sent 745 individuals (in 489 households) the opt-out consent text. At the time of sending out the letter and e-mails, we excluded eleven individuals (from eleven single-person households) from the initial selection because of unknown e-mail addresses or unknown mail addresses; one person stopped participating in the panel after we made the random selection. Table 1 in the next subsection presents the distribution of individuals in the four conditions.

Out of the 745 respondents who were sent the opt-out consent text, only 38 respondents indicated they objected to linking (5.1%). Most sent a very short reply by e-mail; some asked for additional information and objected later. One respondent explicitly indicated that he has no problems with linking (which in fact was not asked for). The opt-out numbers per condition are presented in Table 1. The short e-mail version resulted in the highest opt-out rate (by quite a large margin), and the extended text resulted in lower opt-out rates in the e-mail condition but not in the letter condition.

Table 2 presents the results of a probit analysis in which the decision to opt out is explained by the two factors and the interaction between these factors. The results show significant main effects, as well as a significant interaction effect. The probability of

*Table 1. Number and percentage of respondents who opted out across the four conditions*

| Mode of communication | Length of consent text | | Total |
|---|---|---|---|
| | Short text | Extended text | |
| E-mail | 18 out of 187 (9.6%) | 6 out of 186 (3.2%) | 24 out of 373 (6.4%) |
| Letter | 5 out of 190 (2.6%) | 9 out of 182 (4.9%) | 14 out of 372 (3.8%) |
| Total | 23 out of 377 (6.1%) | 15 out of 368 (4.1%) | **38 out of 745 (5.1%)** |

opting out is lower in the letter condition than in the e-mail condition. The same holds for the extended text, which resulted in lower opt-out rates than the shorter text. The significant interaction effect implies that the effect of an extended text is different for a letter than for an e-mail. A test for a difference in effect on the probability of opting out between 'short text by e-mail' and 'extended text by letter' does not indicate a significant difference. The same holds for 'short text by letter' versus 'extended text by e-mail'. The combination of a short message and e-mail seems to be worst (see Table 1), but in general the results indicate that the way in which respondents were informed did not have much effect on opting out, and overall relatively few did so.

Note that for the experiment we randomly selected 500 households, and we sent the opt-out consent message to all panel members in these households. Conversations among household members might have induced a cluster effect, and the decision to opt out (or not) is not really an individual decision. The 38 respondents who objected were from 25 households, ten of them being a single-person household. There was one household with four (participating) members who all opted out, ten households with two (participating) members who all opted out, and four multi-person households with only one member opting out. If we base the probit analysis on household-level data (with a household being classified as opt out if any member of the household objected) the signs of the probit coefficients are the same as those presented in Table 2. However, due to the lower number of observations only the effect of mode is still significant.

### *4.3. Follow-Up Survey*

In the discussion of the results of the follow-up survey, we refer to the experimental group as Group A and to the control group as Group B. In February and March 2011 a few panel

*Table 2. Probit estimates in a model with opt out as dependent variable and factors mode of communication (Mode = 1 for letter, 0 for e-mail) and length of consent text (Length = 1 for extended, 0 for short) as independent variables, including interaction effect*

| Variable | Estimated coefficient | Z | P > \|z\| |
|---|---|---|---|
| Constant term | −1.30 | −10.3 | 0.000 |
| Mode (1 = letter) | −0.635 | −2.78 | 0.005 |
| Length (1 = extended) | −0.545 | −2.49 | 0.013 |
| Mode * Length | 0.833 | 2.52 | 0.012 |

members from both groups A and B indicated they wanted to stop participating in the panel. There was no significant difference between the number of respondents that decided to stop from Group A (7) and Group B (6), indicating that there was no short-term effect on panel attrition because of the opt-out statement. The panel management system excluded some others (seven for Group A and twelve from Group B) before fielding the questionnaire for other reasons (e.g., because the respondent indicated they would be unavailable for a longer period due to holiday, illness, etc.). This yielded a final selection of 1,489 panel members who were asked to complete the follow-up survey (Group A: 731; Group B: 758).

The overall response rate to the follow-up survey was 73.2% (1,090 out of 1,489). The response rate was almost the same for both groups: 73.5% for Group A (537 out of 731) and 73.0% for Group B (553 out of 758). One respondent in Group A and one respondent in Group B had incomplete data.

#### 4.3.1. Knowledge Questions

We first discuss the results of Part II (the knowledge questions), which was only presented to Group A. Almost 70% of the respondents said they had read the e-mail (or letter). There was no significant difference between the group who received the e-mail and the group who received the letter. That is, whether the opt-out statement was (said to be) read did not depend on mode.

We presented seven statements on the process of linking to the respondents. Table 3 summarizes the responses.

For four of the seven statements a majority of the respondents gave the correct answer. However, for none of the statements was the percentage of respondents giving the correct

*Table 3. Summary of the responses given to the seven statements concerning the record linkage. The correct answer for each statement is displayed in brackets*

| When linking your responses to our questionnaires to information that Statistics Netherlands has available about you . . . | Answer (in %) | | |
|---|---|---|---|
| | Correct | Incorrect | Don't know |
| a) your name, gender, and date of birth will be sent to Statistics Netherlands. [TRUE] | 34.7 | 44.2 | 21.1 |
| b) researchers (from outside Statistics Netherlands) get access to your name, gender, and date of birth. [FALSE] | 68.3 | 11.4 | 20.3 |
| c) your name, gender, and date of birth will be saved with the linked data. [FALSE] | 34.7 | 39.2 | 26.1 |
| d) for each project the linked data will always stay at Statistics Netherlands, and will be destroyed after completion of the specific project. [TRUE] | 39.9 | 16.4 | 43.7 |
| e) results of the study can be traced to you as an individual. [FALSE] | 65.5 | 9.7 | 24.8 |
| f) every researcher can consult the linked data via the Internet. [FALSE] | 66.8 | 7.5 | 25.7 |
| g) the Dutch Data Protection Authority supervises the linking and analyses of the data. [TRUE] | 65.5 | 5.6 | 28.9 |

answer very high. A majority seemed to understand that when linking survey data to administrative data researchers from outside Statistics Netherlands would not gain access to name, gender, and date of birth (68.3%), that results of the study cannot be traced to a specific individual (65.5%), that researchers cannot consult the linked data via the Internet (66.8%), and that the Dutch Data Protection Authority supervises the linking and analysis of the data (65.5%). However, only a minority knew that when linking information is sent to Statistics Netherlands, these variables are not saved with the linked data (34.7%), and that for each project the linked data will always stay at Statistics Netherlands, and will be destroyed after completion of the specific project (39.9%). We examined whether presenting the consent text in the survey only to the respondents who indicated they did not read the e-mail (or letter) affected the answer distributions. We did not find statistically significant associations, that is, the answer distributions did not differ between those who indicated they had read the letter or e-mail (and were not presented with the text again) and those who were presented with the text.

No respondent had all answers incorrect, and 3.5% of the respondents scored the maximum value (that is, all answers correct). When comparing the group of respondents who opted out with those who did not opt out, we found a striking result. Those who objected (opted out) had a significantly lower total score value (1.06 versus 2.51; $p < 0.001$). That is, they seem to have significantly *less* knowledge about the aspects of linking survey records to administrative data. Considering the responses to individual items for the group who opted out, the results not only show that the average frequency of correct answers to the individual items is much lower for the opt-out group (30.3% versus 55.2% for those who did not opt out), but also the average frequency of 'don't know' answers is much higher for the opt-out group (54.6% versus 25.4% for those who did not opt out).

When comparing the total score value of the panel members who received the extended text with those who received the short text, we found a significant difference. Panel members who received the extended text have a better knowledge about the linking process than those who received the short text (score values of 2.95 versus 1.88 respectively; $p < 0.001$). This seems to be a reassuring result: the more details one provides, the better respondents seem to understand what they are consenting to.

As mentioned in Subsection 3.4, we also calculated an alternative to the total score value: the measure of perceived risk of disclosure. We subtracted one point for a 'yes' answer to the statements d and g (see Table 3); for all the other statements a 'yes' answer increased the measure by one point.

Of the 536 respondents, 39 respondents (7.3%) scored $-7$, or the lowest possible perceived risk; the highest value for perceived risk of disclosure was 6 (for only one respondent). There is (again) an obvious and significant difference between the group of respondents who opted out and the group who did not. The group of respondents who opted out have an average value of perceived risk of $-0.94$, while this average for the group of respondents who did not opt out is $-2.72$ ($p < 0.001$). Those who objected to record linkage thus have a significantly higher perceived risk of disclosure.

Comparing the measure for perceived risk for those who received an extended text with those who received a short text, the perceived risk is significantly lower for the extended text ($-3.20$ versus $-2.02$; $p < 0.001$). So providing more detail is not only associated

with higher levels of knowledge, it is also associated with lower perceived risk of disclosure, regardless of the truth.

#### 4.3.2. General Attitudes

The relative frequency distributions of the answers to the questions in Part I of the follow-up survey for both groups A and B are shown in Table 4. A chi-square test shows no significant difference between the answers given by Group A and B to the first question about personal privacy ($p = 0.44$). That is, there is no evidence that presenting the opt-out consent statement makes respondents more concerned about personal privacy in general. Table 4 also shows that the same holds true when asking questions about violations of privacy rights by banks, credit card companies, tax authorities, research on public opinion (either by government or in general), and computers (storing large amounts of information). The consent experiment did not appear to result in any changes in attitudes towards these issues. The experimental and control group also did not differ in concerns about continued confidentiality of information possessed by private and public organizations. We only found that the experimental group felt more strongly than the control group that different government agencies can get information about the respondent

*Table 4. Relative frequency distributions (in %) of the answers to questions on concerns for both the experimental (A) and control (B) group*

| Questions on concerns | Experimental group (A) | Control group (B) |
|---|---|---|
| *Overall, how concerned are you about your (personal) privacy?* | | |
| not at all concerned | 13.2 | 14.5 |
| not very concerned | 45.6 | 49.2 |
| a bit concerned | 37.2 | 32.7 |
| very concerned | 3.9 | 3.6 |
| *Please indicate whether you feel that your privacy is violated by the following entities (yes/no):* | | |
| [frequencies of 'yes' answers are displayed] | | |
| – banks and credit card companies, when they inquire after your financial situation | 46.9 | 46.3 |
| – the government, when you fill out your tax forms | 14.5 | 14.1 |
| – the government, when they conduct research projects among the population | 24.8 | 25.3 |
| – computers, which store a lot of information about you | 80.4 | 80.1 |
| – persons that ask questions as part of public opinion surveys | 39.5 | 41.8 |
| *Do you think that government entities can gather information about you if they try (yes/no)?* | 95.9 | 90.2 |
| [frequency of 'yes' answers is displayed] | | |
| *All sorts of private and public organizations possess personal information about us. How concerned are you that this information will remain confidential?* | | |
| not at all concerned | 6.7 | 6.0 |
| not very concerned | 30.0 | 31.3 |
| a bit concerned | 55.5 | 52.4 |
| very concerned | 7.8 | 10.3 |

if they try to (95.9% versus 90.2%; $p < 0.001$). Overall the conclusion is that *concerns* about privacy hardly change after being exposed to the consent statement.

In terms of *trust* that information gathered about the respondent is treated confidentially, we found a few significant effects. Table 5 shows the frequency distributions of the answers to questions on trust for both the experimental and control group. Trust in research agencies that investigate public opinion and government agencies such as Statistics Netherlands to keep the information they collect from the respondent confidential turned out to be significantly different between those receiving the opt-out statement (Group A) and those not (Group B) ($p = 0.046$ and $p = 0.001$, respectively). Those who were exposed to the consent statement have *higher* levels of trust in these organizations.

## 5. Implementation and Longitudinal Effects

In September 2011 we informed all LISS panel members – except those we selected for the experiment – about the record linkage, and gave them the opportunity to opt out. Based on the results of the experiment and follow-up survey, we decided to send an extended text by e-mail to the balance of the LISS panel. The objective was not only to minimize opt-out rates but also to remove unfounded fears and to ensure that respondents understand what they are consenting to. The short letter had slightly lower opt-out rates than the extended e-mail in the experiment, but this difference was insignificant and the extended text increased knowledge and decreased perceived risk of disclosure, based on the results of the follow-up survey.

In total we sent 6,055 panel members the e-mail message; 551 (9.1%) opted out. This rate is significantly higher than that for the experiment (5.1% overall or 3.2% for the extended text by e-mail). A possible explanation is the fact that at the time the e-mail message was sent out, several cases of fraud related to data integrity were published in the

*Table 5. Relative frequency distributions (in %) of the answers to questions on trust for both the experimental (A) and control (B) group*

| Questions on trust | Experimental group (A) | Control group (B) |
|---|---|---|
| *To what extent do you trust that information gathered about you is treated confidentially by:* | | |
| – research agencies that investigate public opinion | | |
| not at all | 5.6 | 9.6 |
| not very | 27.4 | 29.7 |
| a bit | 50.7 | 46.6 |
| a lot | 16.2 | 14.1 |
| – market research agencies | | |
| not at all | 14.6 | 19.6 |
| not very | 36.9 | 32.2 |
| a bit | 39.6 | 39.1 |
| a lot | 9.0 | 9.1 |
| – government entities such as Statistics Netherlands | | |
| not at all | 3.7 | 4.9 |
| not very | 13.8 | 17.8 |
| a bit | 45.9 | 51.6 |
| a lot | 36.6 | 25.7 |

Netherlands and discussed extensively in the Dutch media. This may have triggered respondents to be more cautious in consenting to data linkage.

There is also the question of what the effects are of asking for consent in the long(er) run. Sending out a consent statement might affect the loyalty of panel members towards the fieldwork organization, and – if concern or distrust is stimulated by such a consent question – response rates may drop and panel attrition may increase. To examine the effects of sending out the opt-out consent statement, we compared the experimental group with all the other panel members who did not receive the consent statement (until September 2011). In the period February 2011–September 2011 the attrition rate on the individual level was equal to 8.3% for the panel members from the experimental group. For the remaining panel members the attrition rate in the same period was equal to 7.5%. The difference is not significant ($p = 0.43$).

In terms of response to the monthly questionnaires, we considered the following indicator. For the six-month period March–September 2011 we divided the number of months in which the panel member completed at least one of the questionnaires for which he was selected by the total number of months in which the panel member was selected for at least one questionnaire. This yields an indicator of participation. From the experimental group 92.8% were selected for questionnaires in all six months; for the non-experimental group this percentage was equal to 91.7%. The average percentage of participation is 72.2% for the panel members from the experimental group, compared to 70.4% for the panel members who were not included in the experimental group. Once again, the difference is not significant ($p = 0.22$).

The fact that the consent request could increase anxiety or distrust may be stronger for the group of panel members who opt out of linkage. To test for this we examined the group of respondents who did not participate in the experiment, but who were sent the opt-out statement in September 2011. We compared attrition and participation rates of the group of panel members who opted out with the group who did not opt out. We took data from the period September 2011–August 2012 (twelve months).

The attrition rate (on the individual level) was equal to 12.0% for the panel members who opted out (in response to the consent statement sent in September 2011), compared to 14.7% for those who did not opt out. This difference is not significant ($p = 0.06$). With respect to the participation rate (as defined above), among those who were active in the panel in the six months prior to September 2011, the opt-out respondents had an average percentage of participation of 84.7% in the twelve-month period after exposure to the statement, which is significantly higher than the 79.7% for those who did not opt out ($p < 0.001$). In other words, those who opted out in response to the presentation of the statement in September 2011 did *not* attrite at a higher rate and had *higher* participation rates in the following twelve months than those who did not opt out.

In summary, we found no evidence that sending an opt-out consent statement leads to a higher attrition or lower participation rates in future surveys in the panel.

## 6. Discussion

Results from our study show that sending a short e-mail message about record linkage between survey and administrative data yields the highest opt-out rate. A short e-mail

message may be considered as an attempt of the survey organization to get things done in a quick and easy way. It also provides the respondents with an easy way to opt out – simply by replying to the e-mail. A short letter or an extended e-mail yields lower opt-out rates. A letter appears to be more legitimate to respondents, and an extended text indicates the request is taken seriously by the survey organization.

The cognitive interviews as well as results from the experiment and follow-up survey indicate that respondents find it difficult to understand what linkage is all about. Whether we choose opt-in or opt-out consent statements, respondents seem to have little idea what is happening with the survey data they provide. The central question is whether it is possible at all to explain this to respondents and, if so, how best to do so. While researchers may be confident they can explain exactly what is happening in the case of record linkage, respondents in general population surveys may have little interest in these explanations. Results from the follow-up study, however, do indicate that the more details one provides the better respondents seem to understand what they are consenting to. These results run counter to the findings of Singer et al. (1992), which were based on a small convenience sample of university students and focused on survey participation. Singer (1978) found no effect of a short versus long statement on participation in an interviewer-administered survey. More research is needed on this topic. Our results also show that providing respondents with opt-out consent does not appear to increase concerns over privacy or trust in the survey organization.

While the proportion of LISS panel members who opted out of record linkage is relatively small (9.1%) compared to what might have occurred if explicit opt-in consent was required, it may still bias the results of projects using linked data. The rich amount of data collected earlier in the LISS panel offers the unique opportunity to get a clear picture of the respondents who opted out of record linkage, and the possible effects of their exclusion from key analyses. The LISS panel provides a further opportunity to explore the long-term effects of exposure to opt-out consent on later participation and attrition.

In summary, using an opt-out consent process in the LISS panel appears to have maximized the value of the linked data for researchers without apparent effects on panel members in terms of increased concerns over privacy or increased rates of attrition.

## Appendix

### Short and Extended Text Versions Used in the Experiment

### SHORT TEXT VERSION (translated from Dutch)

Dear < XXX > ,

As member of the LISS panel, you are helping scientists gain valuable information. As you know, your privacy is guaranteed in all our research projects. Your answers to the questionnaires are only used for scientific research. We strictly comply with the Personal Data Protection Act, and never provide any information to other organizations like the Tax Administration or the UWV.

Some information is difficult or even impossible to acquire through your answers to our questions, for example because it would make the questionnaires extremely long or

complicated. Fortunately, Statistics Netherlands (SN) has additional information about you available; for example about your General Old Age and pension benefits, or data about your health and well-being.

For that reason we will soon start working with SN to combine information they have about you with your answers to our questionnaires. This will be done using highly protected computer systems. The results are always anonymous and cannot be traced to you in any way.

If you do not want us to combine your answers with data about you at SN, then please contact the CentERdata helpdesk on 0800–023 14 15 (free) or via lisspanel@uvt.nl.

With kind regards,

**EXTENDED TEXT VERSION (translated from Dutch)**

Dear < XXX > ,

As member of the LISS panel, you are helping scientists gain valuable information. As you know, your privacy is guaranteed in all our research projects. Your answers to the questionnaires are only used for scientific research. We strictly comply with the Personal Data Protection Act, and never provide any information to other organizations like the Tax Administration or the UWV.

Some information is difficult or even impossible to acquire through your answers to our questions, for example because it would make the questionnaires extremely long or complicated. Fortunately, Statistics Netherlands (SN) has additional information about you available; for example about your General Old Age and pension benefits, or data about your health and well-being.

*How Will We Use the Additional Information?*

Researchers using our data never have access to your name or address details. Our panel management does of course know which panel member number belongs to what person. SN also knows to whom the information that they have belongs. By comparing name, sex and date of birth, it is possible to combine the data of CentERdata and SN.

The information exchange with SN will be done using highly protected computer systems. After the data have been combined, your name, sex and date of birth will be removed from the database. Researchers that study the combined data will therefore never see your name, sex or date of birth. The results are strictly anonymous and cannot be traced to you in any way.

*Who Will Have Access to the Combined Data?*

Researchers can submit a request to SN to use the combined data for scientific research. If the request is accepted, the researcher is required to sign a contract with SN. The data will never leave SN. The researcher can access and use the SN data by means of a fingerprint reader and a secure connection. Research results will first be checked by SN before the researcher is permitted to publish them.

*What Will CentERdata Do?*

We will soon start working with SN to combine information they have about you with your answers to our questionnaires. If you do not want us to combine your answers with data about you at SN, then please contact the CentERdata helpdesk on 0800 – 023 14 15 (free) or via lisspanel@uvt.nl.

With kind regards,

## 7. References

Anderman, C., A. Cheadle, S. Curry, P. Diehr, L. Shultz, and E. Wagner. 1995. "Selection Bias Related to Parental Consent in School-Based Survey Research." *Evaluation Review* 19: 663–674. DOI: http://dx.doi.org/10.1177/0193841X9501900604.

Bates, N.A. 2005. "Development and Testing of Informed Consent Questions to Link Survey Data with Administrative Records." In Proceedings of the American Statistical Association, May 12–15, 2005. 3786–3793, Miami Beach, FL. Available at: http://www.amstat.org/committees/ethics/linksdir/Jsm2005Bates.pdf (accessed May 21, 2014).

Calderwood, L. and C. Lessof. 2009. "Enhancing Longitudinal Surveys by Linking to Administrative Data." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 55–72. Chichester: John Wiley & Sons.

Couper, M.P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75: 889–908. DOI: http://dx.doi.org/10.1093/poq/nfr046.

Dahlhamer, J.M. and C.S. Cox. 2007. "Respondent Consent to Link Survey Data with Administrative Records: Results from a Split-Ballot Field Test with the 2007 National Health Interview Survey." In Proceedings of the Federal Committee on Statistical Methodology Research Conference, November 5–7, 2007, Arlington, VA. Available at: http://www.fcsm.gov/07papers/Dahlhamer.IV-B.pdf (accessed May 21, 2014).

Eaton, D.K., R. Lowry, N.D. Brener, J.A. Grunbaum, and L. Kann. 2004. "Passive Versus Active Parental Permission in School-Based Survey Research." *Evaluation Review* 28: 564–577. DOI: http://dx.doi.org/10.1177/0193841X04265651.

Ellikson, P.L. and J.A. Hawes. 1989. "Active vs. Passive Methods for Obtaining Parental Consent." *Evaluation Review* 13: 45–55. DOI: http://dx.doi.org/10.1177/0193841X8901300104.

Fendrich, M. and T.P. Johnson. 2001. "Examining Prevalence Differences in Three National Surveys of Youth: Impact of Consent Procedures, Mode, and Editing Rules." *Journal of Drug Issues* 31: 615–642.

Gray, M. 2010. "A Review of Data Linkage Procedures at NatCen", Working Paper. London: National Centre for Social Research. Available at: http://www.natcen.ac.uk/media/205504/data-linkage-review.pdf (accessed May 21, 2014).

Hunt, K.J., N. Schlomo, and J. Addington-Hall. 2013. "Participant Recruitment in Sensitive Surveys: A Comparative Trial of 'Opt In' Versus 'Opt Out' Approaches." *BMC Medical Research Methodology* 13, DOI: http://dx.doi.org/10.1186/1471-2288-13-3.

Junghans, C., G. Feder, H. Hemingway, A. Timmis, and M. Jones. 2005. "Recruiting Patients to Medical Research: Double Blind Randomised Trial of 'Opt-in' versus

'Opt-out' Strategies." *British Medical Journal* 331: 940. DOI: http://dx.doi.org/10.1136/bmj.38583.625613.AE.

Pascale, J. 2011. "Requesting Consent to Link Survey Data to Administrative Records: Results from a Split-Ballot Experiment in the Survey of Health Insurance and Program Participation (SHIPP)." *Study Series in Survey Methodology* #2011-03, (U.S. Census Bureau. Washington DC). Available at: http://www.census.gov/srd/papers/pdf/ssm2011-03.pdf (accessed April 2013).

Pascale, J. and T.S. Mayer. 2004. "Exploring Confidentiality Issues Related to Dependent Interviewing: Preliminary Findings." *Journal of Official Statistics* 20: 357–377.

Sakshaug, J., M.P. Couper, M.B. Ofstedal, and D. Weir. 2012. "Linking Survey and Administrative Records: Mechanisms of Consent." *Sociological Methods and Research* 41: 535–569. DOI: http://dx.doi.org/10.1177/0049124112460381.

Sakshaug, J. and F. Kreuter. 2012. "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data." *Survey Research Methods* 6: 113–122.

Sala, E., J. Burton, and G. Knies. 2012. "Correlates of Obtaining Informed Consent to Data Linkage: Respondents, Interview and Interviewer Characteristics." *Sociological Methods and Research* 41: 414–439. DOI: http://dx.doi.org/10.1177/0049124112457330.

Sala, E., G. Knies, and J. Burton. 2013. "Propensity to Consent to Data Linkage: Experimental Evidence from the Innovation Panel on the Role of Three Survey Design Features." *Understanding Society Working Paper Series* 2013-05. Colchester, England: University of Essex, Institute for Social and Economic Research. Available at: http://www.iser.essex.ac.uk/publications/working-papers/understanding-society/2013-05 (accessed June 29, 2014).

Scherpenzeel, A.C. and M. Das. 2011. "'True' Longitudinal and Probability-Based Internet Panels: Evidence from the Netherlands." In *Social and Behavioral Research and the Internet*, edited by M. Das, P. Ester, and L. Kaczmirek, 77–104. New York: Taylor and Francis.

Singer, E. 1978. "Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys." *American Sociological Review* 43: 144–162.

Singer, E. 2003. "Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits." *Journal of Official Statistics* 19: 273–285.

Singer, E. and M.P. Couper. 2011. "Ethical Considerations in Internet Surveys." In *Social and Behavioral Research and the Internet*, edited by M. Das, P. Ester, and L. Kaczmirek, 133–162. New York: Taylor and Francis.

Singer, E., H.-J. Hippler, and N. Schwarz. 1992. "Confidentiality Assurances in Surveys: Reassurance or Threat?" *International Journal of Public Opinion Research* 4: 256–268. DOI: http://dx.doi.org/10.1093/ijpor/4.3.256.

Vellinga, A., M. Cormican, B. Hanahoe, K. Bennett, and A. Murphy. 2011. "Opt-Out as an Acceptable Method of Obtaining Consent in Medical Research: A Short Report." *BMC Medical Research Methodology* 11: 40. DOI: http://dx.doi.org/10.1186/1471-2288-11-40.

Received April 2013
Revised April 2014
Accepted April 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 499–520, http://dx.doi.org/10.2478/JOS-2014-0031

# Predictions vs. Preliminary Sample Estimates: The Case of Eurozone Quarterly GDP

*Enrico D'Elia*[1]

Economic agents are aware of incurring a loss in basing their decisions on their own extrapolations instead of on sound statistical data, but this loss may be smaller than the one related to waiting for the dissemination of the final data. Broad guidelines on deciding when statistical offices should release preliminary and final estimates of the key statistics may come from comparing the loss attached to users' predictions with the loss associated to possible preliminary estimates from incomplete samples. Furthermore, the cost of delaying decisions may support the dissemination of very early estimates of economic indicators, even if their accuracy is not fully satisfactory from a strict statistical viewpoint. Analysing the vintages of releases of quarterly Euro area GDP supports the view that even very inefficient predictions may beat some official preliminary releases of GDP, suggesting that the current calendar of data dissemination requires some adjustment. In particular, actual "flash" estimates could be anticipated, while some later intermediate releases are likely less informative for the users.

*Key words:* Accuracy; data dissemination; Eurozone GDP; forecast; preliminary estimates; timeliness.

## 1. Introduction

The trade-off between accuracy and timeliness of statistical data is a key issue for statistical offices. It has been analysed mainly with reference to estimates of GDP and other "Principal European Economic Indicators" identified by the Economic and Financial Committee of the European Commission, aimed at detecting the turning points of the business cycle earlier. An international conference organised by the UNSTATS (2009) discussed the same topic in depth and the OECD analysed the quality of statistical information within the "Short-Term Economic Statistics Timeliness Framework".

Notably, Altavilla and Ciccarelli (2007) and the European Central Bank (2009) pointed out that the flash estimates of European GDP do not differ significantly from the official first releases published later, so that early estimates are probably more helpful for decision makers than the corresponding final releases. Economic agents also form their informed predictions on the relevant variables while waiting for official data releases. The main aim of this article is to show how the "competition" between the accuracy of users' judgements and the accuracy of early official estimates may provide some guidelines for improving the

[1]Italian Ministry of Economy and Finance, via XX Settember, 97, Rome 00187 and ISTAT, via Cesare Balbo, 16, Rome 00186, Italy. Email: delia.enrico@gmail.com

**Acknowledgments:** The views expressed in this article are those of the author and do not necessarily reflect views at the Italian Ministry of Economy and Finance and ISTAT. The author gratefully acknowledges the valuable suggestions and criticisms made by the referees of this journal, Alberto Zuliani and the participants of a series of seminars. Of course, errors and omissions are the responsibility of the author.

data dissemination policy of statistical offices, particularly for the quarterly estimates of GDP in the Eurozone.

Early estimates of economic indicators are welcomed by decision makers who are not in the position of waiting for the dissemination of the final results of the pertinent statistical surveys before choosing between alternative strategies. In particular, the timing is important in most decision processes concerning investment, consumption, and price setting. Thus, users of statistical data often have to resort to model-based predictions on the final outcome of some statistical surveys on past and current facts, often referred to as "nowcasts", to be distinguished from genuine forecasts about the future. In other words, predictions and preliminary results from surveys can be regarded by decision makers as imperfect substitutes. This fact doubtless offers a novel viewpoint on the trade-off between timeliness and accuracy of statistics, providing some suggestions about the strategy for disseminating statistical data. Particularly, the implicit competition between nowcasts and early estimates should be taken into account together with the usual assessments on production costs, technical capability, transparency, credibility, and legal obligations of statistical offices.

Data users are perfectly aware that the final results of statistical surveys are more accurate than forecasts, nowcasts and early estimates. In principle, the profit expected from decisions based on very precise final statistical data is higher than profit deriving from choices founded on predictions and first releases of pertinent statistics. Nevertheless, waiting for the final results of statistical surveys before deciding is costly as well, since profitable actions are postponed and economic resources are left unused, resulting in further costs. In addition, users know that both the accuracy of their predictions and of preliminary estimates usually improve over time, at least under "normal" conditions when no major shocks hit the economy or the data collection process. Indeed, at the beginning of data collection, say at time $t$, users' predictions are expectedly superior to any pure sample estimate, since the former embody public and private information, while the variance of pure survey estimates based on very few observations is virtually infinite and is subject to small sample bias unless the statistical offices adopt an explicit Bayesian approach and reliable priors, which is infrequent in official statistics.

For their part, statistical offices acknowledge that earlier estimates meet the needs of most users, and are generally technically capable of producing excellent nowcasts, also by exploiting experts' judgements and confidential sources of information. In principle, the statistical offices would be able to release a mass of preliminary data as well, even though they are aware that it could be very costly. Nevertheless, official statisticians recognise that data revised too frequently and too much would confuse users and possibly damage institution credibility. In addition, publishing provisional data, possibly not included in the release calendars agreed to at international level, would raise uncertainty and search costs for users and introduces unduly informational asymmetries in international statistics which can ultimately impair users. Thus it is hoped that the current data release calendar finds the middle ground among many different requirements and constraints, and the specific viewpoint presented here should be correctly considered only as an additional one.

Let us assume that the accuracy of statistical estimates improves as data collection proceeds over time, achieving on average the accuracy of users' forecasts only at time

$t + h_0$, while information available to the users does not improve significantly. It follows that typical users would not exploit and appreciate figures possibly released before $t + h_0$, because these figures are considered less accurate than their own nowcasts. The threshold $h_0$ depends crucially on subjective users' conditions, primarily on their technical capability and knowledge.

Many users may wait intentionally for "official" data as long as preliminary estimates are expected to improve very fast, while the loss of making decisions based on inaccurate forecasts could be large. Moreover, users' extrapolations hardly beat preliminary survey results when a major shock hits the economy, inevitably making model-based predictions less accurate. Nevertheless, the threshold $h_0$ is hardly null, and may be quite large if the accuracy of early estimates does not sufficiently increase over time or occasionally decreases.

Eurozone GDP estimates, analysed in the next sections, derive from a complex procedure that exploits both pure sample information and model-based estimators. As a consequence, comparing official GDP preliminary estimates and users' nowcasts should provide strong evidence in favour of the dominance of official preliminary estimates, supporting the current dissemination policy of Eurostat, since the efficiency of the data elaboration process most likely reduces $h_0$ significantly. Nevertheless, the empirical evidence presented in Section 4 seems to show that even very inefficient predictions may do better than some preliminary estimates of GDP, suggesting that there is scope for improving the calendar of data releases even if representative data users are not very sophisticated. However, this result may be influenced by the particular period of time analysed (2002–2012) and by the small sample of fully comparable data available. Of course, a more comprehensive analysis of costs and benefits of changing the present calendar is also needed. In addition, conclusions depend crucially on the assumed ability of the representative users to form good forecasts and to exploit available information.

The next section exploits some properties of preliminary estimates from incomplete samples to derive an ideal calendar for disseminating preliminary estimates exactly when their accuracy beats the errors size of model-based predictions. The main conclusions are derived under the ideal conditions that no large shock perturbs the economy and that the accuracy of official estimates improves over time. The consequences of departing from this simplified framework are discussed briefly as well. The third section introduces the cost of delaying decisions while waiting for better official estimates. This issue, if taken into consideration, should encourage statistical offices to anticipate the release of data, but also clarifies that the dissemination calendar should adapt to the characteristics of some "representative" user of statistical data, endowed with a given capability and needs. Thus it is crucial to acknowledge that statistical offices must serve different users, including legislators and governmental agencies. The fourth section analyses the different vintages of quarterly GDP estimates in the Eurozone, regularly released by Eurostat, and recommends some adjustments to the current dissemination policy, even under the simplified hypothesis that users form very naïve predictions based on GDP and do not incur costs for delaying their decisions. In particular, the suitability of the three major data releases currently available (respectively 45, 65 and 100 days after the end of the reference quarter) is discussed. Some concluding remarks close the article.

## 2. The Accuracy of Preliminary Sample Estimates and Forecasts

Let $x_{i,t}$ measure a quantitative characteristic of the $i$-th individual at the time $t$, whose unconditional mean is $m_t$. It is assumed that the "representative" economic agent has to base decisions on $m_t$ by using only the incomplete information set $\Omega_{t+h}$ available at time $t+h$. Typically $\Omega_{t+h}$ includes the past releases of the time series of $m_t$ and other aggregate economic indicators related to $m_t$; private information generally unavailable to the statistical offices; "soft" statistics, also produced by private agencies; judgements of experts. Nevertheless, $\Omega_{t+h}$ excludes the observations on $x_{i,t}$ collected and processed by the statistical office after $t$.

Thus, at least two provisional estimates of $m_t$ are ideally available at the time $t+h$:

(a) $f_{t+h} = \mathrm{E}(m_t|\Omega_{t+h})$ the subjective prediction produced by exploiting the information set $\Omega_{t+h}$;
(b) $s_{t+h}$ the preliminary estimate based on the first $M_{t+h}$ observations collected at time $t+h$ by the statistical office.

Within this simplified framework, the representative user has the advantage of exploiting prior beliefs and private information, but has no access to individual records collected by the statistical office. The latter is allowed to use sample observations, but no other potentially useful pieces of information on $m_t$. In principle, statistical offices could develop mixed estimates within an explicit Bayesian framework, also taking into account experts' judgements and other relevant nonsample information. Although the Bayesian approach has many theoretical advantages, it is seldom used to improve sample estimates directly. Statistical offices tend to avoid estimation procedures that risk appearing too subjective, in view of defending and strengthening their neutrality and independence, in compliance with the first principle of the European Statistics Code of Practice (2011). Although Little (2012) points out the possible advantage of adopting an explicit Bayesian approach in official statistics and discusses an application to the US Census data, Bayesian methods are applied in official statistics mainly to treat nonresponses (see Graham et al. 2009), to reduce the disclosure risk in the dissemination of individual data (see Little et al. 2004), to match the units of different surveys statistically (see D'Orazio et al. 2006), but not to improve preliminary estimates directly.

The time series $\{x_{i,t}\}$ can be decomposed as follows

$$x_{i,t} = f_{t+h} + v_{t+h} + e_{i,t} \tag{1}$$

where $v_{t+h} = m_t - f_{t+h}$ is an innovation process, with $\mathrm{E}(v_{t+h}|\Omega_{t+h}) = 0$ and $\mathrm{E}({v_{t+h}}^2|\Omega_{t+h}) = \phi_h^2$ not depending on $t$, even though the unconditional average of $v_{t+h}$, say $\mathrm{E}(\mathrm{v}_{t+h})$, is not necessarily null; $e_{i,t}$ is an idiosyncratic factor with $\mathrm{E}(e_{i,t}) = 0$ and $\mathrm{E}({e_{i,t}}^2) = \sigma^2$. Notably, the two assumptions on $e_{i,t}$ are quite standard, while the hypotheses on $v_{t+h}$ could be violated if some time-specific factor changes the predictability of the relevant events systematically. For instance, forecast accuracy of GDP likely changes at the turning points of the business cycle or when some structural change makes the economic activity more or less erratic. In the latter cases the time invariance of $\phi_h^2$ does not hold, while the variance of $e_{i,t}$ does not necessarily change.

Let individual observations be collected and processed by the statistical office randomly, regardless of whether they are gathered almost continuously over time or in

large batches, as commonly occurs. In this case the subscript $i$ in (1) may denote the collection order of data, without any loss of generality. Thus the preliminary pure sample estimate of $m_t$ at time $t + h$ is

$$S_{t+h} = \sum_{i=1}^{M_{t+h}} w_{i,t+h} x_{i,t}, \tag{2}$$

where the weights $w_{i,t+h}$ are such that $\sum_{i=1}^{M_{t+h}} w_{i,t+h} = 1$ for each $t + h$. Under the previous assumptions on $e_{i,t}$ in (1) and on the random collection of data, the average $\mathrm{E}(s_{t+h})$ evaluated over every possible sample of size $M_{t+h}$ equals $m_t$. Furthermore, if the individuals' deviations from the average are mutually independent, the usual assumption $\mathrm{E}(e_i e_j) = 0$ for $i \neq j$ applies, so that the standard deviation of $s_{t+h}$ is

$$\sigma_h = \frac{\sigma}{\sqrt{M_{t+h}}} \tag{3}$$

in the simplest case of equally weighted observations.

Within a Bayesian framework, the estimator $s_{t+h}$ and its variance should take into account the priors on $m_t$, so that $s_{t+h}$ would be a weighted average of the sample mean (2) and the mean of the assumed probability distribution of $m_t$. Moreover, if the data are drawn from a normal population and the prior distribution of $m_t$ is normal as well, the posterior variance of $s_{t+h}$ is

$$\sigma_h = \frac{\sigma}{\sqrt{M_0 + M_{t+h}}} \leq \frac{\sigma}{\sqrt{M_{t+h}}} \tag{4}$$

where $M_0 > 0$ measures the confidence on the prior, that is the ratio between the variance of $e_{i,t}$ and the variance of the probability distribution assumed for $m_t$. The same result holds for the Theil–Goldberger mixed least square estimator of $m_t$, regardless of the probability distribution of data and priors. The parameter $M_0$ in (4) can be interpreted as the size of the virtual sample from which the prior distribution of $m_t$ has been estimated.

According to (3), $\sigma_h$ is virtually infinite before the survey begins, since no observation has yet been collected and $M_{t+h}$ is null. (4) also implies that $\sigma_h$ peaks at its maximum when $M_{t+h}$ equals zero, and is almost certainly large, unless the confidence of the statistical office in its priors is implausibly strong. In any case, during the survey, $M_{t+h}$ is a nondecreasing function of $h$, for instance: $M_{t+h} = M(h)$ with $\frac{dM}{dh} \geq 0$, regardless of the reference period $t$. It follows from (3) and (4) that $\frac{d\sigma_h}{dh} \leq 0$ holds at least under "normal" conditions, in which data collected when $h$ takes some special values are not systematically biased and volatile. Note that this is not the case when most influential units are surveyed just at the beginning and the end of the data collection process, for instance because some units are able to provide the data only according to a special calendar (e.g.: just after the balance sheets or periodic reports have been published). In such unlucky cases, $\sigma_h$ may even increase with $h$ during some phases of the survey process. The case in which $\frac{d\sigma_h}{dh} \geq 0$ will be discussed only briefly, since it would be even more supportive of the advantage of nowcasts over official preliminary estimates.

The profit loss associated to the use of preliminary estimates, say $S(h)$, can be assumed to be a nondecreasing function of $\sigma_h$, say $S(h) = L(\sigma_h)$ with $\frac{dL}{d\sigma_h} \geq 0$ and $L(0) = 0$. The function $L(\sigma_h)$ depends largely on the subjective conditions of data users and on the

specific decision to be based on statistical data. In particular, the inaccuracy of a variable could be almost negligible in some cases, and potentially harmful in others. For instance, estimating the level and dynamics of GDP correctly is very important when deciding investment, but not export strategies. Nevertheless, the formal properties of $L(\sigma_h)$ utilised in the following sections are not influenced by such subjective factors.

Notably, $L(\sigma_h)$ is not necessarily a linear transformation of the standard deviation of errors $\sigma_h$, and in particular could be flat for a wide range of $\sigma_h$. It implies that $S(h)$ and $L(\sigma_h)$ are not necessarily quadratic functions of errors, as often assumed. The main limitation of the relationship $S(h) = L(\sigma_h)$ is that data users are assumed to be equally adverse to positive and negative estimation errors, in contrast to what Granger and Pesaran (2000) argue. However, if statistical data are used to design fiscal policies, the government is more likely to be worried about overestimating GDP growth, since less income entails larger budget imbalances, due to larger social expenditure and lower tax revenues. Also, most firms acting in a competitive market fear overestimating the potential market much more than underestimating it, since overestimation calls for unduly large investment and related financial costs. By contrast, in oligopolistic markets, plants could be oversized intentionally to prevent the entry of possible competitors, so that entrepreneurs would be more averse to underestimating market size. In general, the symmetric relationship $S(h) = L(\sigma_h)$ can be considered a feasible approximation of the true loss function of the representative agent only for a small size of errors.

The main advantage of relating $S(h)$ to $\sigma_h$ is that it makes it possible to compare users' predictions and preliminary sample estimates, disregarding the specific functional form of $L(\sigma_h)$, that is, the nature of decisions to be made by the representative user. As $\sigma_h$ is not a continuous function of $h$, $S(h)$ may also share this discontinuity. For instance, if data are collected in batches, $S(h)$ is very likely a piecewise continuous function, in all probability characterized by sudden drops after each batch of data has been processed, or when information on the most important units can be collected. In any case, $S(h)$ is suitable to be estimated empirically by statistical offices from the track of data collection, and can be approximated by users from the revisions of data, compared to some benchmark release, which can be considered the ultimate estimation, hopefully closest to the true value of the relevant variables. Furthermore, assuming that $S(h)$ is a nondecreasing function of $\sigma_h$ implies that $\frac{dS(h)}{dh}$ shares the sign of $\frac{d\sigma_h}{dh}$, apart from possible discontinuities. For instance, Table 1 and Figure 1 provide some empirical evidence on the negative relationship between $\sigma_h$ and the dissemination delay $h$ of the preliminary estimates of quarterly GDP in the Eurozone released by Eurostat, compared to the official estimate released 400 days after each reference quarter. It is worth noticing that in the case examined here the condition $\frac{d\sigma_h}{dh} \leq 0$ holds even before the latest economic crisis, when the industrial structure and the heterogeneity among firms' performances was completely different. In addition, the accuracy of preliminary estimates of GDP seems to improve at decreasing rates, as if the data elaboration process is much more efficient at the beginning of the statistical survey and each additional observation makes only a minor contribution to the accuracy of the sample estimates.

Since $\Omega_{t+h} \supseteq \Omega_{t+h-1}$ by definition, it follows that $\frac{d\phi_h}{dh} \leq 0$, at least on average and in "normal" times, namely when news available at time $t+h$ prevails on "noise", as questioned by Blanchard et al. (2009). The function $\phi_h$ can be also discontinuous, with sudden drops when some influential piece of information is usually available only when $h$

takes some special values. The assumption $\frac{d\phi_h}{dh} \leq 0$, apart from some possible discontinuity points, relies crucially on the fact that the representative user is able to keep, or hopefully to improve over time, its capacity to understand and exploit available information efficiently. Notably, full rationality of economic agents is not strictly required for $\frac{d\phi_h}{dh} \leq 0$. For instance, it is enough that they are "rationally inattentive" as argued by Sims (2003), that is, they intentionally disregard part of the available information because collecting and elaborating it exceeds the profit expected from further improving their decisions. In any case, we will see that the hypothesis $\frac{d\phi_h}{dh} \leq 0$, although very likely and desirable, is not strictly necessary in designing an ideal calendar for data release.

Like $\sigma_h$, $\phi_h$ can also be measured empirically, for instance from direct surveys on users' judgements, or assuming a reasonable mechanism for the formation of nowcasts, as done in Section 4. Given the relation between the expected profit loss and the accuracy of data used to make a decision, one can define $F(h) = L(\phi_h)$. Thus $F(h)$, similarly to $S(h)$, can be considered a nondecreasing transformation of errors' size at time $t + h$. This property allows us to compare $\sigma_h$ and $\phi_h$ instead of the subjective and unknown functions $F(h)$ and $S(h)$.

If $v_{t+h}$ and $e_{i,t}$ are not correlated, as assumed above in "normal" times, the decomposition (1) implies that

$$\mathrm{E}(\sigma_h^2) = \mathrm{E}\left[\frac{1}{M_{t+h}}\sum_{i=1}^{M_{t+h}}(x_{i,t} - f_{t+h})^2\right] - \phi_h^2 \tag{5}$$

where the E(.) operator applies to the time series of the relevant variables. The expression in square brackets in (5) is larger than $\mathrm{E}(\sigma_h^2)$, since only the arithmetic average $m_t$ minimizes the sum of squared discrepancies $(x_{i,t} - f_{t,h})$, thus $\phi_h^2$ can be seen as the difference between the estimated variance among observations around the forecast $f_{t+h}$ on one hand and the variance $\sigma_h^2$ around the true average $m_t$ on the other. Therefore $\phi_h^2$ is most likely small compared to $\sigma_h^2$, as long as $f_{t+h}$ is a reasonable forecast of $m_t$.

As noted above, rational agents are assumed to be able to make forecasts even before data collection has begun, when $\sigma_h^2$ is virtually infinite, so that $\phi_h^2 < \sigma_h^2$ for $h \leq 0$. As time goes on, predictions may improve, thanks to the availability of other relevant pieces of information, but probably at a slower pace compared to a survey. Otherwise, forecasts would do better than statistical surveys all the time and implausibly the latter would have only a little value for the representative user. Excluding the latter implausible case, the assumptions that $\phi_h^2 < \sigma_h^2$ for $h \leq 0$ and $\frac{d\sigma_h}{dh} \leq \frac{d\phi_h}{dh}$ (disregarding possible singular discontinuities) subsequently imply that $\phi_h^2 = \sigma_h^2$ at some point in time, say $t + h_0$. Noticeably, the condition $\frac{d\sigma_h}{dh} \leq \frac{d\phi_h}{dh}$ ideally does not require that both $\phi_h^2$ and $\sigma_h^2$ are nonincreasing functions of $h$. Furthermore $h_0$ could be very large, so that nowcasts could retain their advantage for a long while. What is really crucial is that the initial advantage of nowcasts (if any) tends to reduce as the dissemination delay increases.

If statistical offices actually release the sample estimates available before $t + h_0$, users are likely to continue basing their decisions on their own forecasts in order to minimise their expected profit loss, unless this new piece of information improves $\Omega_{t+h0}$ because users can incorporate even the very inaccurate preliminary data released before $t + h_0$ into their nowcasts. The expected loss associated to these "data-adjusted" nowcasts determines a downward shift of the $\phi_h$ function and a new intersection point between $\phi_h$ and $\sigma_h$, say at

$h = h_1$. The size of the shift can be ideally measured carrying out a survey among the users, or by assuming some nowcasting model, as in the next sections. If the downward shift of $\phi_h$ is substantial, economic agents would welcome even earlier provisional releases of data by the statistical office. In contrast, if intermediate data releases improve users' predictions only to a lesser extent, users will in the meantime continue to base their decision on their own past nowcasts even after the dissemination of official data. As a consequence, comparing the functions $\phi_h$ and $\sigma_h$ after each data release may provide operative guidelines for refining the dissemination plans of statistical offices. In particular, the first data release could be anticipated if it causes a large downward shift of $\phi_h$, even if the inaccuracy of sample estimates is large. On the other hand, intermediate preliminary estimates that do not improve $\phi_h$ sufficiently should be avoided, since they are very likely costly for the statistical offices and less appreciated by the users. It is worth noting that $\phi_h$ and $\sigma_h$ can be compared even if they present some discontinuities; thus the approach proposed here to design an ideal data release calendar seems quite general. For instance, Table 3 and Figure 2 show an example of the interplay between the publication of preliminary estimates and the elaboration of users' nowcasts based on simple univariate time series extrapolations of quarterly GDP in the Eurozone. In particular, the dashed line in Figure 2 represents the accuracy of nowcasts adjusted after each data release that improves almost every time new data are published.

Indeed, the downward shift of the function $\phi_h$ would be null if statistical offices provided the best nowcast by applying efficient model-based estimators to the collected data, so that users' forecasts could hardly be better than the preliminary data published at time $t + h$. The improvement in nowcasts can be seen as a special case of efficiently exploiting the data collected up to $t + h$ by integrating missing data in the full sample by means of a model-based estimator, as discussed in Särndal and Lundström (2005). In any case, users can only combine available forecasts, as suggested by Clemen (1989) and Yang and Zou (2004), while the statistical offices possibly may combine the same forecasts and the provisional results of their surveys.

Provided that the survey ends at $t+H$, the relative performance of the two estimators $s_{t+h}$ and $f_{t+h}$ depends on the time schedule of the survey, which determines the coverage ratio $\frac{M_{t+h}}{M_{t+H}}$ on one hand and the ratio $\gamma_h = \frac{\phi_h}{\sigma_H}$ of the mean square error of prediction to the variance of $m_t$ among observations at the end of the survey on the other. The ratio $\gamma_h$ ranges from 0 to infinity: In particular, $\gamma_h$ is null if the time series $m_t$ is purely deterministic and tends to infinity if individuals are identical. For instance, the changes of the average age of a stationary population can be virtually predicted without any error even though the age differs greatly among individuals. By contrast, the yield of a homogeneous set of equities can hardly be predicted, even if they have the same market price.

As assumed cautiously above, let the forecast accuracy improve over time less than $\frac{\sigma_h}{\sigma_H}$, namely less than $\sqrt{\frac{M_{t+H}}{M_{t+h}}}$ according to (3). Since rational agents prefer their own forecast to preliminary estimates of $m_t$ as long as $\sigma_h \geq \phi_h$, it follows that

$$\sqrt{\frac{M_{t+h}}{M_{t+H}}} \geq \frac{\sigma_H}{\phi_h}. \tag{6}$$

The inequality (6) has a number of interesting consequences. First of all, it implies that the subsample estimator is more efficient after some threshold $h$ only if $\gamma_0$ is not null,

otherwise a rational agent would always be better off by making decisions based on his own nowcasts. Conversely, the preliminary results from incomplete samples are the best choice at any time only in the limiting case in which even one single observation provides better information than any forecast, so that $\sigma_h$ is null for whatever small dissemination delay $h$. Secondly, the threshold $\frac{M_{t+h}}{M_{t+H}}$ that makes valuable the publication of preliminary results may be unexpectedly large, even when the prediction accuracy is quite poor compared to the final sample mean variance $\sigma_H$. For instance, if $\phi_0$ is as (implausibly) large as ten times $\sigma_H$, the minimum subsample for data publication would be larger than 10% of the complete sample.

## 3. The Cost of Delaying Decisions

Other than the cost of taking decisions based on inaccurate data, often economic agents also have to consider the additional cost of delaying decisions, as argued by Granger and Machina (2006). This is the typical case when the "first mover" has some advantage over the followers. For example, if the potential market is given, the first firm entering the market is able to serve the most profitable segment of demand, while the followers have to make do with supplying only the others. Furthermore, purchasing and investment decisions are usually supposed to have an optimal timing, mainly related to economic fluctuations. Winston (2008) provides a comprehensive survey of economic models in which decision timing is a major factor.

In some special cases, taking into account the cost of delaying decisions may imply that users incur smaller overall losses if they base their decisions on timely but very inaccurate nowcasts instead of delayed preliminary and final official estimates of the relevant variables. In fact, the loss of delaying decisions may grow so fast over time that agents cannot afford to wait for more accurate but late survey results.

The cost of delaying decisions, waiting for more accurate information, is presumably a function of time passed from the reference period of relevant information, say $D(h)$. The function $D(h)$ achieves its minimum at $h = 0$, when assumedly $D(0) = 0$ without any loss of generality, and the cost of delaying decisions very likely does not decrease with $h$, that is $\frac{dD}{dh} \geq 0$.

As already noted in Section 2, both the accuracy of nowcasts and surveys may vary discontinuously over time, because most valuable data and information are often gathered only at specific points in time, and these points are often unpredictable, in particular for administrative sources. However, here $F(h)$, $S(h)$ and $D(h)$ are assumed to be continuous functions of $h$ only to make the problem more tractable analytically and show the role of the cost of delaying decisions in "normal" times.

In any case, if $F(h)$ and $S(h)$ cross for the first time at the delay $h_0$, as assumed in Section 2, the rational agents exploiting only predictions incur the minimum overall loss $L_f$ at $h_f$; thus $L_f$ is approximately

$$L_f = (F(h_c) + D(h_0)) + (f' + d')(h_f - h_0) + (f'' + d'')(h_f - h_0)^2 \tag{7}$$

where $x' = \frac{dX}{dh}\Big|_{h=h_0}$ and $x'' = \frac{d^2X}{dh^2}\Big|_{h=h_0}$.

By contrast, users that base their decisions on the preliminary results of surveys face the minimum loss, say $L_s$, $h_s$ periods after the reference time, namely

$$L_s = (S(h_c) + D(h_c)) + (s' + d')(h_s - h_0) + (s'' + d'')(h_s - h_0)^2. \tag{8}$$

According to (7) and (8), the losses $L_f$ and $L_s$ achieve their minima when

$$h_f = h_0 - \frac{1}{2}\frac{f' + d'}{f'' + d''} \tag{9}$$

and

$$h_s = h_0 - \frac{1}{2}\frac{s' + d'}{s'' + d''} \tag{10}$$

that is when

$$L_f = (F(h_0) + D(h_0)) - \frac{1}{4}\frac{(f' + d')^2}{f'' + d''} \tag{11}$$

and

$$L_s = (S(h_0) + D(h_0)) - \frac{1}{4}\frac{(s' + d')^2}{s'' + d''}. \tag{12}$$

In principle, according to (11) and (12) the minimum loss could be achieved either basing decisions on forecasts or on sound statistical data, depending on the shape of the functions $S(h)$, $F(h)$ and $D(h)$. Indeed, since $F(h_0) = S(h_0)$ by definition, the condition for $L_f \leq L_s$, together with (11) and (12), implies

$$\frac{(s' + d')^2}{s'' + d''} \geq \frac{(f' + d')^2}{f'' + d''}. \tag{13}$$

(13) entails that the decision makers would be better off basing their decisions on their own predictions even when the accuracy of nowcasts improves over time only very slowly, and much slower than the results of surveys, namely when $f' \cong 0$ and $s'' \geq f''$ hold, so that (13) reads

$$s' \geq -2d' \tag{14}$$

namely if the marginal improvement of survey accuracy (i.e. $-s'$) does not exceed twice the loss attached to postponing decisions by one unit of time more (i.e. $d'$). Notably, the condition (14) derives from hypotheses that are very unfavourable to the use of forecasts and are less likely to occur in the real world. In any case, the inequality (14) fully confirms the assumption that agents prefer basing their decisions on predictions when the cost of delay increases very fast and the expected error size of surveys does not decrease too quickly over time. In the real world, nowcasts could improve quite fast, while the preliminary results of some survey may not. Thus the scope for utilising nowcasts is arguably even larger.

It is worth noting that the result (14) does not take into consideration the possibility that disseminating preliminary survey results might dramatically increase the accuracy of

forecast. Otherwise, it could happen that the minimum loss associated to predictions is always lower than that deriving from making decisions based only on survey results, since, in this case, the curve $F(h) + D(h)$ lies below $S(h) + D(h)$ by definition.

Unfortunately, $D(h)$ cannot be related to $\sigma_h$ or $\phi_h$, in contrast to $S(h)$ and $F(h)$, thus the condition (14) strictly depends on the specific decision problem faced by economic agents. As a consequence, this factor cannot be considered in the next section. Nevertheless, (14) implies that users may appreciate preliminary estimates released much earlier than $h_0$, when the accuracy of sample estimates crosses the accuracy of users' predictions.

## 4. Analysing the Releases of Quarterly GDP Estimates for the Eurozone

In the European Union, quarterly national accounts are released according to a "minimal" calendar established by EC Regulation N° 1392/2007. However, the statistical offices of the member states and Eurostat tend to provide data in an even timelier manner than prescribed by this Regulation. At the moment, three main releases are published for each quarter:

1. The first release, 45 days after the end of the reference quarter, named "flash estimate" and consisting of GDP growth estimates for the latest quarter only. No component of GDP is published at this stage;
2. The "second release" about 65 days after the end of the reference quarter, including a basic breakdown GDP. A more complete set of data, including an estimate of domestic employment, follows about ten days after the "second release";
3. The "third release" is scheduled at around 100 days after the end of the quarter. It provides more detailed breakdowns for the latest quarter.

Quarterly data are open to backwards revision at each release, and data on the previous three years are usually subject to major revisions when annual data are released by March for the "excessive deficit notification" prescribed by the European rules. Furthermore, seasonal adjustment procedures may lead to some minor revisions of quarterly data even older than three years. As a consequence, many different "vintages" of GDP estimates are available for each quarter: Combining the sequence of the three releases listed above, the GDP estimate for a given quarter is possibly subject to eleven revisions during the subsequent twelve months.

It is worth noting that the national account estimates derive from a very sophisticated process that exploits both the results of pure preliminary sample estimates on a large number of statistical indicators and a range of model-based procedures aimed at integrating missing data and treating possible outliers (European Communities (1999) reports the methodologies and best practices for estimating quarterly national accounts in Europe). Thus quarterly GDP vintages almost certainly improve their accuracy over time much faster than a sequence of pure non-Bayesian estimates from incomplete samples such as that considered in Section 2 for illustrative purposes. Thanks to the mass of nonsampling information embodied in each release of data, the $\sigma_h$ function associated to the GDP vintages can be expected to decrease faster than a pure sample estimate, so that the comparison between $\sigma_h$ and $\phi_h$ is very unfavourable to users' predictions at any time. The comparison is even less favourable to users' predictions if the growth rates in the same

quarter of the previous year are considered, since this transformation of original time series tends to reduce two major sources of revision, namely the best information on the level of GDP, mainly related to back revision of annual data, and the changes of data induced by running seasonal adjustment procedures on longer time series.

The different "vintages" of year-on-year growth rates of volume GDP, seasonally and working-day adjusted, for the twelve countries of the Eurozone are collected and published regularly on the Eurostat website, starting from the rate of 2003Q1 (the revision triangle can be downloaded from http://epp.eurostat.ec.europa.eu/portal/page/portal/national_accounts/methodology/quarterly_accounts in Excel format). Older data are considered much less comparable over time and across the member states. As of the end of 2012, the last available data on GDP revisions refer to 2011Q4 because final releases of later data are unavailable.

The so-called "triangle of revisions" published by Eurostat shows that the largest revisions of GDP estimates occur within six to nine months after the reference quarter, but in principle GDP can be revised many times for about three years after the reference quarter, following the regular revisions of annual data. In addition, seasonal adjustment procedures may induce further minor changes of data even after 3–4 years. However, no economic agent is probably in the position to wait for such a long period of time before making a decision; accordingly, here the benchmark for evaluating the accuracy of preliminary estimates has been set arbitrarily to 400 days after the reference quarter (that is after about 13 revisions), also to save degrees of freedom to carry out further statistical analysis. A comparison of real-time data with their third-year benchmark (corresponding to the latest release admitted for the excessive deficit notification) will be discussed briefly below.

The revisions of GDP represent a challenging case study for simulating the interplay between users' nowcasts and official data releases sketched in Section 2. Since this article aims at testing the possible advantages of users' estimates over current official estimates, a number of assumptions unfavourable to users' nowcasts have been adopted throughout the simulation exercise. In particular, the revision of annualized growth rates of GDP are considered, and users' estimations are simulated by using intentionally simple and inefficient procedures that exclude any piece of information other than the time series of GDP vintages.

Table 1 reports some statistics on the accuracy of preliminary estimates of GDP in the Eurozone evaluated *vis à vis* the 400-day benchmark estimated on the sample 2003Q1 to

*Table 1. The accuracy of preliminary estimates of GDP*

| | Average error | | RMSE | | 5th centile | | 95th centile | |
|---|---|---|---|---|---|---|---|---|
| Dissemination delay | Full sample | Until 2008Q2 | Full sample | Until 2008Q2 | Full sample | Until 2008Q2 | Full sample | Until 2008Q2 |
| 45 days | −0.024 | −0.052 | 0.209 | 0.179 | −0.336 | −0.336 | 0.268 | 0.268 |
| 65 days | −0.031 | −0.058 | 0.159 | 0.132 | −0.246 | −0.246 | 0.233 | 0.096 |
| 100 days | −0.022 | −0.041 | 0.137 | 0.123 | −0.224 | −0.224 | 0.192 | 0.143 |
| 101–200 days | −0.022 | −0.036 | 0.098 | 0.089 | −0.136 | −0.197 | 0.157 | 0.092 |
| 201–250 days | −0.018 | −0.026 | 0.073 | 0.071 | −0.115 | −0.148 | 0.105 | 0.070 |
| 251–350 days | −0.011 | −0.014 | 0.040 | 0.040 | −0.071 | −0.071 | 0.049 | 0.058 |

2011Q4 and on the pre-crisis subsample ranging from 2003Q1 to 2008Q2. The latter is part of a period often called the Great Moderation, because business cycle fluctuations, and average growth rate of GDP, were very weak. Thus both preliminary estimates and nowcasts were exposed only to minor unpredictable shocks. On the contrary, the post-2008 sample includes the largest economic crisis since World War II, and has provided many surprises for forecasters and statisticians.

Although the figures reported in Table 1 should be considered cautiously because only 24 degrees of freedom are available for the computation of statistics, some evidence is reasonably clear. First of all, preliminary estimates show a weak downward bias in both periods, although not significant from a pure statistical point of view, possibly because the statistical offices are usually more concerned with overestimating GDP growth rates rather than with revising the data upward during the following years. Strikingly, this evidence was even stronger before the last economic crisis, supporting the view that the accuracy of official estimates has not been influenced overly by the large adverse shocks that hit the economy after 2008Q2. In any case, the negative bias tends to vanish as the delay of preliminary estimates increases from 45 days to 250 days and over.

Also Table 1 shows that the root mean square error (RMSE) of preliminary estimates decreases quite fast, as conjectured in Section 2: In the full sample it falls from 0.21 percentage points for the flash estimates to 0.04 percentage points for the oldest vintage considered here; before the crisis, the RMSE ranged from 0.18 to 0.04, that is not much lower than the same statistic calculated for the full sample of data. This evidence supports the hypothesis that the preliminary estimates of GDP are very robust to large shocks. Furthermore, in nine cases out of ten, between 2003 and 2011, the revisions range from $-0.34$ to 0.27 percentage points for the flash estimates, and only from $-0.07$ to 0.06 percentage points for the 250–350 day releases, and the analysis of the pre-crisis period reveals similar results.

The same evidence is confirmed by the nonparametric estimate of the function $\sigma_h$ reported in Figure 1, even again taking into account that few degrees of freedom are available particularly for the estimation on longer dissemination delays. The local second-degree

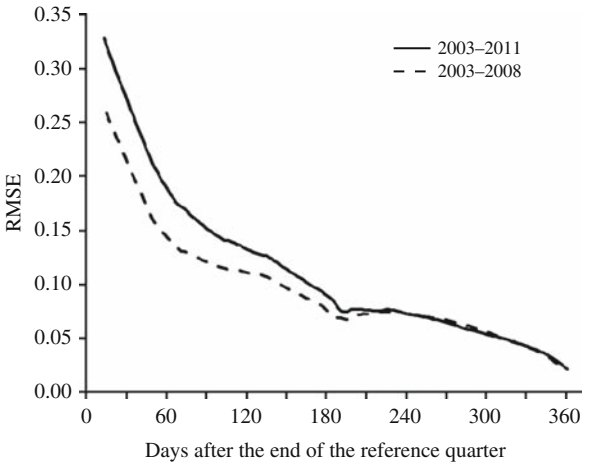



*Fig. 1. The accuracy of quarterly GDP preliminary estimates*

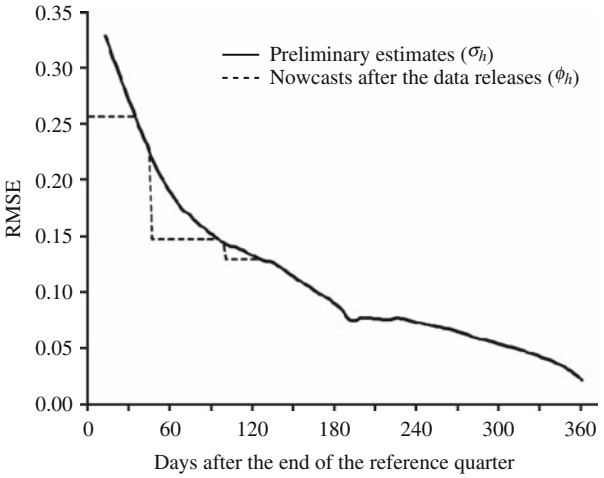



*Fig. 2. Comparing the accuracy of nowcasts and preliminary estimates*

polynomial estimator described by Fan (1992) has been adopted. For each vintage, the interpolation is based on a series of weighted least square estimators in which the observations close to the reference vintage are weighted by a "kernel function". The "bandwidth" of the weighted observations has been determined according to the formula proposed by Fan and Gijbels (1996). The main drawback of this methodology is that it assumes a continuity of interpolated functions that could be unrealistic, as argued in Section 2.

Furthermore, according to nonparametric analysis, the RMSE of revisions decreases with the dissemination delay both during the pre-crisis period and the full period 2003–2011, even though the decline is faster during the first 60–90 days and slower afterwards, supporting the view that the statistical offices are able to exploit the most informative data by the beginning of the estimation process. The virtual RMSE of preliminary data released just at the end of the reference quarter would be 0.33 percentage points, almost 50% larger than the actual RMSE of flash estimates. Nevertheless, this value is likely underestimated, since it definitively comes from a purely backwards extrapolation of the observed rate of changes of $\sigma_h$ between 45 and 65 days after the end of the reference quarter, and is not consistent with (3) and (4). In any case, the RMSE of preliminary estimates apparently halves within about 80 days, regardless of the period considered, and divides by four within about 180 days.

Most results found comparing GDP revisions to their 400-day benchmarks are confirmed by considering the three-year benchmarks instead, although the degrees of freedom for estimating RMSE and other statistics drop dramatically. In particular, $\sigma_h$ is still decreasing as the dissemination delay increases, although the RMSE of flash estimates picks up to 0.371, about 80% more than the RMSE computed versus the 400-day benchmark. The latter benchmark is still subject to revisions, the average size of which is 0.252 percentage points during the next 600 days. The nonparametric estimation shows that the $\sigma_h$ curve is almost parallel to the one computed for the 400-day benchmark, beyond the second release of data. Detailed descriptive and nonparametric statistics for the three-year benchmark are not reported here for sake of brevity, and are available from the author.

In order to compare the pure statistical estimates to users' predictions and nowcasts, a forecasting model has been assumed. To make the exercise more challenging, the forecasts made before the end of each reference quarter and the adjusted nowcasts based on the following preliminary estimates are simulated by using intentionally very simple time series models, estimated inefficiently by means of ordinary least squares on real data available at the moment of each simulation. This procedure intends to mimic the actual behaviour of an unsophisticated user who exploits only official information readily available on GDP and disregards any other evidence, such as timely short-term statistics, "soft data" on business and household confidence, possible private information, and so on. Thus, in principle, the experiment is strongly biased toward the superiority of official estimates and, in principle, should support the actual data dissemination policy adopted by Eurostat, since subsequent official estimates potentially embody more information than that used by the imaginary naïve user considered in our simulation.

Some insight on actual accuracy of forecasts and nowcasts on GDP made by more sophisticated users is provided by Barhoumi et al. (2008), Diron (2008), Angelini et al. (2011) and Frale et al. (2011), who developed very short-term forecasts and nowcasts of Eurozone GDP, and by Pain and Sédillot (2005), who applied similar methods to other OECD countries. Joint nowcasts and short-term forecasts of inflation and GDP were proposed by Giannone et al. (2008). In those papers, the RMSE of nowcasts based on real-time information likely available to data users ranges from 0.2 to 0.6 percentage points, with a gain of using additional information peaking as high as 40% of naïve predictions. Jansen et al. (2012) also estimated that consensus forecasts collected by ECB among experts are slightly more accurate, with a further cut of $\phi_h$ by 10% compared to the best statistical models.

In this simulation experiment, the forecast $Y^*_t$ on the yearly growth rate of GDP made before the end of the reference quarter $t$ derives from the simple AR model

$$Y^*_t = c_{t,v} + a_{t,v} Y_{t-l,v} + u_{t,v} \tag{15}$$

where $c_{t,v}$, and $a_{t,v}$ are parameters estimated by using only the latest data available at time $t-1$, not including $Y_t$; $Y_{t-1,v}$ is the latest release of the GDP growth rate at time $t-1$; $u_{t,v}$ is a random disturbance, likely autocorrelated, being a forecasting error, and possibly heteroscedastic. Even though the assumed characteristics of $u_{t,v}$ would require appropriate methods for estimating (15) efficiently, our imaginary user is supposed to use only ordinary least squares. In any case, scarce degrees of freedom available for the simulation (on occasion less than ten) would make it unfeasible to use other proper estimation methods. This practice creates forecasts even worse than those possibly produced by the Model (15) itself. In order to allow some degree of freedom to the estimates, the results of the first ten regressions have been discarded and the first forecasting period has been set to 2006Q1. The first column of Table 2 reports the main results of the regression run to predict GDP at the end of the simulation period. It is apparent that the naïve model (15) fits the data quite well, even though the RMSE is as large as the average yearly growth of GDP during the last decade, mainly due to very few large outliers. Furthermore, forecasts tend to revert to the average after each deviation, as the estimate of the parameter $a_t$ is significantly below 1; thus, in principle, the model is incapable of predicting sudden turning points of GDP growth rate correctly. The coefficient of the dummy variable is not

significant, at least for the last estimation period. The results of the regressions run to produce the forecasts and nowcasts for each point in time are not reported here and are available on request.

When a new release of GDP figures, say $Y_{t,v}$, is published, users can improve their nowcast of the GDP growth by also taking into account the previous revisions and the past dynamics of GDP. In this exercise, this "adjusted" estimate, say $Y^*_{t,v}$, has been simulated by using the model

$$Y^*_{t,v} = c_{t,v} + a_{t,0,v-l}Y_{t,v-l} + a_{t,l,v-l}Y_{t-l,v-l} + a_{t,l,v}Y_{t-l,v} + d_{t,v}D_t + v_{t,v}, \qquad (16)$$

where the parameters are estimated on the sample of data actually available when the $(v-1)th$ vintage of data is released; $D_t$ is a dummy variable that is 1 at time $t-1$ and zero elsewhere, which serves to "sterilize" the forecast from the interpolation error made at time $t-1$. The rationale for (16) is that the revisions of GDP are hardly ever purely random and serially uncorrelated, so that there is room for improving the accuracy of the official estimates by also taking into account the typical time series structure of revisions. Similar evidence is also reported by Fixler and Grimm (2006) for the US GDP, and by Frale and Raponi (2012) for the case of Italy.

The main results of estimating the models (15) and (16) from the largest available samples are reported in Table 2. It is apparent that every model fits the data quite well, but there is strong evidence that the models are over-parameterized. In fact, regression results show that only the coefficients of $Y_{t,v-1}$ are statistically significant. In addition, the coefficients of $Y_{t,v-1}$ are higher than 1 at any reasonable confidence level, confirming the tendency of statistical offices to revise GDP growth upwards at each release of data. Apparently the underestimation is fairly substantial, ranging from 6% for the flash estimates to 3% for the 100-day releases. By contrast, other regressors are not statistically significant: This result was expected for the dummy variable, which serves only to sterilize the effects of possible outliers in the most recent estimation period, while it is unexpected for $Y_{t-1,v-1}$ and $Y_{t-1,v}$. Indeed, the few available degrees of freedom of estimates and the strong collinearity

*Table 2. The main results of regressions used to simulate users' nowcasts and forecasts (the statistics refer only to the longest sample available for each model)*

| Regressors | Forecast one quarter ahead | Data release | | |
|---|---|---|---|---|
| | | 45 days | 65 days | 100 days |
| $Y_{t,v-1}$ | | 1.059 | 1.033 | 1.030 |
| | | (0.014) | (0.010) | (0.011) |
| $Y_{t-1,v-1}$ | | −0.044 | 0.297 | 0.013 |
| | | (0.155) | (0.185) | (0.200) |
| $Y_{t-1,v}$ | 0.895 | −0.008 | −0.313 | −0.039 |
| | (0.079) | (0.156) | (0.187) | (0.205) |
| Dummy variable | | 0.018 | −0.002 | 0.174 |
| | | (0.074) | (0.050) | (0.549) |
| Constant | 0.133 | −0.011 | −0.015 | −0.005 |
| | (0.199) | (0.014) | (0.010) | (0.010) |
| Adjusted $R^2$ | 0.795 | 0.999 | 0.999 | 0.999 |
| RMSE | 1.036 | 0.072 | 0.048 | 0.051 |

*Standard error of estimates in parentheses.*

between the regressors may mask the true influence of those variables. In fact, excluding them from the regressions significantly worsens the accuracy of adjusted nowcasts.

The overall performance of one-step-ahead forecasts and adjusted nowcasts are summarised in Table 3. The most interesting result is that, excluding a single large forecast error in 2009Q2 (about three percentage points below the true value), the predictions made before the end of the reference quarter are unexpectedly accurate, although they are intentionally naïve and extrapolation based. In fact, simulated users' forecasts are less downward biased than most preliminary estimates and exhibit a RMSE that is roughly comparable to flash estimates. This evidence merits further attention, since the simulation period comprises the data on the last global crisis, when large unexpected shocks hit the European economy and a "double dip", including three turning points, occurred. However, Model (15) also produced a number of large positive and negative errors compared to the preliminary official estimates, as confirmed by the value of the 5th and 95th centiles of the distribution of errors that almost doubled the corresponding statistics computed for the flash estimates. As a result, comparing the accuracy of the simulated forecasts to nonparametric interpolation of $\sigma_h$ it emerges that even naïve users' predictions would be able to compete against preliminary estimates of GDP possibly released about 30 days after the end of the reference quarter. This is really surprising, even taking into account that the out-of-sample interpolation of $\sigma_h$ likely underestimates the accuracy of estimates when $h$ is below the first dissemination delay actually observed. The horizontal piece of the dashed line in Figure 2 shows how early the simulated $\phi_h$ function crosses the $\sigma_h$ function for the first time.

As argued in Section 2, this is a situation in which very timely official releases of GDP data, for instance just several weeks after the end of the reference quarter, would not be as "competitive" from the point of view of a representative economic agent. Nevertheless, if Eurostat decided to release such data, users might exploit this new piece of information elaborating even better nowcasts, hopefully surpassing their previous projections one step ahead.

In fact, the second row of Table 3 suggests that when flash estimates are published, users are able to greatly improve the accuracy of their adjusted nowcasts. The RMSE of nowcasts based on flash estimates is located amid the RMSEs of the official estimates of GDP released respectively 65 days and 100 days after the reference period. More precisely, Figure 2 suggests that after the flash estimates, the $\phi_h$ function shifts downwards

*Table 3. The accuracy of nowcasts and adjusted preliminary estimates of GDP*

| | Average error | RMSE | 5th centile | Median | 95th centile |
|---|---|---|---|---|---|
| Pure forecast one quarter ahead[(a)] | −0.012 | 0.262 | −0.449 | −0.015 | 0.542 |
| Adjusted preliminary estimates | | | | | |
| 45 days | −0.013 | 0.147 | −0.239 | −0.020 | 0.189 |
| 65 days | −0.004 | 0.148 | −0.165 | −0.049 | 0.182 |
| 100 days | 0.004 | 0.130 | −0.137 | −0.025 | 0.205 |

*The statistics are computed on the sample 2006Q1–2011Q4 to have at least a ten observations for running each regression.*

[(a)]Excluding only the large forecast error on 2009Q2 (−2.931). Considering the full sample, the average error is −0.234 and the RMSE is 1.444.

substantially and intersects with the $\sigma_h$ function when the dissemination delay is 95 days. In contrast, when the 65-day official estimates are released, users' adjusted nowcasts do not improve much, as is apparent from the third row of Table 3. Therefore, the second release of GDP data has very likely only a minor impact on users' decisions based on the dynamics of output in the Eurozone. However, the 65-day release of data includes a breakdown of data that expectedly improve the information set available to economic agents; thus the second release of GDP is welcomed by users focusing on sectorial dynamics rather than on the overall economic performance of the Eurozone.

The publication of the third release of quarterly GDP, 100 days after the end of the reference quarter, seems to increase the accuracy of users' nowcasts further, as Figure 2 and the last row of Table 3 make evident. Nevertheless, the improvement is relatively too small to change users' decisions significantly, so that they could became "rationally inattentive" as argued by Sims (2003), even if the accuracy of later official data releases of GDP is expected to increase. In any case, evidence for longer dissemination delays could be influenced by the scarce degrees of freedom available for estimation.

To summarise, the approach proposed in Section 2 and the empirical evidence presented in this section suggest producing a very early estimate of GDP as soon as possible before the first official release, possibly after a few weeks, followed by a second release only 3–4 months later. Noticeably, the thresholds above were determined assuming that the typical user of data does not make use of very sophisticated forecasting methods and large information sets, and that after each nowcast, made when official data are disseminated, the nowcast does not improve further. Otherwise, the horizontal pieces of the dashed line in Figure 2 would be downward sloped, so that $\phi_h$ would cross the curve of the accuracy of official releases later than 30 or 100 days after the end of the reference quarter. In addition, the cost of waiting, considered in Section 3, could prompt Eurostat to disseminate even more timely data to better meet the needs of users that are not in the position to wait overly long before making their decisions.

A parallel simulation exercise carried out on data revisions versus the three-year benchmark provided very similar results, although taking into account the drop in the degrees of freedom available to simulate users' nowcasts. Indeed, this outcome was almost fully expected, since the revisions made during the first 400 days are most likely uncorrelated to those occurring in the following two years, which are related mainly to the availability of very detailed structural information available only after years, and are likely less correlated to the short-term indicators mostly used to compute earlier estimates of GDP. Thus the size of revisions over the two benchmarks differs almost by a constant term, roughly explained by the difference between the 400-day estimate and the "definitive" 1000-day data release, as remarked on above. Given that users' nowcasts cannot depend on data available only in the future, their accuracy versus the definitive data worsens only by a constant term as well, so that the relative comparison versus the accuracy of official data is almost unchanged. Full details of this experiment are available from the author.

## 5. Concluding Remarks

By regarding the results of statistical surveys as an input for decisions, we are able to provide some guidelines in adjusting the calendar to users' needs for data release.

In general, rational agents would appreciate less accurate data in advance instead of delayed perfect statistics, and the "impatience" of agents depends mainly on their capacity to make reliable early estimates of the relevant variables autonomously. In fact, provisional data are assumed to improve agents' decisions only if the data are capable of enhancing their own estimates and forecasts. Otherwise, rational agents would be better off continuing to base their decisions on their extrapolations. It follows that the size of forecast errors should be an important benchmark for statistical institutes in deciding when data should be released, taking into account the forecasting capability of "representative" data users, including government and professional users. As a consequence, regular surveys of users' nowcasts could be helpful in enhancing current release calendars.

The real-data simulation experiment presented in Section 4 shows how the proposed approach may help to improve the current dissemination calendar of quarterly Euro area GDP. In particular, "flash" estimates seem only slightly more accurate than naïve users' forecasts made during the reference quarter, thus earlier (and coarser) releases would very likely be appreciated by users since such data could improve their nowcasts. By contrast, the intermediate release of data 65 days after the end of the reference quarter apparently is less informative on the current dynamics of GDP, since the data's accuracy does not surpass the nowcasts already based solely on flash estimates. Of course, the breakdown of data provided by the second release is almost certainly valuable. In any case, statistical offices should balance such suggestions with the cost of producing more estimates and their institutional duties. The empirical evidence presented in Section 4 also suggests that there is only little scope for users to wait for definitive data published after three years before making their decisions, since the revisions made beyond 400 days after the reference quarter are generally small, apart from general methodological changes that appear virtually independent compared to the first revisions.

Further support for statistical agencies disseminating preliminary results of their surveys comes from the fact that rational agents often balance the cost of making a decision based on inaccurate data with the cost of delaying their decisions. If timing is crucial in making a decision, even very noisy and inaccurate preliminary data would be appreciated under most circumstances. However, according to the approach sketched above, designing a dissemination calendar requires first of all the identification of the forecasting ability of and the cost of postponing decisions to a "representative" data user. Notably, this conceptual framework seems fully consistent with the 11th principle of the European Statistics Code of Practice that states: "User satisfaction is monitored on a regular basis and is systematically followed up", as well as with the 13th principle that provides that "Preliminary results of acceptable aggregate accuracy can be released when considered useful."

In principle, the release of preliminary and final statistical data could be adapted dynamically to the possible changes of the accuracy of nowcasts, the variance of sample estimates and the cost of delaying decisions. Since predictions hopefully improve over time, the publication of preliminary estimates from incomplete samples should be anticipated progressively. Furthermore, even less accurate statistical data about, for instance, the turning points of the business cycle could be appreciated by users when their forecasts become more uncertain. Nevertheless, such a flexible dissemination policy would not comply with statistical offices' commitment to following a fully predictable

strategy in order to strengthen their credibility and independence. Moreover, data "inflation" could impair users, raising their search costs. Nevertheless, there is still room for flexibility in data release, provided that "Statistical release dates and times are preannounced", as stated by the 6th principle of the European Statistics Code of Practice, and "[d]ivergence from the dissemination time schedule is publicised in advance, explained and a new release date set", as pointed out by the 13th principle.

The comparison of users' estimates versus official preliminary sample estimates may also help official statisticians to decide the timing for the dissemination of disaggregated data. In fact, agents who need a given breakdown of data to make a decision, for example at $N$ "digits" level of the NACE classification of economic activity, necessarily compare the loss associated to the use of preliminary survey results at $N$ digit level, say $S^*(N)$, to the loss of using some model-based estimation which exploits only data already available, such as data broken down at $N-n$ digits, say $F^*(N-n)$. Thus, at time $t+h$, statistical data disaggregated at level $N$ would be long-awaited by agents only if $F^*(N-n) \geq S^*(N)$; otherwise users would be better off if statistical agencies had released earlier data, disaggregated at level $N-n$ instead, that improve users extrapolations. However, more research is probably needed to thoroughly investigate the issue of how and when preliminary disaggregated data should be disseminated.

Further refinements of the approach presented in this article and many more simulation experiments are required before implementing these concepts in official statistics. In particular, the cost of delaying decisions should be quantified to be compared to the loss related to the inaccuracy of data utilised in the decision process. Furthermore, the advantages of model-based preliminary estimates directly released by the statistical offices, also exploiting internal and confidential information sources, should be explored, although this practice is often criticized by those defending a strict separation between official statistics and forecasting. Finally, an extensive analysis of releases of other statistical indicators is required. In any case, the suitability of releasing earlier preliminary data should be balanced with other considerations sketched above, mainly concerning the institutional role of statistical offices and the cost incurred by users in collecting and elaborating more information.

## 6. References

Altavilla, C. and M. Ciccarelli. 2007. "Information Combination and Forecast (st)ability. Evidence from Vintages of Time-Series Data." *Working Paper Series ECB,* No. 864. Available at: http://ideas.repec.org/e/pal73.html (accessed July 31, 2014).

Angelini, E., G. Camba-Mendez, D. Giannone, L. Reichlin, and G. Rünstler. 2011. "Short-Term Forecasts of Euro Area GDP Growth." *The Econometrics Journal* 14: 25–44. DOI: http://dx.doi.org/10.1111/j.1368-423X.2010.00328.x.

Barhoumi, K., S. Benk, R. Cristadoro, A. Reijer, P. Jakaitiene, P. Jelonek, and A. Rua. 2008. "Short-Term Forecasting of GDP Using Large Monthly Datasets: a Pseudo Real-Time Forecast Evaluation Exercise." *Occasional Paper Series ECB,* No. 84. Available at: http://ideas.repec.org/p/ecb/ecbops/20080084.html (accessed July 31, 2014).

Blanchard, O.J., J.P. L'Huillier, and G. Lorenzoni. 2009. "News, Noise, and Fluctuations: An Empirical Exploration." *NBER Working Paper,* No. w15015, Available at: http://www.nber.org/papers/w15015 (accessed July 31, 2014).

Clemen, R. 1989. "Combining Forecasts: a Review and Annotated Bibliography." *International Journal of Forecasting* 5: 559–583. DOI: http://dx.doi.org/10.1016/0169-2070(89)90012-5.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching: Theory and Practice*. New York: Wiley.

Diron, M. 2008. "Short-Term Forecasts of Euro Area Real GDP Growth: An Assessment of Real-Time Performance Based on Vintage Data." *Journal of Forecasting* 27: 371–390. DOI: http://dx.doi.org/10.1002/for.1067.

European Central Bank 2009. "Revisions to GDP Estimates in the Euro Area." *Monthly Bulletin* 4: 85–90. Available at: https://www.ecb.europa.eu/pub/pdf/mobu/mb200904en.pdf (accessed July 31, 2014).

European Communities 1999. *Handbook on Quarterly National Accounts*, Luxembourg: Office for Official Publications of the European Communities. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/CA-22-99-781/EN/CA-22-99-781-EN.PDF (accessed July 31, 2014).

European Statistics Code of Practice 2011. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF.

Fan, J. 1992. "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association* 87: 998–1004. DOI: http://dx.doi.org/10.2307/2290637.

Fan, J. and I. Gijbels. 1996. *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

Fixler, D.J. and B.T. Grimm. 2006. "GDP Estimates: Rationality Tests and Turning Point Performance." *Journal of Productivity Analysis* 25: 213–229. DOI: http://dx.doi.org/10.1007/s11123-006-7640-x.

Frale, C. and V. Raponi. 2012. "Revisions in Official Data and Forecasting." *Working Papers of Dipartimento del Tesoro,* No. 3.

Frale, C., M. Marcellino, G.L. Mazzi, and T. Proietti. 2011. "EUROMIND: A Monthly Indicator of the Euro Area Economic Conditions." *Journal of the Royal Statistical Society: Series A* 174: 439–470. DOI: http://dx.doi.org/10.1111/j.1467-985X.2010.00675.x.

Giannone, D., L. Reichlin, and D. Small. 2008. "Nowcasting: The Real-Time Informational Content of Macroeconomic Data." *Journal of Monetary Economics* 55: 665–676. DOI: http://dx.doi.org/10.1016/j.jmoneco.2008.05.010.

Graham, P., J. Young, and R. Penny. 2009. "Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models." *Journal of Official Statistics* 25: 245–268.

Granger, C.W.J. and M.J. Machina. 2006. "Forecasting and Decision Theory." In *Handbook of Economic Forecasting*, edited by G. Elliott, C.W.J. Granger, and A. Timmermann. Amsterdam: Elsevier.

Granger, C.W.J. and M.H. Pesaran. 2000. "Economic and Statistical Measures of Forecast Accuracy." *Journal of Forecasting* 19: 537–560. DOI: http://dx.doi.org/10.1002/1099-131X(200012)19:7<537:AID-FOR769>3.0.CO;2-G.

Jansen, W.J., X. Jin, and J. de Winter. 2012. "Forecasting and Nowcasting Real GDP: Comparing Statistical Models and Subjective Forecasts." *De Nederlandsche Bank Working Paper,* No. 365. Available at: http://www.dnb.nl/en/binaries/Working%20Paper%20365_tcm47-283164.pdf (accessed July 31, 2014).

Little, R.J.A. 2012. "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics* 28: 309–334.

Little, R.J.A., F. Liu, and T.E. Raghunathan. 2004. "Statistical Disclosure Techniques Based on Multiple Imputation." In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, edited by A. Gelman and X.L. Meng, 141–152. New York: John Wiley & Sons.

Pain, N. and F. Sédillot. 2005. "Indicator Models of Real GDP Growth in the Major OECD Economies." *OECD Economic Studies,* No. 40. Available at: http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=4a74b653-6721-45c4-897d-9aa24c0c2037%40sessionmgr113&vid=2&hid=128. (accessed July 31, 2014).

Särndal, C.-E. and S. Lundström. 2005. *Estimation in Surveys With Nonresponse*. New York: John Wiley & Sons.

Sims, C.A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50: 665–690. DOI:http://dx.doi.org/10.1016/S0304-3932(03)00029-1.

UNSTAT, 2009. "International Seminar on Timeliness, Methodology and Comparability of Rapid Estimates of Economic Trends." Available at: http://unstats.un.org/unsd/nationalaccount/workshops/2009/ottawa (accessed July 31, 2014).

Winston, G.C. 2008. *The Timing of Economic Activities*. Cambridge: Cambridge University Press.

Yang, Y. and H. Zou. 2004. "Combining Time Series Models for Forecasting." *International Journal of Forecasting* 20: 69–84. DOI: http://dx.doi.org/10.1016/S0169-2070(03)00004-9.

Received April 2012
Revised March 2014
Accepted March 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 521–532, http://dx.doi.org/10.2478/JOS-2014-0032

# Developing Calibration Weights and Standard-Error Estimates for a Survey of Drug-Related Emergency-Department Visits

*Phillip S. Kott*[1] *and C. Daniel Day*[2]

This article describes a two-step calibration-weighting scheme for a stratified simple random sample of hospital emergency departments. The first step adjusts for unit nonresponse. The second increases the statistical efficiency of most estimators of interest. Both use a measure of emergency-department size and other useful auxiliary variables contained in the sampling frame. Although many survey variables are roughly a linear function of the measure of size, response is better modeled as a function of the log of that measure. Consequently the log of size is a calibration variable in the nonresponse-adjustment step, while the measure of size itself is a calibration variable in the second calibration step. Nonlinear calibration procedures are employed in both steps. We show with 2010 DAWN data that estimating variances as if a one-step calibration weighting routine had been used when there were in fact two steps can, after appropriately adjusting the finite-population correct in some sense, produce standard-error estimates that tend to be slightly conservative.

*Key words:* Frame variable; response model; prediction model; general exponential model; finite population correction.

## 1. Introduction

The Drug Abuse Warning Network or DAWN (Substance Abuse and Mental Health Services Administration 2012) was a national stratified random sample of US hospitals used to estimate annual drug-related emergency-department visits and related statistics. This article describes a calibration-weighting strategy for the DAWN that was never implemented because the survey was discontinued after 2012. Nevertheless, we feel this strategy and our contemplated approach to variance/mean squared error estimation contained some innovative features worth sharing.

The DAWN sample was drawn from a list frame provided by the American Hospital Association (AHA). The frame was stratified by location, size, and ownership type (public *vs.* private). Hospitals were oversampled within 13 metropolitan areas, for which domain estimates were published when respondent sample sizes were deemed large enough.

[1] RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. Email: pkott@rti.org
[2] Substance Abuse and Mental Health Services Administration, 1 Choke Cherry Road, Rockville MD 20857, U.S.A. Email: charles.day@samhsa.hhs.gov

**Acknowledgments:** The views expressed in this article do not necessary reflect those of the Substance Abuse and Mental Health Services Administration. The authors express their gratitude to an associate editor and several referees whose comments improved the presentation considerably.

In the estimation strategy used operationally for DAWN, the weight for a respondent began with the hospital's design weight. A nonresponse adjustment factor was applied to each weight to account for those hospitals that were sampled but did not participate in the DAWN survey. This was followed by a sample balancing – often called a "poststratification" adjustment – to improve the efficiency (reduce the variances) of most of the resulting nearly (i.e., asymptotically) unbiased estimates. Both steps employed simple weighting-class adjustments requiring *ad hoc* collapsing schemes when there were too few respondents in a class or the class adjustment factor was deemed too large.

In this article, we will describe alternative approaches to these two adjustments. For simplicity, we will ignore the subsampling of visits and visit-level nonresponse adjustments that took place within some DAWN hospitals.

The new nonresponse adjustment factors use a calibration-weighting routine that implicitly models the probability that a hospital responds to (participates in) the DAWN survey. It does this by assuming hospital response is a function of its characteristics, such as its size, measured by annual emergency-department (ED) visits on the AHA frame, ownership (public or private), region, and the population density of the county in which it is located. If the response model is correctly specified, as we assume it is, then employing this calibration-weighting routine produces nearly unbiased estimates of DAWN totals.

The new sample-balancing adjustment factors are produced using a version of nearly pseudo-optimal calibration (Kott 2011) that forces each final weight to no less than 1. Sample balancing exploits the fact that the variables measured by the DAWN survey, such as annual drug-related ED visits, are functions of characteristics known for all hospitals on the AHA frame. Calibrating the respondents' weights so that the estimated totals of (some of) those characteristics computed from the respondent sample exactly equal corresponding frame (AHA) totals tends to increase the efficiency of estimated DAWN totals, which remain nearly unbiased.

Evaluation of the nonresponse pattern in DAWN data from 2010 lead us to treating the hospitals from the 13 metropolitan areas as one subpopulation and the remaining hospitals as a separate subpopulation. For brevity's sake we restrict our attention in this article to nonresponse modeling and weight adjustments for the former subpopulation. Similar methods can be used for the subpopulation of remaining hospitals. The impact of finite-population correction on variance estimation is much less of an issue in that subpopulation.

Although the DAWN published domain estimates for many of the 13 metropolitan areas, we investigated domain estimates within the four US census regions instead. This kept the respondent sample sizes within domains more respectable given that much of the theory underpinning calibration weighting is asymptotic.

Since many DAWN hospitals were sampled with certainty (before nonresponse), we restrict our attention in this article to linearization-based variance estimators of nearly unbiased estimated totals that require finite population correction. Most software designed to estimate variances using linearization-based methods only capture the *increase* in variance from the respondent sample size being smaller than the before-nonresponse sample size and from the final weights being more variable than the original weights. We will describe linearization-based methods that also capture the *decrease* in variance resulting from hitting calibration targets as well as from finite population correction.

The software package SUDAAN 11® (RTI 2012) can produce linearization-based measures that estimate variances appropriately when there is a single step of calibration weighting, but not (easily) when there are multiple calibration steps. We will discuss a simplified variance estimator for the DAWN given our two-step calibration scheme that can be implemented in SUDAAN 11. The resulting estimated variances tend to be slightly conservative when applied to DAWN data from 2010.

Calibration weighting for the DAWN is discussed in Section 2. Section 3 addresses variance estimation after calibration weighting. Section 4 contrasts alternative variance estimators using DAWN data, while Section 5 offers some concluding remarks.

## 2. Calibration Weighting for the DAWN

### *2.1. Nonresponse Adjustment*

Let $d_k$ be the design weight for a sampled DAWN hospital $k$. For our purposes, this was the population size of the stratum (say $h$) containing $k$ divided by its sample size ($N_h/n_h$). The strata within a metropolitan area were determined by size class (up to three within an area) and ownership type.

Following Folsom (1991), our nonresponse-adjusted weight for a DAWN respondent $k$ has the form:

$$a_k = d_k\left[1 + \exp\left(\mathbf{g}^T\mathbf{x}_k\right)\right], \tag{1}$$

where $\mathbf{x}_k$ is a vector of the respondent's characteristics to be described shortly, and $\mathbf{g}$ is determined using Newton's method (successive linear approximation) so that the calibration equation

$$\sum_R a_j\mathbf{x}_j = \sum_R d_j\left[1 + \exp\left(\mathbf{g}^T\mathbf{x}_k\right)\right]\mathbf{x}_j = \sum_S d_j\mathbf{x}_j \tag{2}$$

holds where $R$ is the respondent sample and $S$ the sample before nonresponse.

The value

$$p_k = p(\mathbf{g}^T\mathbf{x}_k) = 1/\left[1 + \exp\left(\mathbf{g}^T\mathbf{x}_k\right)\right]$$

implicitly estimates the probability that $k$ is a respondent given its characteristics in vector $\mathbf{x}_k$.

Although $p(\mathbf{g}^T\mathbf{x}_k)$ is a logistic function of $\mathbf{g}^T\mathbf{x}_k$, this method in not the same as finding $\mathbf{g}$ using either maximum likelihood (i.e., so that $\sum_S\left\{\left[1 + \exp\left(\mathbf{g}^T\mathbf{x}_k\right)\right]^{-1} - I_j\right\}\mathbf{x}_j = \mathbf{0}$, where $I_j = 1$ if $j \in R$ and 0 otherwise) or quasi-maximum (i.e., so that $\sum_S d_j\left\{\left[1 + \exp\left(\mathbf{g}^T\mathbf{x}_k\right)\right]^{-1} - I_j\right\}\mathbf{x}_j = \mathbf{0}$). Kim and Riddles (2012) show why the calibration approach in Equation (2) can lead to estimated totals with smaller variances than maximum-likelihood-based alternatives.

Preliminary analyses of 2010 DAWN data strongly suggested that the probability of response was better modeled as the log of the AHA emergency-department visits than as a direct function of ED visits. This is a more sensible result than it may appear to be. It means that a one percent increase in the size measure lead to an $r$ percent increase in the odds of response, all other things being equal.

After extensive model searching, we ultimately assumed unit response to be a logistic model of an $\mathbf{x}_k$ vector containing the log of the number of AHA emergency-department visits, which we denote $\log(q_k)$, dummy variables for each of the 13 metropolitan areas, $d_{1k}, \ldots, d_{13k}$, an indicator for a public (as opposed to private) hospital, $d_{Pk}$, an interaction term between the public indicators and one of the area dummies $d_{Pk}\, d_{13k}$, and the log of the population density within the ZIP code containing the hospital (from US Census Bureau 2012) with imputation of missing values when needed, $t_k$. Note that $q_k$ must always be positive, which it was, so that $\log(q_k)$ can be defined.

Although we assume we know the correct form of the model governing the response probabilities for each hospital, $\rho_k = p(\boldsymbol{\gamma}^T\mathbf{x}_k) = 1/[1 + \exp(\boldsymbol{\gamma}^T\mathbf{x}_k)]$, we can only estimate the parameter $\boldsymbol{\gamma}$ with $\mathbf{g}$ in Equation (2). We further assume that whether or not a hospital $k$ responds given $\mathbf{x}_k$ is independent of whether another hospital responds.

### 2.2. *Sample Balancing*

Like most government surveys, the DAWN produces a number of estimates. It is possible that a weight adjustment will decrease the variances of some estimates while increasing those of others. Nevertheless, we chose to focus our sample-balancing efforts on reducing the variance of a single estimate: the total number of drug-related emergency-department visits. This can be viewed as the "flagship" variable of the DAWN survey. Not only is it important in its own right, but it is related to many of the DAWN survey variables.

Using the nonresponse-adjusted weights from the previous step (the $a_k$), ignoring strata (and thus the need to collapse strata with only a single responding hospital) but otherwise using a routine sensitive to the sampling design, we fit linear models of drug-related emergency-department visits, $y_k$, using covariates available on the AHA frame.

The model we liked best effectively modeled not $y_k$ but $y_k/q_k$ as a function of four census-region dummies, $u_{1k}, \ldots, u_{4k}$, $\log(q_k)$, and $u_{1k}d_{Pk}$ through $u_{4k}d_{Pk}$. Observe that $y_k/q_k$ is the ratio of the number of drug-related emergency-department visits to a proxy of all emergency-department visits (using a previous year's data). The final model fit $y_k$ as a linear function of $q_k u_{1k}, \ldots, q_k u_{4k}$, $q_k \log(q_k)$, and $q_k u_{1k} d_{Pk}$ through $q_k u_{4k} d_{Pk}$.

Following the advice in Kott (2011), we set final calibration weights at

$$w_k = a_k \frac{\ell_k(u_k - 1) + u_k(1 - \ell_k)\exp(B_k[a_k - 1]\mathbf{h}^T\mathbf{z}_k)}{(u_k - 1) + (1 - \ell_k)\exp(B_k[a_k - 1]\mathbf{h}^T\mathbf{z}_k)}, \tag{3}$$

where $\mathbf{z}_k = (q_k u_{1k}, \ldots, q_k u_{4k}, q_k \log(q_k), q_k u_{1k} d_{Pk}, \ldots, q_k u_{4k} d_{Pk})^T$, $B_k = (u_k - \ell_k)/[(1 - \ell_k)(u_k - 1)]$, $\ell_k = 1/a_k$, and $\mathbf{h}$ is found so that the calibration equation, $\sum_R w_j \mathbf{z}_j = \sum_U \mathbf{z}_j$, holds.

The fraction on the right-hand side of Equation (3) is a particular version of the general exponential model of Folsom and Singh (2000):

$$f(\mathbf{h}^T\boldsymbol{\delta}_k; u_k, c_k, \ell_k) = \frac{\ell_k(u_k - c_k) + u_k(c_k - \ell_k)\exp(A_k\mathbf{h}^T\boldsymbol{\delta}_k)}{(u_k - c_k) + (c_k - \ell_k)\exp(A_k\mathbf{h}^T\boldsymbol{\delta}_k)}. \tag{4}$$

This version is centered at 1 (all $c_k$ are 1) with all $A_k = B_k$. With some work, one can see that the right-hand side of Equation (4) is nearly equal to $1 + \mathbf{h}^T\boldsymbol{\delta}_k$ when $\mathbf{h}^T\boldsymbol{\delta}_k$ is small, which it should be assuming we have already appropriately adjusted for

nonresponse (and there are no frame coverage issues). By setting $\ell_k = 1/a_k$, no weight can be less than 1. Finally, letting $\boldsymbol{\delta}_k = [a_k - 1]\mathbf{z}_k$ will tend to produce more efficient estimates than the conventional setting $\boldsymbol{\delta}_k = \mathbf{z}_k$.

If no restriction is put on the upper size of the weight adjustment in Equation (3), that is, if all $u_k = \infty$, then

$$w_k = 1 + (a_k - 1)\exp(B_k[a_k - 1]\mathbf{h}^T\mathbf{z}_k).$$

The third census region has only 32 respondents. Without restricting the $u_k$ some of those have relatively large $w_k q_k$ values. This suggested to us setting $u_k$ in this region to $.105Q/q_k$, where Q was the sum or the $q_j$ in the region. This restricts the size of $w_k q_k = a_k u_k q_k$ to 10.5% of $Q$. We chose 10.5% because a restriction to 10% was not possible without the calibration equations failing to hold.

## 3. Variance Estimation

Both the weight-adjustment functions, whether $a_k/d_k$ in Equation (1) or $w_k/a_k$ in Equation (3), are versions of Folsom and Singh's general exponential model:

$$f(\phi; u_k, c_k, \ell_k) = \frac{\ell_k(u_k - c_k) + u_k(c_k - \ell_k)\exp(A_k\phi)}{(u_k - c_k) + (c_k - \ell_k)\exp(A_k\phi)}$$

where $A_k = (u_k - \ell_k)/[(c_k - \ell_k)(u_k - c_k)]$. For variance estimation under a correctly specified response model, one needs the derivative of $f(.)$ with respect to $\phi$, which is

$$f'(\phi; u_k, c_k, \ell_k) = \frac{(u_k - f_{1k})(f_{1k} - \ell_k)}{(u_k - c_k)(c_k - \ell_k)} \tag{5}$$

where $f_{1k} = f(\phi; u_k, c_k, \ell_k)$.

### *3.1. One Calibration-Weighting Step*

If we only calibrated for nonresponse, a good estimator for the variance of $t_{y,a} = \Sigma_R a_k y_k$, assuming the response model is correctly specified, would be

$$v(t_{y,a}) = \sum_{h=1}^{H}\sum_{k\in S_h}\left(1 - \frac{n_h}{N_h}\right)\left(\frac{n_h}{n_h - 1}\right) \times \left[\left(\theta d_k \mathbf{x}_k^T\mathbf{b}_1 + a_k e_{1k}\right) - \frac{1}{n_h}\sum_{j\in S_h}\left(\theta d_j \mathbf{x}_j^T\mathbf{b}_1 + a_j e_{1j}\right)\right]^2 + \sum_{k\in R} d_k\left(f_{1k}^2 - f_{1k}\right)e_{1k}^2, \tag{6}$$

where $a_k = 0$ when hospital $k$ is not in the set of responding hospitals $R$, $S_h$ denotes a stratum ($h = 1, \ldots, H$) containing $n_h$ sampled hospitals and $N_h$ total hospitals, $n$ is the total number of sampled hospitals (in our case, 367), $f(\mathbf{g}^T\mathbf{x}_k; \infty,2,1) = f_{1k} = a_k/d_k = 1 + \exp(\mathbf{g}^T\mathbf{x}_k)$ is the weight-adjustment factor, $f'(\mathbf{g}^T\mathbf{x}_k; \infty,2,1) = \exp(\mathbf{g}^T\mathbf{x}_k)$,

$$\begin{aligned}\mathbf{b}_1 &= \left[\sum_R d_k f'(\mathbf{g}^T\mathbf{x}_k;\infty,2,1)\mathbf{x}_k\mathbf{x}_k^T\right]^{-1}\sum_R d_k f'(\mathbf{g}^T\mathbf{x}_k;\infty,2,1)\mathbf{x}_k y_k \\ &= \left[\sum_R d_k \exp(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k\mathbf{x}_k^T\right]^{-1}\sum_R d_k \exp(\mathbf{g}^T\mathbf{x}_k)\mathbf{x}_k y_k, \\ e_{1k} &= y_k - \mathbf{x}_k^T\mathbf{b}_1, \text{ and } \theta = 1.\end{aligned} \tag{7}$$

Table 1 displays the sample and respondent sizes for our 2010 DAWN data within strata. The certainty strata from across the metropolitan areas have been combined.

See, for example, Kott and Liao (2012) for a fuller explanation of why Equation (6) provides a nearly unbiased estimator for the variance of $t_{y,a}$ when unit response is a logistic function of $\mathbf{x}_k$. The argument there parallels an earlier one in Kott (2006) where instead of the respondent sample being calibrated to the full sample as in Equation (2), the respondent (or full) sample was calibrated to the population using $\sum_R a_j\mathbf{x}_j = \sum_U \mathbf{x}_j$. Equation (6) was proposed in Kott (2006) with $\theta = 0$. The article shows that by injecting $f'(\mathbf{g}^T\mathbf{x}_k;\infty,2,1)$ into $\mathbf{b}_1$, one is able to avoid accounting for the $p_k$ only being estimates of the hospital response probabilities.

Were a simple random sample drawn *with* replacement within the $H$ strata or if the sampling fraction $(n_h/N_h)$ in each stratum were small enough to ignore, a good variance estimator would be

$$v_{WR}(t_{y,a}) = \sum_{h=1}^{H}\sum_{k\in S_h}\left(\frac{n_h}{n_h-1}\right)\left[\left(\theta d_k\mathbf{x}_k^T\mathbf{b}_1 + a_k e_{1k}\right) - \frac{1}{n_h}\sum_{j\in S_h}\left(\theta d_j\mathbf{x}_j^T\mathbf{b}_1 + a_j e_{1j}\right)\right]^2 \tag{8}$$

The added variance due to nonresponse is contained within what looks like a naïve single-phase variance estimator in Equation (8). The added variability due to the response/nonresponse phase comes from the $a_k = d_k f_{1k} I_k = d_k I_k/p_k$, where $I_k$ is the response indicator for hospital $k$, and $p_k$ remains the hospital's implicitly estimated probability of response. Since the $I_k$ are independent across hospitals, the naïve single-phase variance estimator fully captures the added variance due to nonresponse (for which $\sum_R d_k^2\left(f_{1k}^2 - f_{1k}\right)e_{1k}^2$ would be a good estimator).

### 3.2. *Two Calibration-Weighting Steps*

Kott and Liao (2012) also provide a nearly unbiased variance estimator for $t_{y,w} = \Sigma_R\, w_k y_k$ when unit response is a logistic function of $\mathbf{x}_k$:

$$\begin{aligned}v(t_{y,a}) = &\sum_{h=1}^{H}\sum_{k\in S_h}\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h}{n_h-1}\right) \\ &\times\left[\left(d_k\mathbf{x}_k^T\tilde{\mathbf{b}}_1 + a_k f_{2k}\tilde{e}_{1k}\right) - \frac{1}{n_h}\sum_{j\in S_h}\left(d_j\mathbf{x}_j^T\tilde{\mathbf{b}}_1 + a_j f_{2j}\tilde{e}_{1j}\right)\right]^2 \\ &+\sum_{k\in R} d_k\left([f_{1k}f_{2k}\tilde{e}_{1k}]^2 - f_{1k}f_{2k}\tilde{e}_{1k}^2\right),\end{aligned} \tag{9}$$

where

$$\tilde{\mathbf{b}}_1 = \left[\sum_R d_k \exp\left(\mathbf{g}^T\mathbf{x}_k\right)\mathbf{x}_k\mathbf{x}_k^T\right]^{-1} \sum_R d_k \exp\left(\mathbf{g}^T\mathbf{x}_k\right)\mathbf{x}_k f_{2k} e_{2k},$$

$$f_{2k} = f([a_k - 1]\mathbf{h}^T\mathbf{z}_k; u_k, 1, 1/a_k),$$

$$e_{2k} = y_k - \mathbf{z}_k^T\mathbf{b}_2,$$

$$\tilde{e}_{1k} = e_{2k} - \mathbf{x}_k^T\tilde{\mathbf{b}}_1,$$

$$\mathbf{b}_2 = \left(\sum_R a_j f'([a_j - 1]\mathbf{h}^T\mathbf{z}_j; u_j, 1, 1/a_j)[a_j - 1]\mathbf{z}_j\mathbf{z}_j^T\right)^{-1}$$

$$\sum_R a_j f'([a_j - 1]\mathbf{h}^T\mathbf{z}_j; u_j, 1, 1/a_j)[a_j - 1]\mathbf{z}_j y_j,$$

and $f'(.)$ is defined using Equation (5). To a large extent, Equation (9) is Equation (6) but with $y_k$ replaced by $f_{2k}\, e_{2k}$ causing $\tilde{\mathbf{b}}_1$ and $\tilde{e}_{1k}$ to replace $\mathbf{b}_1$ and $e_{1k}$. Recall that $f_{2k}$ is very close to 1 under the assumption that we modeled the nonresponse correctly.

Observe that if $\tilde{\mathbf{b}}_1 = \mathbf{0}$, we would have the simplified expression:

$$v(t_{y,a;S}) = \sum_{h=1}^{H}\sum_{k\in S_h}\left(1 - \frac{n_h}{N_h}\right)\left(\frac{n_h}{n_h - 1}\right)\left[a_k f_{2k} e_{2k} - \frac{1}{n_h}\sum_{j\in S_h} a_j f_{2j} e_{2j}\right]^2$$

$$+ \sum_{k\in R} a_k\left(f_{1k}[f_{2k}e_{2k}]^2 - f_{2k}e_{2k}^2\right).$$

This is almost the variance estimator one would get by ignoring the first calibration step and pretending the $a_k$ were the design weights:

$$v(t_{y,a;S'}) = \sum_{h=1}^{H}\sum_{k\in S_h}\left(1 - \frac{n_h}{N_h}\right)\left(\frac{n_h}{n_h - 1}\right)$$

$$\times\left[a_k f_{2k} e_{2k} - \frac{1}{n_h}\sum_{j\in S_h} a_j f_{2j} e_{2j}\right]^2 + \sum_{k\in R} a_k\left([f_{2k}e_{2k}]^2 - f_{2k}e_{2k}^2\right).$$

The difference is the $f_{1k}$, which appears in $v(t_{y,a;S})$ but not in $v(t_{y,a;S'})$ and makes the former larger than the latter except when all the sampling fractions are ignorably small or there is no nonresponse.

Now suppose instead we assume a linear prediction model consistent with treating $\mathbf{b}_1$ as $\mathbf{0}$. In particular,

$$y_k|\mathbf{z}_k, \mathbf{x}_k = \mathbf{z}_k^T\boldsymbol{\beta}_2 + \varepsilon_{2k}, \tag{10}$$

where the $\varepsilon_{2k}$ was uncorrelated random variables each with a mean of zero and a variance of $\kappa q_k$ for some unknown $k$, whether or not the hospital was sampled or responded when sampled.

It is not hard to see that the model variance of $t_{y,w}$ as an estimator for $\Sigma_U y_k$ given the respondent sample is $\sum_R (w_k^2 - w_k)\kappa q_k$. Similarly, the variance estimator in Equation (6) will have nearly the same prediction-model expectation if the $N_h$ is replaced by

$$N_h^* = n_h \frac{\sum_{R_h} a_k{}^2 f_{2k}{}^2 q_k}{\sum_{R_h} a_k f_{2k}{}^2 q_k} \tag{11}$$

when the respondent sample in stratum $h$ is not empty (otherwise, set $N_h^*$ to, say, 1000). Since the variance estimator is nearly unbiased given any respondent sample, it is also nearly unbiased on average across all respondent samples, that is, under the combination of the assumed response and prediction models and the original sampling mechanism. Note that when all the stratum sample fractions are ignorably small, this variance estimator coincides with $v(t_{y,a;S})$ (but not generally otherwise).

## 4. An Application

In this section, we compare variance estimators computed after:

1. Calibrating only for nonresponse pretending the sample was drawn with replacement;
2. Calibrating only for nonresponse;
3. Calibrating for both nonresponse and sample balance but pretending the sample was drawn with replacement;
4. Calibrating for both nonresponse and sample balance;
5. Calibrating for both nonresponse and sample balance but pretending the sample was drawn with replacement and using the simplified version of variance estimation described in the subsection 3.2;
6. Calibrating for both nonresponse and sample balance using the simplified version of variance estimation described in the subsection 3.2.

Since the estimated totals are different when we only calibrate for nonresponse, we compare estimated coefficients of variation (*cv*s) rather than estimated variances. Henceforth, we will abbreviate an estimated coefficient of variation as cv. The fourth variance-estimation method above produced nearly unbiased estimates of the variances for the following six estimated totals we investigated at the US and census-region levels:

*all drug-related hospital visits,*

*alcohol-related visits,*

*illicit-drug-but-not-alcohol-related visits,*

*psychotherapeutics-related visits,*

*stimulant-related visits, and*

*drug-related visits ending in death*

computed within each census region and across the four regions.

We computed some variance estimates pretending the sample was drawn with replacement since that is how many variances are estimated in practice, either because

*Table 1. Population, sample, and respondent sizes in subpopulation 1 (13 "metro" areas)*

| Stratum | Population Size | Sample Size | Respondent Size | $N_h$* (Equation (11)) |
|---|---|---|---|---|
| Certainties | 254 | 254 | 123 | 683.74 |
| Probability Strata | | | | |
| **East** | | | | |
| *Metro Area 1* | | | | |
| Stratum 1 | 10 | 8 | 6 | 14.34 |
| Stratum 2 | 10 | 7 | 4 | 15.26 |
| Stratum 3 | 10 | 3 | 2 | 13.51 |
| *Metro Area 2* | | | | |
| Stratum 1 | 4 | 2 | 1 | 3.96 |
| Stratum 2 | 8 | 6 | 5 | 11.78 |
| Stratum 3 | 14 | 9 | 4 | 22.33 |
| **South** | | | | |
| *Metro Area 3* | | | | |
| Stratum 1 | 6 | 5 | 4 | 5.65 |
| Stratum 2 | 44 | 28 | 15 | 108.92 |
| *Metro Area 4* | | | | |
| Stratum 1 | 18 | 5 | 4 | 22.67 |
| **Midwest** | | | | |
| *Metro Area 6* | | | | |
| Stratum 1 | 6 | 3 | 1 | 43.34 |
| Stratum 2 | 7 | 3 | 1 | 31.02 |
| *Metro Area 7* | | | | |
| Stratum 1 | 5 | 3 | 0 | 1000.00 |
| Stratum 2 | 7 | 3 | 2 | 15.67 |
| *Metro Area 8* | | | | |
| Stratum 1 | 19 | 4 | 2 | 74.72 |
| **West** | | | | |
| *Metro Area 9* | | | | |
| Stratum 1 | 6 | 5 | 3 | 8.97 |
| *Metro Area 10* | | | | |
| Stratum 1 | 10 | 9 | 3 | 19.49 |
| *Metro Area 11* | | | | |
| Stratum 1 | 4 | 3 | 0 | 1000.00 |
| *Metro Area 13* | | | | |
| Stratum 1 | 4 | 3 | 3 | 4.09 |
| Stratum 2 | 5 | 4 | 4 | 5.21 |
| Total | 451 | 367 | 187 | |

Metro Areas 5 and 12 have no probability strata (all certainties)

stratum sampling fractions are very small, as they are *not* here, or because the assumption makes variance estimation both easy and conservative. It also lets us see what damage, if any, resulted from our prediction-model-based treatment of finite-population correction.

Both pretending samples were drawn with replacement (WR) and treating them as drawn without replacement (WOR), the relative increase in the *cv*'s from only calibrating for nonresponse are displayed in the first two columns of Table 2. We looked at relative differences in the *cv*s because the different weights from using one or two calibration-

weighting steps lead to different estimated totals. We measured relative differences by taking the log of the ratio of the *cv*s being compared (e.g., $\log(cv_A/cv_B)$) because that measure is symmetric.

It is easy to see there is considerable *cv* reduction in most, but not all, cases from the sample balancing in the second calibration-weighting step. The *cv* of the estimates of the

*Table 2. Relative increase in estimated coefficients of variation (cv) due to adjusting only for nonresponse or using the simplified variance estimator*

| | Adjusting only for nonresponse | | Simplified variance estimator | |
|---|---|---|---|---|
| Estimator | WR $\log(cv_1/cv_3)$ | WOR $\log(cv_2/cv_4)$ | WR $\log(cv_5/cv_3)$ | WOR $\log(cv_6/cv_4)$ |
| **All regions** | | | | |
| Drug-related visits | 48.73 | 45.21 | 1.27 | 7.04 |
| Alcohol-related visits | 22.26 | 17.89 | 0.59 | 4.96 |
| Illicit-drug-related visits | 19.82 | 13.77 | 0.78 | 7.30 |
| Psychotherapeutics-related visits | 21.57 | 16.57 | 0.70 | 6.45 |
| Stimulant-related visits | 38.49 | 34.92 | 2.31 | 8.90 |
| Resulted in death | 4.93 | −8.81 | −0.13 | 7.60 |
| **East** | | | | |
| Drug-related visits | 76.78 | 83.14 | 2.57 | 8.72 |
| Alcohol-related visits | 37.12 | 35.96 | 1.41 | 5.25 |
| Illicit-drug-related visits | 48.44 | 47.85 | 1.50 | 10.31 |
| Psychotherapeutics-related visits | 44.71 | 48.04 | 3.34 | 5.90 |
| Stimulant-related visits | 50.88 | 56.26 | 3.05 | 10.06 |
| Resulted in death | 11.79 | 17.21 | 0.24 | 2.33 |
| **South** | | | | |
| Drug-related visits | 82.93 | 87.53 | 0.31 | 2.46 |
| Alcohol-related visits | 26.79 | 26.00 | 0.61 | 2.76 |
| Illicit-drug-related visits | 18.46 | 16.29 | 0.78 | 1.61 |
| Psychotherapeutics-related visits | 61.01 | 62.57 | −0.05 | 1.33 |
| Stimulant-related visits | 78.39 | 83.29 | 0.56 | 2.97 |
| Resulted in death | 23.05 | 21.03 | 0.46 | 0.98 |
| **Midwest** | | | | |
| Drug-related visits | 118.44 | 102.45 | −0.58 | 21.37 |
| Alcohol-related visits | 106.02 | 76.14 | −0.79 | 19.74 |
| Illicit-drug-related visits | 96.51 | 70.14 | 1.28 | 24.55 |
| Psychotherapeutics-related visits | 44.18 | 29.80 | −0.31 | 17.53 |
| Stimulant-related visits | 98.91 | 84.06 | −0.55 | 20.09 |
| Resulted in death | −14.25 | −16.70 | 0.54 | 15.32 |
| **West** | | | | |
| Drug-related visits | 66.50 | 49.16 | 0.44 | 0.02 |
| Alcohol-related visits | 49.07 | 37.72 | 0.24 | 10.91 |
| Illicit-drug-related visits | 52.15 | 45.74 | −0.05 | 22.66 |
| Psychotherapeutics-related visits | 47.78 | 36.27 | 0.41 | 0.17 |
| Stimulant-related visits | 56.66 | 43.43 | 0.62 | −0.06 |
| Resulted in death | 9.52 | 6.16 | −0.43 | −2.77 |
| Mean | 47.93 | 42.30 | 0.70 | 8.22 |
| Min | −14.25 | −16.70 | −0.79 | −2.77 |
| Max | 118.44 | 102.45 | 3.34 | 24.55 |

number of deaths from drug-related visits both across the US and in the Midwest are larger after sample balancing. All other *cv*s are smaller, over 40% smaller on average.

Columns 3 and 4 show that using the simplified variance estimator described in the last subsection (Equation (6) with the $N_h$ replaced by the $N_h$* in Table 1) increases the *cv*s more often than not. When it is not conservative, the simplified method is never more than 3% lower than its nearly unbiased counterpart in the 30 *cv*s we computed. The results tend to be more conservative and much more variable when the without-replacement version of the variance estimator is used, and we employ Equation (11) to counteract what would otherwise be an over-correction for the large sampling fractions in most strata. Replacing $q_k$ in Equation (11) by $q_k^2$ would make the simplified *cv*s a bit less conservative (not shown). The average upward bias would drop to 4.67%, with a minimum of −7.01% and a maximum of 21.13%.

## 5. Some Concluding Remarks

We have shown how to produce calibration weights for the 2010 DAWN respondent sample of hospitals in two steps – the first to remove the bias from unit nonresponse assuming that we modeled response correctly as a logistic function of covariates, and the second to provide sample balance and thereby increase the statistical efficiency of most estimated totals. We have also shown how to compute nearly unbiased measures of the standard errors of DAWN-estimated totals, providing a simplified version that, although not nearly unbiased, appears to be mostly conservative and is easily computed using SUDAAN 11.

The reason why the simplified version tends to be conservative is that it replaces a respondent-sample derived estimate for a parameter ($\mathbf{b}_l$) by $\mathbf{0}$. To the extent that there are efficiency gains to be made from the nonresponse calibration-weighting step *in addition* to those made in the sample-balancing step − and there may not be any (we are effectively regressing a residual, $e_{2k}$ on $\mathbf{x}_k$, in the nonresponse-adjustment step) − this simplification will tend to underestimate the true standard error of the two-step calibration.

Since we were able to compute a nearly unbiased measure of the standard errors of two-step-calibrated estimates, an obvious question is why bother introducing a simplified version of the computation? The obvious reason is that statisticians will not be able to mimic what we have done for variance estimation without great effort. Moreover, this effort grows for estimated ratios, like the fraction of drug-related hospital visits involving alcohol.

Some may wonder why we did not perform the calibration-weighting steps in the reverse order: sample balancing first, followed by nonresponse adjustment. That clearly could be done, but we will not follow up on it here. Something to consider before reversing the calibration steps, however, is that upper bounds on the final weights cannot be set in the nonresponse-adjustment step unless one is willing to change the form of the response model being fit. This runs the risk of introducing nonresponse bias. No such risk exists when setting upper bounds in the sample-balancing step.

## 6. References

Folsom, R.E. 1991. "Exponential and Logistic Weight Adjustments for Sampling and Nonresponse Error Reduction." In *Proceedings of the American Statistical Association, Social Statistics Section*, 197–202.

Folsom, R.E. and A.C. Singh. 2000. "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification." In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 598–603. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/2000_099.pdf (accessed July 1, 2014).

Kim, J.K. and M. Riddles. 2012. "Some Theory for Propensity Scoring Adjustment Estimator." *Survey Methodology* 38: 157–165.

Kott, P.S. 2011. "A Nearly Pseudo-optimal Method for Keeping Calibration Weights from Falling Below Unity in the Absence of Nonresponse or Frame Errors." *Pakistan Journal of Statistics* 27: 391–396.

Kott, P.S. 2006. "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors." *Survey Methodology* 32: 133–142.

Kott, P.S. and D. Liao. 2012. "One Step or Two? Calibration Weighting from a Complete List Frame with Nonresponse." Under review by *Survey Methodology* (presented at the Symposium on the Analysis of Survey Data and Small Area Estimation, in honour of the 75th Birthday of Professor J. N. K. Rao).

RTI International 2012. *SUDAAN Language Manual, Release 11.0*. Research Triangle Park, NC: RTI International.

Substance Abuse and Mental Health Services Administration 2012. *Drug Abuse Warning Network (DAWN)*. Available at: http://www.samhsa.gov/data/DAWN.aspx (accessed July 1, 2014).

US Census Bureau 2012. *ZIP Code™ Tabulation Areas (ZCTAs™)*. Available at: https://www.census.gov/geo/reference/zctas.html (accessed July 1, 2014).

Received November 2012
Revised February 2014
Accepted May 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 533–545, http://dx.doi.org/10.2478/JOS-2014-0033

# Access to Sensitive Data: Satisfying Objectives Rather than Constraints

*Felix Ritchie*[1]

The argument for access to sensitive unit-level data produced within government is usually framed in terms of risk and the legal responsibility to maintain confidentiality. This article argues that the framing of the question may restrict the set of possibilities; a more effective perspective starts from the data owner's principles and user needs. Within this principles-based framework, the role of law changes: It becomes an 'enabling technology', helping to define the solution but playing no role in setting the objectives.

This shift in perspective has a number of consequences. The perception of 'costs' and 'benefits' is reversed. Law and established practice are distinguished and appropriately placed within a cost-benefit framework. The subjectivity and uncertainty in risk assessments is made explicit. Overall, all other things being equal, the expectation is that a move towards objective-based planning increases data access and improves risk assessment.

This alternative perspective also addresses the problem of the public-good nature of research outputs. It encourages the data owner to engage with users and build a case for data access taking account of the wider needs of society.

The UK data access regime is used as the primary example of the arguments in this article.

*Key words:* Confidential data; data access; data security; public goods; risk.

## 1. Introduction

It is nowadays widely accepted that access to confidential or sensitive microdata collected by government is essential for the research needed to produce an evidence base for policy; see Trewin et al. (2007) for a discussion. This data is usually collected either by statistical agencies to produce aggregates, or by government departments as part of their work. In both cases, use of the underlying microdata directly allows the collecting body to leverage their investment in data collection at minimal additional cost.

General agreement on the principle of research access is common; but principles can take a back seat when implementation is considered. In particular, the confidentiality of the data becomes paramount, and access to data focuses on how that confidentiality can be maintained. However, there is ample evidence to suggest that government is likely to be

[1] Bristol Business School, University of the West of England, Frenchay Campus Bristol BS16 1QY, Bristol, UK. Email: felix.ritchie@uwe.ac.uk

**Acknowledgments:** This article is developed from a presentation for the Statistics New Zealand Official Statistics Forum in March 2010. I am grateful to SNZ and Motu for funding my visit and giving me the opportunity to draw out some of the themes here. The germ of this article arose from discussions with Richard Welpton of the Secure Data Service. I am also grateful to Tanvi Desai for detailed comments on an earlier draft, and to the referees and editors for incisive comments, particularly in relation to the appropriate role of risk.

collectively and individually risk-averse (see, for example, OAG 1998; House of Lords 2006; Pfeifer 2008; Buurman et al. 2012; Hall 2013) and so decisions taken may not be socially optimal.

This article argues that changing the perspective to concentrate on the principles governing data access can help to improve the quality of decisions taken, as well as clarifying exactly what risks are being run and what the benefits are. The basis of this argument, well-attested in psychology, is that the framing of the question affects the answers that are generated.

The next section proposes a perspective on access which emphasises the predominance of objectives over constraints. This leads to a model where law and technology are 'enablers': That is, they inform, and may constrain, decisions to be taken on how to implement an objective, but do not define the objective itself. The following two sections discuss these in more detail, and Section 5 examines how this facilitates an understanding of the role of risk.

Section 6 considers how this change of perspective can inform the debate on the 'public goods' problem of data access identified by Ritchie and Welpton (2012). The article concludes by noting that the arguments advanced here run counter to the natural decision-making structures in government, and so an efficient system of confidential data access may need an active and engaged sponsor.

For simplicity, the article throughout refers to the options of National Statistics Institutes (NSIs), who are generally the main or only holders of confidential government research data. However, it should be clear that the arguments apply to any owner of confidential data considering giving access to that data for research.

The author has been involved in data access in the UK for over a decade, has formally and informally advised the OECD and Eurostat, and has worked on data confidentiality with NSIs in many different countries. The examples used in this study are mostly drawn from the author's experience in the UK, as it is difficult to ascertain whether an individual's perspective truly reflects the experience of an organisation or country without having worked there. However, I am confident that the characterisation of NSI behaviour in this article, while simplified, is a fair reflection.

## 2. The Framework Principles

### 2.1. *Constrained Decision Making: The 'Constraint Model'*

The usual decision-making process for giving access to confidential data can be framed as in Figure 1, which we will refer to as the 'constraint model'.

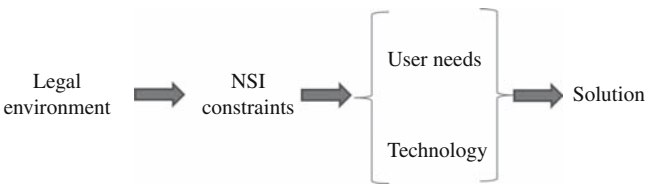



*Fig. 1. The 'Constraint Model'*

That is, organisations ask:

- What does the law say we can do?
- Given that, what do we need to ensure?
- What technologies are available to satisfy those constraints, and what are the needs of the user that can be satisfied within those constraints?
- How do we employ technology to meet the identified user needs in the best way?

The problem with the Constraint Model is the first step. Clearly, acting within the law is a requirement of any agency. The problem is that 'the law' is rarely a simple, unambiguous construct with only one possible outcome; a statement of practical law is an interpretation in relation to a specific set of circumstances. However, focusing on a particular interpretation constrains the set of solutions to a subset of outcomes, particularly if the circumstances surrounding the interpretation are not explicitly made known.

Consider the UK experience in 2003. The Office for National Statistics (ONS), the UK's NSI, was reviewing the options for giving academic researchers access to confidential business data. The prevailing legal opinion was that this was not possible: The Act governing such access strictly limited access to employees of the UK government.

This seems crystal clear, until it is considered that the question being asked is the implicit one, "can academics, *in their own right*, have access to business microdata?" This is a very specific and, as it turned out, very limiting question. An alternative question was put to the government legal advisors: "Can academics become Civil Servants for the purposes and duration of their research?" There were several positive responses to this question; a form of secondment was taken as the most workable. As a result of changing the perspective, an outcome, previously considered impossible, was achieved with a solution in keeping with both the spirit and letter of the law. For details, see Ritchie (2009).

The specific legal arrangements continued to evolve as different circumstances came to light. ONS's Legal Services unit periodically reviewed the secondment arrangements, and the team providing access was required from time to time to amend its procedures to address potential areas for challenge. For example, the team was asked to demonstrate that access was through 'fair and open competition', and to specify the criteria for determining whether access contributed to ONS's 'benefit'.

The important lesson from this is that the attitude of the NSI determined the outcome; that is, whether access could be granted or not. Both the research team and the legal team shared the same aim: to see wider research use being made of confidential data, lawfully. The specifics of implementation were just that: specifics of implementation, not a universal statement of law.

### *2.2. Principles-Led Decision Making: The 'Objective Model'*

This focus on objectives rather than implementation leads to a rather different framework for access, as displayed in Figure 2:

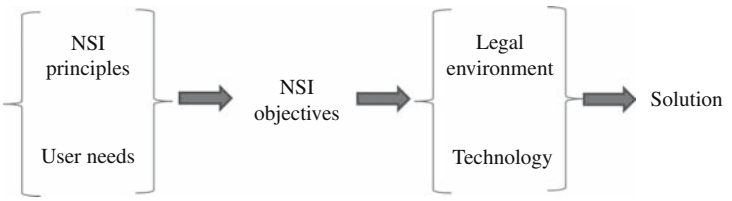



*Fig. 2. The 'Objective Model'*

The questions now are, in order,

- What are our operating principles, and what do users want?
- What how do we turn this into a set of objectives?
- What legal and technological options are available?
- How do we employ these alternatives to meet the objectives in the best way?

This 'Objective Model' puts the aims of the NSI and the user at the start of the decision process. Law has the same status as technology: Just as all implementations are limited by the existing technology, so they are constrained by the existing law. But technology and law are both used to *achieve* the objectives; they do not count towards the *definition* of those objectives.

One of the implications of the Objective Model is that multiple legal and technological solutions may meet those objectives; there is no need to identify 'the' legal or technical solution. Several solutions might coexist; the aim is to find the combination of solutions that meets the objectives best.

A second implication is the primacy of the 'user need' (with the NSI itself as one class of user). User demands can be stereotyped as "give me all the data now, on my desktop, with no restrictions", but this is an exaggeration. Researchers are generally aware that not all tasks need all data, particularly as more detail typically involves more restrictions. As an example, the UK Data Archive provides many datasets in both anonymised and detailed form, with the latter having more access restrictions. A bona fide UK researcher would have little trouble getting access to either, but the usage of the anonymised files massively outstrips that of the restricted-access detailed files. This model, of some sort of data archive holding files for distribution balanced by more restricted access to more detailed data, is relatively common across countries, indicating that users can make balanced judgements about costs and benefits.

The dichotomy characterised by the Constraint/Objective Models may be unfair to individual NSIs, but in the author's experience, based on work with numerous NSIs and international organisations, this state generally prevails in the real world. There are units within NSIs that consider user needs and then consider how to meet them; but the majority still seek to identify the legal framework and then assess which user needs can be accommodated within that framework. An even smaller minority are prepared to consider NSI objectives and user needs jointly without reference to legal limits on implementation.

### *2.3. Semantics or Substance?*

It could be argued that this is a largely semantic argument; that is, the real questions are always about implementation, and the same solution could be derived by individuals

working from the different models. A rational organisation with all the necessary information would always come to the same conclusion, whatever its conceptual stance.

An analogy is with constrained optimisation. Take the typical undergraduate economics problem of maximising utility subject to a budget constraint, resulting in an optimal utility of, say, U*. It can be demonstrated that minimising the expenditure needed to achieve that given level of utility U* recreates the budget constraint from the first problem, assuming the constraints are binding. Hence these are referred to as the 'primal' and the 'dual', with each generating 'shadow prices' for the cost of the constraint (see e.g., Varian 1992).

This focus on the equivalence of solution hides an important outcome. In the maximisation problem, the shadow prices are the benefit to be gained by loosening the budget constraint. In the minimisation problem, the shadow prices reflect the cost of any further increases in utility. These are clearly two different concepts, and the way the problem is posed reflects the analysts' interest.

Similarly, the Constraint and Objective Models imply fundamentally different mindsets: the difference between "what can we do?" and "what would we like to do?" In coming to a solution, the perception of what has been given up to achieve that outcome differs, even if the outcome is the same.

However, this is not simply an alternative perspective. In the mathematical problem the choice of maximisation or minimisation does not affect the problem parameters, only the interpretation of results; but in the human world, the outcome can be substantially affected by the way that the question is framed.

Since the 1970s the psychological literature has repeatedly demonstrated the importance of framing effects; see, for example, Kahneman (2012, especially chap. 34) for an overview, Mellers et al. (1999) for experimental evidence, or De Martino et al. (2006) for a discussion of the psychological basis. This is also recognised in environmental and behavioural economics, politics, and marketing, for example – all subjects where the focus is on understanding what influences the decisions of people.

The relevance of this to the Constraint/Objective Model discussion is that the initial framing of the problem will lead to either 'losses' or 'gains' being identified. Losses tend to be felt more keenly than gains, and the certainty of outcomes affects decisions. As result, there is a tendency, all other things being equal, to stick to the starting point; see Kahneman et al. (1991) for examples. Samuelson and Zeckhauser (1988) identify this as 'status quo bias'.

Consider Figure 3, below.

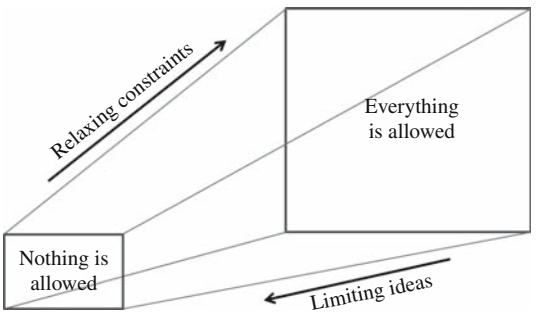



*Fig. 3. The importance of the starting point*

The Constraint Model could be considered as starting from 'nothing is allowed'; as potential candidates for solutions are evaluated for conformity with the legal environment, constraints can be relaxed. The 'Objective Model' is starting from 'everything is allowed', and withdrawing from that position as solutions are shown to be unlawful or impractical. The theories of framing suggest that, while two NSIs starting from different perspectives could come to the same conclusion, the more likely outcome is that they will differ in their implementation. All other things being equal, the Objective Model is more likely to lead to more data access. Hence, this is not a semantic discussion: The conceptual stance of the organisation affects the outcome.

## 3. Law as an Enabler

Once law is seen as a tool to inform decisions about implementation, rather than a governing framework, some useful results appear.

First, attention focuses on the purposes of legal advice. Lawyers are professionally cautious: That is, one of their duties is to ensure that clients are warned about liabilities and consequences. Advice is likely to focus on avoiding negative outcomes. In the Constraint Model, legal advice sets the ground rules for all subsequent decisions. This places an inappropriate burden on lawyers, who are unlikely to be experts in data access. In the Objective Model, legal advice is taken in the context of specific solutions. Lawyers are not being asked to speculate on potential future interpretations, and any advice is reviewed in the context of the objectives. Both the giving and receiving of advice is more effective.

Second, changes to the law can be evaluated more easily. If the legal environment changes, the constraints on the NSI change; the set of feasible delivery options changes. Under the Constraint Model, the 'value' of the changed law is whether the outcomes being produced are now better for society. In the Objective Model, NSI objectives are invariant to law; therefore, a test of the likely effectiveness of any new law is simply whether it improves the way the NSI meets its objectives.

Again taking the case of the UK, in 2008 the Statistics and Registration Act came into force. This formally gave ONS a function of supporting research for the public benefit, and provided a simple universal legal gateway for access to ONS microdata. This greatly simplified the process through which researchers gained access to data, clarified the role of researchers' use of ONS data, and brought all ONS data under the same legal framework for research. ONS objectives were largely unaltered, and so the impact of the law was a straight efficiency gain.

Third, the difference between law and established practice can be clarified and challenged. In the context of the diagrams above, 'law' includes the NSI's procedures, which often go beyond the law into areas where the NSI feels it has an ethical or operational responsibility even if no legal responsibility exists. Fixed ways of working, particularly when in place for a long time, can also easily be confused with law. Even when procedures are explicitly recognised as NSI policy decisions, they can still be seen as immutable.

For example, at an OECD meeting on international data sharing, the author discussed country attitudes with a representative from a European NSI. The representative initially stated that such data sharing was not allowed in that country's law; after some minutes of discussion, it transpired that the true position was that it was legally possible but the NSI would not allow it. This is a small difference but a very significant one.

Under the Constraint Model, challenging established practice is hard. A new or changed objective needs to demonstrate that it fits into the current understanding of the legal environment, which may be partly defined by established practice. But this begs the question: Why should objectives need to justify their value by reference to specific implementations? Surely the implementation has to address the objective, not the other way around?

Under the Objective Model, established practice has to justify its existence using fair and proper criteria for how well it addresses an objective: cost, benefit, effectiveness, legality, impact of disruption against the alternative solutions. These are also more easily quantifiable: The impact of a change on access rules on IT expenditure, for example, can be readily identified. Under the Objective Model cost-effective practice is what matters; the value of 'established practice' is only reduced costs of learning or change.

## 4. Technology as an Enabler

Technology (meaning all the practical matters surrounding access to data, including cost decisions) as an enabler is relatively self-explanatory. The technological options can be broadly grouped into six types:

- *Anonymisation* of the data: This is used for public files, such as those on the web.
- *Licensing* of researchers, sometimes combined with a degree of anonymisation, is still the most common way for researchers to get access to microdata.
- *Secure 'research data centres'* (RDCs), laboratory facilities at the NSI or the researcher's base; for many countries, this is still the only way to get access to detailed data.
- *Remote access*, where 'virtual' RDCs allow users to manipulate data unhindered by geography; although the technologies are common, implementation varies greatly from restricted-site access only to direct access from the internet.
- *Remote job submission*, where users send statistical programmes to be run and get back results, are relatively uncommon, but a number of NSIs have been exploiting web technologies to develop friendly interfaces.
- *Synthetic data*, which has the same characteristics as the real data but has been imputed from statistical models; the resulting dataset is then intended to be safe for distribution.

Most countries employ a number of these options, and often these solutions are combined. For example, some US Census Bureau data is made available at restricted on-site RDCs, but synthetic equivalents are accessible through a virtual RDC.

NSIs tend to be risk-averse and avoid new solutions, but in most of these areas a prospective data manager can draw on a wealth of international experience in implementation. As Ritchie (2013) notes, for strategic planning purposes an NSI can assume that an 'off-the-shelf' solution is available to meet its objectives. The everything-is-possible answer does not help planning, so Ritchie (2009) reduces the solution set by employing the concept of the 'data access spectrum'. This suggests identifying a finite number of access options defined by class of user, and then developing appropriate legal or technological solutions based on NSI costs and the resulting risk profile. In the UK this model has been used both to classify existing operations and to justify the development of a third-party remote access system and an improved off-site RDC model.

## 5. Perceptions of Risk

The shift between perceptions is important for the evaluation of risk. Risk is often discussed as if it is measurable, and sometimes this is the case. For example, the large field of research on ensuring that datasets are anonymised to an 'appropriate' level quantifies risk as the probability of identification given intruder and protection scenarios; see Duncan et al. (2001) for a typical example.

However, the risk inherent in the data is only one element in a data access solution. The commonly-used 'VML Security Model' (also called the 'Five Safes' model) classifies risks into those arising from people, projects, the settings, the data and the outputs (see Ritchie 2013 for an expanded discussion). These risk elements interact, and most are not amenable to quantification; for example, how are the risks inherent in an NSI's procedures for approving projects to be objectively assessed? The problem is made harder because NSIs generally have a very good record of managing confidentiality; examples of misuse of NSI research data are few and far between, and so there is no historical guide to the probability of confidentiality breaches. Risk is therefore a subjective measure, in general, which means it will be affected by the framing discussed earlier.

We earlier characterised the Constraint Model perspective as starting from 'nothing is allowed' and then evaluating individual ways to increase access in the context of the legal framework. Any solution therefore increases risk compared to doing nothing. Solutions may be compared to each other for their riskiness, but the default is always to do nothing.

Under the Objective Model, there is no risk baseline. If the NSI has an objective to make data available to a class of users, the default is to hand the data over. All solutions then involve placing some restrictions on that default by, for example, anonymising the data or restricting access. The aim of this is to reduce the risk of a breach of confidentiality, but from an uncertain level. As result, the only comparison that can be made is a subjective comparison to an alternative of equally subjective measurement.

Standard methods do of course exist, such as risk-utility models for evaluation of dataset vulnerability, as well as technical tools like tau-Argus, all adding an element of objectivity. However, like all models, these are parameterised subjectively; see Skinner (2012) for a perspective on perceived versus actual objectivity in NSI decision making. Moreover, while there are guidelines for good practice for nondata factors such as access environments and researchers (Brandt et al. 2010), these are entirely subjective. The Objective Model forces this subjectiveness to be acknowledged.

As noted earlier, losses tend to be weighed more heavily than gains. In the Constraint Model, the losses in security are balanced by gains in data access; in the Objective Model, gains in security are being balanced against losses in access. All other things being equal, the Constraint Model is likely to deliver lower access and higher security, the Objective Model more access and less security.

The psychological literature provides an additional insight. Certainty, all other things being equal, tends to have a higher weight than uncertain outcomes when comparing positive outcomes (see, for example, Viscusi et al. 1987). Consider now the options for an NSI. Benefits and changing risk are uncertain and subjective. The only fixed point is zero risk, which will have more weight in deliberations than the uncertain benefits and risks. Therefore, if the starting point is 'nothing is allowed', the outcome is likely to be more

restrictive than starting from an open solution and progressively adding restrictions. In the 'everything is allowed' case where there is no clear default measure, risks and benefits are more likely to be equally weighted, if still subjective.

Figure 4 summarises this discussion:

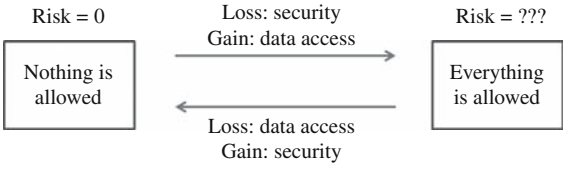



*Fig. 4. Measures and changes in risk*

All other things being equal, the Constraint Model, which is closer to the 'nothing is allowed' option, is likely to place more weight on the loss of security and on the do-nothing option. By contrast, the Objective Model, faced with a set of competing but subjective costs and benefits, is more likely to weight the two fairly; and because the 'losses' are in access and 'gains' are in security, it is more likely to favour access.

## 6. Objectives, Constraints, and the Public Goods Problem

This article has considered how a change in perceptions to the Objective Model may improve NSI outcomes and operations. A natural question is why the Constraint Model predominates in NSI thinking. Part of the answer lies in the communal nature of research output.

Ritchie and Welpton (2012) argue that one reason why NSIs tend to focus on protecting data rather than maximising value is a 'public goods' problem arising from the unequal distribution of risks and benefits. The benefit from making confidential data available for research largely accrues to the wider public, but the risk of being blamed for something going wrong is typically borne by the NSI. For example, if a licensed user is sent a confidential dataset and loses it, the NSI may well get blamed for distributing the data no matter how well founded its distribution policy is. In contrast, if data is not released, or is only used by the NSI for its own purposes, then the NSI minimises risk; but the wider public loses the benefit of that data and runs an increased risk of bad decision making.

In these circumstances it is rational for NSIs to take a cautious approach to data release, and consider their own priorities over the wider public benefit. The NSI's main function is to protect its interests: Risk avoidance becomes the goal, a conservative legal stance appeals, and the Constraint Model predominates. Even if the NSI takes the perspective of the Objective Model, it is still likely to underestimate benefits and overweight risks.

Ritchie and Welpton (2012) argue that one way to address the public-goods problem is to 'negotiate' the level of access with users; as part of that negotiation, issues of risk and responsibility are also addressed. Users are in a better position to identify the benefits from access; but then they need to acknowledge and accept joint responsibility for the risks being run by the NSI. For example, one of the key influences in establishing the ONS remote RDC was the explicit support from the UK Treasury, who made extensive use of the research outputs in their work.

This approach sits comfortably with the Objective Model, which puts agreement with users at the forefront of the decision-making process and views risk as something to be managed, not minimised. For the NSI to set objectives, it needs to consult with users – and this can be used to get the buy-in necessary to ensure a collective responsibility for the data release policy. If that buy-in is not forthcoming, the NSI is arguably justified in ignoring those user needs. Hence the Objective Model is consistent with the customer engagement necessary to avoid underprovision of data access from society's perspective.

Ritchie and Welpton (2012) propose an alternative: the use of third parties to provide the data access services. In many countries, data distribution has been outsourced for years to third party providers in the form of data archives. However, the bulk of confidentiality protection in these cases is vested in the data. More interesting are the recent moves towards allowing third parties to provide distributed-access services such as remote RDCs; for example, the NORC Data Enclave in the US, the UK Secure Data Service, the IAB RDC-in-RDC in Germany, or the DARA project developing a remote access system for Eurostat. In these, the risk dimensions of people, settings and outputs become much more important than data protection.

The advantage of third parties is that the transparency of contractual agreements forces both parties to identify and acknowledge the risks and the acceptable level of risk management to be employed. This model presents difficulties for the Constraint Model perspective because of the need to refer to the zero-risk baseline. In essence, the Constraint Model specifies inputs to third-party processes (i.e., limits on working), whereas the Objective Model emphasises outputs (targets to be achieved, irrespective of how they are achieved). This gives third-party providers more flexibility in delivering the required outcome; and of course it ties in with the requirement to identify user needs as an initial step.

## 7. Conclusion

This article has dichotomised the decision-making process for data access into the 'Constraint Model' and 'Objective Model'. Whilst this is clearly an oversimplification, it nevertheless usefully illustrates some different approaches to setting the objectives and solving problems associated with data access. In doing so, this idealised worldview also suggests that NSIs may be missing opportunities for both their benefit and the wider public.

This article has argued NSI decision-making processes tend to focus initially on what is allowed rather than what is desirable; the incentives for NSIs do not encourage exploration of the boundaries of their duties. This is not to argue that NSIs are deliberately acting against the public interest; as Buurman et al. (2012) demonstrate, risk aversion and 'public spirit' are two different concepts. Nevertheless, the tendency to risk aversion, however well intentioned, can mean that access to the data collected by NSIs and similar bodies is often unnecessarily restricted.

An alternative perspective focuses on the NSI objectives, and uses this to address questions of constraints in implementation, rather than the other way round. In this perspective, law, NSI procedures, and technology all become 'enablers': options for or constraints on implementation which affect the delivery of objectives, but not the objectives themselves.

This is not simply a semantic discussion; the subjectivity of decisions in this area means that the perspective of the NSI directly affects the outcomes achieved. In addition, basic human nature means that decisions about relative risk and uncertainty are affected by the starting point.

Focusing on objectives also provides a framework to bring users into the discussion on access principles, increasing the chance of community buy-in and reducing the NSI's incentives to implement an overly risk-averse release policy. The objective-based worldview opens up the NSI to wider and deeper engagement with users. Knowledgeable users who recognise the risks but can also express the benefits can help to reduce the public-goods problem associated with research data access.

There are signs that attitudes might be changing. At the 2013 meeting of the major biennial UN conference for government statisticians (http://www.unece.org/?id=31938), a session was held on "Moving from risk avoidance to risk management". The session papers described a number of positive developments in data access, using a variety of technologies. Of particular relevance here, while most papers focused on the idea of 'widening access' – that is, starting from a position of needing to justify any relaxation on data security – the Italian NSI (and, to a lesser extent, Eurostat and Mexico) took a strongly user-centred approach to work backwards from general objectives to specific implementations; see ISTAT (2013).

There is therefore a strong argument that NSIs could benefit from a change in perspective, and some shifts are happening. However, this user/objective-centric approach runs counter to the natural decision-making structures in government; these tend to be cautious, and reflect the Constraint Model. Ritchie (2013) notes that international sharing of confidential data has largely been driven by energetic individuals, rather than any corporate vision (the Eurostat DARA project http://www.safe-centre.info/ is an exception). An efficient system of confidential data access may therefore need active and engaged sponsors at a senior level to have any realistic prospect of success.

## 8. References

Brandt, M., L. Franconi, C. Guerke, A. Hundepool, M. Lucarelli, J. Mol, F. Ritchie, G. Seri, and R. Welpton. 2010. *Guidelines for the Checking of Output Based on Microdata Research*. Final report of ESSnet sub-group on output SDC, Eurostat. Available at: http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf (accessed 10th June 2014).

Buurman, M., J. Delfgaauw, R. Dur, and S. van den Bossche. 2012. *Public Sector Employees: Risk Averse and Altruistic?* CESifo Working Paper: Behavioural Economics, No. 3851. Available at: http://www.econstor.eu/handle/10419/61046 (accessed 10th June 2014).

De Martino, B., D. Kumuran, B. Seymour, and R. Dolan. 2006. "Frames, Biases, and Rational Decision-Making in the Human Brain." *Science* 313: 684–687. Available at: http://www.sciencemag.org/content/313/5787/684.full.pdf?sid=e7dcf2c8-5bbb-4d89-97f2-5344613de9bf (accessed 10th June 2014).

Duncan, G., S. Keller-McNulty, and L. Stokes. 2001. *Disclosure Risk vs Data Utility: the R-U Confidentiality Map*. NISS Technical Report no. 121. Available at: http://citeseerx.

ist.psu.edu/viewdoc/download;jsessionid=6BF9C4E902605252F4302A43786EF152?doi=10.1.1.79.1598&rep=rep1&type=pdf (accessed 10th June 2014).

Hall, K. 2013. "Can Government Change its Risk-Averse Take on Security?" *Computer Weekly*, February 7, 2013. Available at: http://www.computerweekly.com/news/2240177688/Can-government-change-its-risk-averse-take-on-security (accessed 10th June 2014).

House of Lords 2006. *Government Policy on the Management of Risk*. Select Committee on Economic Affairs, 5th Report of Session 2005–06, Available at: http://www.publications.parliament.uk/pa/ld200506/ldselect/ldeconaf/183/183i.pdf (accessed 10th June 2014).

ISTAT 2013. *Micro-data: A Crucial Asset for Statistical Systems*. UNECE/CES 61st Plenary Session, item 4(b). Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2013/31.pdf (accessed 10th June 2014).

Kahneman, D. 2012. *Thinking, fast and slow*. London: Penguin Books.

Kahneman, D., J. Knetsch, and R. Thaler. 1991. "Anomalies: the Endowment Effect, Loss Aversion and Status Quo Bias." *Journal of Economic Perspectives* 5:193–206. Available and reprinted at: http://www.jstor.org/stable/1942711 (accessed 10th June 2014).

Mellers, B., A. Schwartz, and I. Ritov. 1999. "Emotion-Based Choice." *Journal of Experimental Psychology: General* 128:332–345. Reprinted at http://www.researchgate.net/publication/215515670_Emotion-based_choice/file/79e4150b79f973939f.pdf (accessed 10th June 2014).

OAG 1998. *Innovation in the Federal Government: The Risk not Taken*. Public Policy Forum discussion paper, Office of the Auditor General of Canada. Available at: http://www.oag-bvg.gc.ca/internet/English/meth_gde_e_10193.html (accessed 10th June 2014).

Pfeifer, C. 2008. *Risk Aversion and Sorting into Public Sector Employment*. IZA Discussion Papers no. 3503. Available at: http://ftp.iza.org/dp4401.pdf (accessed 10th June 2014).

Ritchie, F. 2009. "UK Release Practices for Official Microdata." *Journal of the International Association of Official Statisticians.* 26(3/4): 103–111. DOI: http://dx.doi.org/10.3233/SJI-2009-0706.

Ritchie, F. 2013. "International Access to Restricted Data – a Principles-Based Standards Approach." *Statistical Journal of the International Association of Official Statisticians.* 29: 289–311. Reprinted at DOI: http://dx.doi.org/10.3233/SJI-130780.

Ritchie, F. and R. Welpton. 2012. "Data Access as a Public Good." In *Work session on statistical data confidentiality 2011,* UNECE/Eurostat. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/presentations/21_Ritchie-Welpton.pdf (accessed 10th June 2014).

Samuelson, W. and R. Zeckhauser. 1988. "Status Quo Bias in Decision Making." *Journal of Risk and Uncertainty* 1: 7–59. Available at: http://dtserv2.compsy.uni-jena.de/__C125757B00364C53.nsf/0/F0CC3CAE039C8B42C125757B00473C77/%24FILE/samuelson_zeckhauser_1988.pdf (accessed 10th June 2014).

Skinner, C. 2012. "Statistical Disclosure Risk: Separating Potential and Harm." *International Statistical Review* 80: 349–368. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2012.00190.x/pdf (accessed June 10, 2014).

Trewin, D., A. Andersen, T. Beridze, L. Biggeri, I. Fellegi, and T. Toczynski. 2007. *Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice*. Geneva: UNECE /CES. Available at http://www.unece.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf (accessed 10th June 2014).

Varian, H. 1992. *Microeconomic Analysis*. 3rd ed. New York: W.W. Norton.

Viscusi, K., W. Magat, and J. Huber. 1987. "An Investigation of the Rationality of Consumer Valuations of Multiple Health Risks." *Rand Journal of Economics* 18: 465–479. Available at: http://www.jstor.org/discover/10.2307/2555636?uid=3738032&uid=2&uid=4&sid=21102275515957 (accessed 10th June 2014).

Received July 2012
Revised September 2013
Accepted January 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 547–562, http://dx.doi.org/10.2478/JOS-2014-0034

# Are All Quality Dimensions of Equal Importance when Measuring the Perceived Quality of Official Statistics? Evidence from Spain

*Alex Costa*[1], *Jaume García*[2], *and Josep Lluis Raymond*[3]

Quality has become the key concept in official statistics. There is a general consensus that we have to consider several components when assessing the quality of statistical information. Relevance, accuracy, timeliness, punctuality, comparability, coherence, accessibility and clarity are the dimensions most frequently mentioned. In this article we use regression analysis to evaluate the contribution of these different dimensions when assessing the overall quality of statistical products. We do this using the information collected in the structured consultation with users and experts from both inside and outside the Spanish Central Administration carried out by the Working Group of the Spanish High Council on Statistics, responsible for the preliminary draft of the proposals and recommendations of this council for the Spanish Multiannual Statistical Programme 2013–2016. We find that the above-mentioned dimensions have different weights in the overall assessment of perceived quality (with accuracy and reliability having the highest weight, and relevance having the lowest) and that the structure differs between both types of users.

*Key words:* Quality; dimensions; consultation; weight; regression.

## 1. Introduction

When discussing the first of the general issues raised in the report by the Working Party of the Royal Statistical Society (1991), namely the importance of retaining public confidence in official statistics, Fellegi (1991) mentioned the existence of a virtuous circle in official statistics: Public confidence is a prerequisite for high-quality statistics and high-quality statistics must ultimately be the foundation for public confidence. Since then, we can find

[1] Institut d'Estadística de Catalunya, Via Laietana 58, 08003 Barcelona, Spain. Email: acosta@idescat.cat
[2] Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. Email: jaume.garcia@upf.edu
[3] Department d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, Edifici B, Campus Bellaterra, 08193 Bellaterra, Barcelona, Spain. Email: josep.raymond@uab.cat
**Acknowledgments:** We wish to thank the Spanish Statistics Institute (*Instituto Nacional de Estadística*, INE) for allowing us to access the data used in this paper and, in particular, Carlos Paulogorran, Paz Sánchez and Pilar Ochoa from INE, who helped us with the preparation of the data set. We are also grateful for the comments made by Ada Ferrer and Juan M. Rodríguez on a preliminary version of this paper. We also benefited from the comments by three anonymous referees, an associate editor and the Co-Editor-in-Chief. The usual disclaimer applies. Alex Costa was Director of Planning, Coordination and Statistical Dissemination of INE between June 2009 and February 2012. Jaume García was President of INE between April 2008 and December 2011. Josep Lluis Raymond was the Chairman of the Working Group of the Spanish High Council on Statistics (SHCS) which was responsible for the preliminary draft of the proposals and recommendations of the SHCS for the Spanish Multiannual Statistical Programme 2013-2016.

explicit references to quality as a key element in official statistics in most of the relevant documents of the statistical institutions. In that regard, the preamble of the European Statistics Code of Practice (European Statistical System 2011) includes an explicit reference to quality when defining the mission of the European Statistical System (ESS), taken from the Quality Declaration of the ESS (European Statistical System 2001), which considers the provision of high-quality information on the economy and society on European, national and regional levels. Moreover, the first of the principles governing international statistical activities (United Nations Statistical Commission 2006) refers to quality, stating that high-quality international statistics are a fundamental element of global information systems.

As mentioned by Vale (2010), the concept of quality in official statistics has moved from the traditional approach in which quality corresponds to how closely statistics reflect reality (e.g., the mean square error of an estimator), that is accuracy and reliability, to a situation in which the performance of statistical services is evaluated in terms of how they respond to users' needs (Castles 1991). This new concept fits perfectly with the ISO 9000 definition of quality – the ability of a set of characteristics to satisfy requirements – and is explicitly reflected in the European Statistics Code of Practice when talking about the statistical output, as well as in the first of the Fundamental Principles of Official Statistics (United Nations Economic Commission for Europe 1992). In fact, this actual concept of quality reinforces the consideration of statistics produced by public institutions as a public good (Giovannini et al. 2009).

As mentioned in the final report of the Leadership Expert Group on quality (Eurostat 2002), quality consists of a number of features reflecting users' needs and can be defined along a number of dimensions (i.e., quality is a multidimensional concept), which constitute the product quality. Nowadays, there is almost complete agreement among international statistical institutions about the components (or dimensions) of statistical quality, as is shown in Vale's (2010) mapping of quality components in international statistical organisations (p. 6). In particular, Article 12 of Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics refers to the criteria which shall apply to guarantee the quality of results. These are: relevance; accuracy; timeliness; punctuality; accessibility and clarity; comparability; and coherence. These are the dimensions referred to in Eurostat's definition of quality of statistics introduced in 2003 (Eurostat 2003). However, since statistical offices are confronted with a heterogeneous typology of users, who have different needs and also different perceptions of quality, the importance of the components (or dimensions) will not necessarily be uniform and may differ among users. In fact, Brackstone (2001) thinks of the dimensions of quality in a hierarchical fashion. Relevance is at the top, timeliness and accessibility are not important without relevance, and accuracy, interpretability and coherence only become important when the other three dimensions are satisfied.

In this article we use the data collected in the structured consultation with users and experts from inside and outside the Spanish Central Administration carried out by the Working Group of the Spanish High Council on Statistics (SHCS) (*Consejo Superior de Estadística*) in 2010 when preparing the preliminary draft of the proposals and recommendations of the SHCS for the Spanish Multiannual Statistical Programme (SMSP) (*Plan Estadístico Nacional*) 2013–2016. The aim of these consultations was to provide

evidence on the importance of the different components when assessing data quality and, in particular, on whether different groups of users attach the same (or different) weights to the quality dimensions. Principles 11 to 15 of the European Statistics Code of Practice are used as references of the quality dimensions in the consultation questionnaire.

This study has some similarities with some recent exercises carried out in Greece (Nikolaidis 2012) and Portugal (Zilhao and Ribeiro 2012), but differs from them in terms of the way in which users' information is collected, the type of users considered and, most importantly, in terms of the main objective: the empirical analysis of the relationship between the general assessment of quality of official statistics and that of its different components.

The article is organised as follows. In the next section we describe the main features of the structured consultation. The basic results are presented in the third section. In Section 4 we report some evidence of how the dimensions of quality are related to its general assessment. The article ends with a summary of the main conclusions.

## 2. The Structured Consultation and the Institutional Framework

The SMSP is the main legal tool used by the Spanish Central Administration to define the statistical production to be developed by either its statistical services or any other entity dependent on them or in collaboration with the regional or local administration, covering a period of four years. The SMSP is approved by the government and afterwards is developed and carried out by means of the annual programmes, which are also approved by the government.

Additionally, the SHCS is an advisory body for the Central Administration statistical services in which informants, producers and users of official statistics are represented, for example trade unions and employers' organisations and other social, economic and academic groups, together with ministries and the Spanish Statistics Institute (*Instituto Nacional de Estadística*, INE). One important task of the SHCS is to contribute towards identifying the statistics to be included in the SMSP to improve the coverage of users' needs for information prior to the first draft of the SMSP. The SHCS also delivers opinions on the proposal of the SMSP before it is approved by the government.

In that regard, in March 2010 the SHSC set up a working group which was to be responsible for the preliminary draft of the proposals and recommendations for the SMSP 2013–2016. In its first meeting, the SHSC Working Group defined two main tools to evaluate the official statistics included in the actual SMSP (the Spanish official statistics from the Central Administration) as part of the first step before producing the draft: a Compliance Report by Eurostat, with an evaluation of the Spanish official statistics from the point of view of the European Statistical Regulations; and a Structured Consultation addressed to a wide range of users and experts from inside and outside the Administration. It was the first time in Spain that both the statistics produced by INE and those produced by the statistical services of the ministries were evaluated. In the past, similar exercises, but not as exhaustive, have been carried out only for the INE statistics.

The key concept of the consultation is quality, which the participants were asked to assess as such, in addition to different dimensions and aspects described below. Following the notion of data quality in the Handbook of Data Quality Assessment Methods and Tools

by Eurostat (Ehling and Körner 2007), we must consider three elements: the characteristics of the statistical product; the perception of the statistical product by the user; and the characteristics of the statistical production process. The components (or dimensions) of the product's quality were used in the consultation as a framework to assess users' perceptions, although we know that the quality assessment by different users is not necessarily the same.

As explicitly stated in the questionnaires, the components (or dimensions) of quality included are those of the ESS, which are defined in Principles 11 to 15 of the European Statistics Code of Practice. They are the following, including the literal description contained in the questionnaire:

*Relevance*: Official statistics must meet users' needs. You must consider whether the objectives of each statistic are related to the users' expectations and to the potentialities of the data source.

*Accuracy and Reliability*: Official statistics must portray reality in an accurate and reliable way. You must consider the degree of closeness with reality based on the methodology used, paying attention to the sampling and nonsampling errors plus the biases associated with the different stages of the process.

*Timeliness and Punctuality*: Official statistics must be released in a timely and punctual way. Official statistics must be disseminated with the shortest time possible having passed between the date they become available and the event they describe, and according to a previously announced calendar.

*Coherence and Comparability*: Official statistics must be comparable over time, through space and between domains. Official statistics must allow related data from different sources to be used and combined in a reliable manner.

*Accessibility and Clarity*: Official statistics must be presented in a clear and understandable form, disseminated in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.

The overall quality assessment follows the five aforementioned quality dimensions. Note that the dimensions are complex and highly abstract concepts and consequently that measurement could be an issue that could affect the interpretation of the results. However, as will be discussed below, since the participants in the consultation are users with a high level of expertise and experts, we assume that measurement problems are minimised.

The evaluation of the quality of the product and its components in the consultation refer to single statistics. Results can be aggregated at a sectoral level by assigning each statistic to a sector. In fact, each participant in the consultation could evaluate all the statistics in a particular sector, that is, each questionnaire refers only to one sector. The 279 statistics included in the Structured Consultation came from the Spanish Annual Statistical Programme 2010, and were classified in 22 sectors.

There are also some questions in the questionnaire which correspond to assessments at the sectoral level. In particular, one question refers to the extent to which available statistics meet users' needs in a particular sector (coverage). This is proxying relevance at a sectoral level, but it is not necessarily equivalent to the aggregation of the relevance scores for each statistic in a sector, since it could happen that some relevant information is not covered by any statistic. There is also a question seeking the assessment of the quality of several dissemination channels (press releases, yearbooks, short-term newsletters, web

pages and individual requests) at the sectoral level. All the assessments in the Structured Consultation are made according to an ordinal scale: very low (1); low (2); medium (3); high (4); and very high (5).

With respect to the selection of the sample, the difficulties with this type of users' satisfaction survey are well known. Considering the main objective in the Structured Consultation was to evaluate quality (and its components) and that in order to do this substantial knowledge of the official statistics is required, a purposive sample was used in which participants were selected from those proposed by the members of the SHSC Working Group as experts for a particular sector and with good prospects of collaboration. There are of course other alternative ways of selecting the sample, such as for instance a random sample of those making specific requests for information to the producers of official statistics but, given the data availability and the emphasis on having people with knowledge of the statistics, we ended up with a purposive sample. If possible, as a pending task for future research, it could be informative to compare the results of both approaches.

The sample of participants can be classified in three different groups: users with a high level of expertise, not coming from neither the Central nor the Regional Administration; users and experts from the Central Administration, who could also be producers (in particular the ministries and other public institutions) of some of the statistics; and users and experts from the Regional Administration, who are mainly users of the official statistics of the Central Administration, but are also producers of some regional information and collaborate in the production of some official statistics at national level. The questionnaires were sent as an Excel spreadsheet by e-mail to the participants, some of them receiving more than one questionnaire, each one including several single statistics. The data collection was carried out between June and July 2010. Table 1 presents the distribution of participants.

Table 2 summarises basic information about the questionnaires and the observations according to the three types of participants we considered. In total, 717 questionnaires (corresponding to sectors) were sent to the 236 participants with a response rate of 88.6%, and 599 of them were completed and included in the empirical analysis, generating 4,711 valid observations (complete assessments of single statistics), that is, on average each

*Table 1. Distribution of participants (users and experts) in the consultation*

| | Number of Participants |
|---|---|
| *Outside the Administration* | **130** |
| Universities | 79 |
| Unions and employers' organisations | 20 |
| Other (media, research institutions etc.) | 31 |
| *Central Administration* | **75** |
| INE | 32 |
| Ministries | 28 |
| Other (agencies, institutes etc.) | 15 |
| *Regional Administration* | **31** |
| Regional Statistical Offices | 17 |
| Other (regional ministries, regional agencies etc.) | 14 |
| **Total** | **236** |

*Table 2. Distribution of the number of questionnaires and observations*

| | Sent questionnaires | Received questionnaires | % | Valid questionnaires | % | Valid observations |
|---|---|---|---|---|---|---|
| Outside the Administration | 232 | 165 | 71.1 | 143 | 61.6 | 1,216 |
| Central Administration | 155 | 140 | 90.3 | 135 | 87.1 | 1,275 |
| Regional Administration | 330 | 330 | 100.0 | 321 | 97.2 | 2,220 |
| Total | 717 | 635 | 88.6 | 599 | 83.5 | 4,711 |

participant answered 2.5 valid questionnaires corresponding to 20 single statistics. It is important to note the significantly lower response rate among users outside both the Central and the Regional Administration.

## 3. Evidence from the Structured Consultation

In Tables 3 and 4 we report the basic results of the Structured Consultation: the distribution and the average of the scores for the different dimensions of quality and the dissemination channels (Table 3), and compared for the three types of participants in the consultation (Table 4). It is important to note that traces of straightlining in the responses, that is, not differentiating between the response categories for the dimensions and that of the overall assessment, have been observed for about 20 per cent of the sample. However, correction for this does not affect the empirical results significantly.

Based on the evidence from Tables 3 and 4, we can make the following comments:

– The frequency distributions of the different variables show that the mode for all the assessments is the score "high", except for the dimension Relevance which is "very high". All the average scores are between 3 ("medium") and 4 ("high"), again with

*Table 3. Frequencies (%) and averages of the scores of the different quality variables*

| | Very low 1 | Low 2 | Medium 3 | High 4 | Very high 5 | Average |
|---|---|---|---|---|---|---|
| Coverage | 1.4 | 7.1 | 35.4 | **51.6** | 4.5 | 3.51 |
| Quality of statistics | 0.7 | 5.5 | 28.9 | **49.5** | 15.2 | 3.73 |
| *Quality dimensions* | | | | | | |
| Relevance | 0.7 | 4.4 | 15.8 | 36.4 | **42.7** | 4.16 |
| Accuracy and Reliability | 1.5 | 6.6 | 27.5 | **47.2** | 17.2 | 3.72 |
| Timeliness and Punctuality | 1.8 | 8.2 | 25.7 | **40.8** | 23.5 | 3.76 |
| Coherence and Consistency | 1.9 | 8.0 | 28.6 | **43.1** | 18.4 | 3.68 |
| Accessibility and Clarity | 2.2 | 6.9 | 23.6 | **44.5** | 22.8 | 3.79 |
| *Dissemination channels* | | | | | | |
| Web page | 3.8 | 5.7 | 21.6 | **43.2** | 25.7 | 3.81 |
| Individual request | 11.6 | 6.8 | 21.0 | **37.6** | 23.0 | 3.54 |
| Press releases | 7.8 | 15.8 | 23.9 | **37.0** | 15.5 | 3.37 |
| Yearbooks | 7.6 | 13.5 | 28.1 | **39.9** | 10.8 | 3.33 |
| Short-term newsletters | 7.4 | 13.5 | 34.7 | **35.5** | 8.9 | 3.25 |

Note: Mode in bold type

*Table 4. Average scores by type of participants in the consultation*

| | Total | Users not in Admin. | Central Admin. | Regional Admin. |
|---|---|---|---|---|
| Coverage | 3.51 | 3.40 | 3.76 | 3.44 |
| Quality of statistics | 3.73 | 3.63 | 3.90 | 3.69 |
| *Quality dimensions* | | | | |
| Relevance | 4.16 | 4.19 | 4.36 | 4.03 |
| Accuracy and Reliability | 3.72 | 3.59 | 3.95 | 3.66 |
| Timeliness and Punctuality | 3.76 | 3.60 | 3.83 | 3.81 |
| Coherence and Consistency | 3.68 | 3.56 | 3.76 | 3.70 |
| Accessibility and Clarity | 3.79 | 3.69 | 3.93 | 3.76 |
| *Average of dimensions* | *3.82* | *3.72* | *3.97* | *3.79* |
| *Dissemination channels* | | | | |
| Web page | 3.81 | 3.81 | 3.95 | 3.37 |
| Individual request | 3.54 | 3.36 | 3.92 | 3.38 |
| Press releases | 3.37 | 3.14 | 3.63 | 3.38 |
| Yearbooks | 3.33 | 3.30 | 3.71 | 3.16 |
| Short-term newsletters | 3.25 | 3.21 | 3.63 | 3.07 |
| *Average of channels* | *3.46* | *3.36* | *3.77* | *3.27* |

the exception of that of relevance (4.16). The perceived quality of the official statistics (3.73) is higher than the assessment of the coverage for the sectors (3.51) and also than that of dissemination (3.46 if measured as the average of the scores for the different channels). In fact, almost two thirds of the participants rate the quality of the statistical information as "high" or "very high".

– The average score of the assessment of the extent to which official statistics meet users' needs at a sectoral level (coverage) is significantly below the average score of relevance (3.51 vs 4.16), which also refers to meeting users' needs but for a single statistic. In fact, the score of coverage is also lower than that related to quality of the statistical product (3.51 vs 3.73). This significant difference between the scores of coverage and relevance shows that although available official statistics adequately meet the requirements of the users in those fields mentioned, there is still a lack of statistical information at the sectoral level. Coverage may be poor even if each of the statistics is highly relevant.
– The average score of the five dimensions of the quality of statistical information (3.82) is above the overall average (3.73). According to a t-test for the equality of the means of two variables, the differences of these scores with respect to the overall average are statistically significant, except for Accuracy and Reliability. In any case, except for Relevance, the averages of these scores and that of the overall assessment appear to be quite homogenous given that they do not differ more than two decimal points. Three dimensions (Relevance; Timeliness and Punctuality; Accessibility and Clarity) have average scores of above 3.73.
– The assessment of quality of dissemination by channel (contrary to what was observed in the assessment of the quality of statistical information) is clearly heterogeneous. The average value of the scores of the different channels (3.46), see Table 4, is derived from a positive evaluation of web pages (3.81), a medium-high evaluation of

individual requests (3.54) and less positive ratings for the other channels: press releases (3.37), yearbooks (3.33) and newsletters (3.25). It is interesting to note that the accessibility and clarity in the dissemination of information, as a dimension of the quality of statistical products, has a similar average to the web page channel (both are valued at around 3.8). Therefore, the quality of access does not seem to be significantly influenced by the less positive ratings of the other channels, probably as a consequence of the fact that, given the profile of the participants in the consultation, they rely more on those channels that have a higher evaluation (web pages and individual requests).

- As mentioned in Ehling and Körner (2007), evidence from Table 4 shows that different types of users and experts of official statistics perceive product quality differently. The average scores for all the variables are higher for the group of participants from the Central Administration, who also have a profile of producers. On the other hand, except for the assessment of relevance, the average scores of quality dimensions for the group of users not in the Administration are the lowest. Almost all the differences between the averages among the different groups of participants are statistically significant (t-test).
- The differences between the averages of the score of Relevance for the Central Administration and the Regional Administration are more important than for the other dimensions. This could be explained by the fact that those participants belonging to the Regional Administration are mainly users of the official statistics produced by the Central Administration, and they have a genuine interest in the territorial information at a more disaggregated level than that of the official figures. This is also corroborated at the sectoral level. The average of the coverage variable is quite similar between users and participants from the Regional Administration and is substantially different from the participants of the Central Administration.
- The genuine interest of the Regional Administration in the territorial information could also explain the substantial difference between the average scores for accuracy and reliability. The rating of these characteristics of the quality of the statistical product worsen when we consider more disaggregated geographical areas. However, since users not in the Administration also have a low score (3.59), part of this observed difference could be explained by some overrating of this dimension by the Central Administration, given that they are also responsible for the statistical production.
- It must be pointed out that the results for the total sample in Table 4 do not change significantly if we weight the observations of each group of participants differently. For instance, if we calculate the total values as the average of the estimates for each group, rather than the average of all responses, since regional administrators reported for many more statistics than other groups, then the average total scores are: 3.74 (overall assessment of quality), and 4.19, 3.73, 3.75, 3.67 and 3.69 (scores for the quality dimensions respectively). This pattern of different weights having no relevant effects applies to all the estimates in the article.
- When considering participants' assessments of the dissemination channels, it can be observed that the contents of the web page are scored highly by all three groups. Something similar occurs for the individual requests, although the score of users not in the Administration is significantly lower than that of the web page. This fact may be reflecting the difficulties of accessing microdata in some cases, in particular

administrative data. Finally, the evidence from the score of press releases seems to indicate that they are aimed more at users of the Regional Administration, who produce some regional data.

## 4. The Relationship Between the Quality of the Statistical Product and Its Dimensions

The main objective of this article is to try to evaluate how the different dimensions of the quality of statistical products influence the overall assessment, that is, to quantify what weight each dimension has when comprising the overall assessment of quality. In that regard, the simplest and most intuitive way of relating a variable (overall assessment) and its components (the different dimensions), given that they have the same scale (1–5), is by interpreting the overall assessment as an average of the scores of the different dimensions. This simplest interpretation of how the overall assessment is formed implies an equal weight for each of the dimensions.

We can compare the unweighted average scores of the five dimensions for all the participants in the consultation to the reported overall assessment. These averages are the result of calculating the average of the scores for the dimensions and recoding them on a scale of 1 to 5 by rounding the value of this average: being less than 1.5 (1); between 1.5 and 2.5 (2); between 2.5 and 3.5 (3); between 3.5 and 4.5 (4); and more than 4.5 (5). We obtain the result that in 82.6% of the cases the average prediction coincides with the observed assessment; in 13.5% of the cases it is higher, whereas in only 3.9% of the cases it is lower. This asymmetry can be interpreted as a downwards "correction" of the prediction, which can be thought of as subtracting a constant term from the average.

To provide further evidence about the type of relationship we expect to observe between overall quality and its components, we examine the correlation coefficients between these variables, reported in Table 5. It can be observed that the correlation coefficients between the different dimensions are positive and that they are above 0.5 except for the cases involving the relevance dimension (around 0.35). Furthermore, when looking at correlations with the overall assessment a similar pattern emerges, that is, correlations above 0.7 except for the relevance dimension (0.52). This shows that the influence of the dimensions will probably not be uniform and that of relevance will be lower than those

*Table 5. Correlation matrix of the assessment variables*

| | C | Q | R | A-R | T-P | C-C | A-C |
|---|---|---|---|---|---|---|---|
| Coverage (C) | 1.00 | | | | | | |
| Quality (Q) | 0.44 | 1.00 | | | | | |
| Relevance (R) | 0.16 | 0.52 | 1.00 | | | | |
| Accuracy and Reliability (A-R) | 0.28 | 0.75 | 0.35 | 1.00 | | | |
| Timeliness and Punctuality (T-P) | 0.34 | 0.71 | 0.34 | 0.51 | 1.00 | | |
| Comparability and Coherence (C-C) | 0.35 | 0.79 | 0.38 | 0.64 | 0.60 | 1.00 | |
| Accessibility and Clarity (A-C) | 0.34 | 0.72 | 0.36 | 0.54 | 0.61 | 0.68 | 1.00 |

Note: Correlations are calculated for the average values of the assessments of the quality dimensions by each participant in the consultation (not for each single statistic), given that the assessment of coverage is unique for each participant and sector.

of the others. On the other hand, when looking at correlation with the coverage assessment, correlations are much lower, reflecting the fact that the coverage assessment implicitly takes into account those needs in a sector which are not met by the actual statistics (those for which we evaluate the quality components).

By using multidimensional scaling (Kruskal and Wish 1978), we can transform the correlations (proximities or similarities) between the seven assessment variables (objects) in the matrix of Table 5 into seven vectors in $R^N$, usually N equal to 2 or 3, preserving the "hidden structure" of the data in terms of similarities and where the output is a spatial map, in our case a two-dimensional spatial map.

The seven points in Figure 1 correspond to the seven vectors in $R^2$, the elements of which correspond to *Dimension 1* and *Dimension 2*. Moreover, the smaller the correlation (proximity) between two assessment variables, the further apart the corresponding points (vectors) should be on the map, facilitating the visualisation of the structure of similarities of the data. In particular, in our case we identify five assessment variables that are really close (overall quality, accuracy and reliability, timeliness and punctuality, comparability and coherence, and accessibility and clarity). Since we are interested in measuring how the different components of quality contribute to the perceived overall quality, the structure of the spatial map leads us to expect that the four components included in this group will have a more substantial influence on the overall assessment than relevance.

All the previous evidence seems to indicate that a simple and intuitive mechanism such as a weighted average (to capture the potentially different influence of each component) corrected by a constant term (to take into account the overprediction which was suggested by the use of the simple average) can approximate the relationship between the overall assessment of quality and its components, that is, a mechanism based on a regression model with a constant term where the dependent variable is the score of the quality of the statistical product, the regressors are the scores of the five dimensions of quality defined in the European Code of Practice, and its coefficients are restricted to add one (weighted average).

We must point out that nonlinear specifications, such as Cobb-Douglas (a weighted geometric average with a constant term) or Translog functions, which allow for an

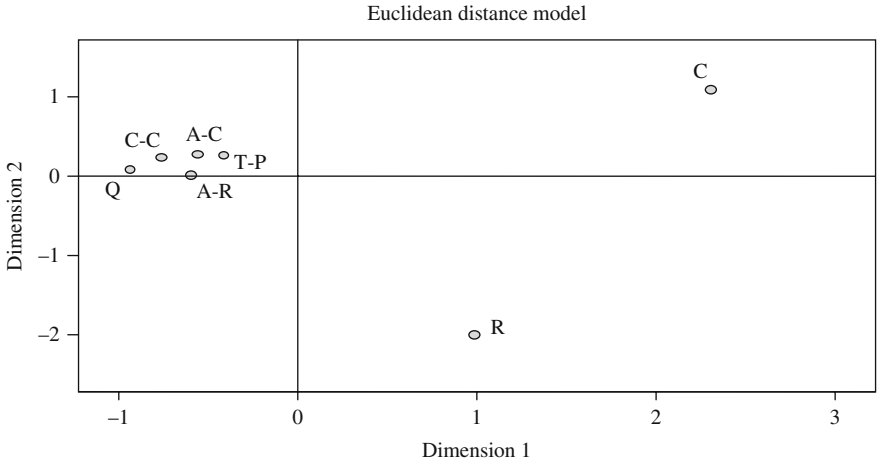



*Fig. 1. Multidimensional scaling for the assessment variables*

interpretation as a production function of the relationship between the overall assessment and its components, were tried and none of these alternatives could offer a relevant better fit than the simple linear relationship we propose from a practical point of view. If a simple model can offer a similar fit capacity to more complex models, while at the same time facilitating the interpretation of the results, the Occam's razor principle suggests that the selection of the more parsimonious formulation is appropriate.

Using the acronyms of the variables in Table 5, the estimated model has the following specification:

$$Q_{ij} = \beta_0 + \beta_1 R_{ij} + \beta_2 (A - R)_{ij} + \beta_3 (T - P)_{ij} + \beta_4 (C - C)_{ij} + \beta_5 (A - C)_{ij} + u_{ij} \quad (1)$$

where indexes $i$ and $j$ correspond to the participant and the single statistics, respectively, $u$ is the error term and the $\beta s$ are the coefficients, which can be interpreted as the expected value of the weights in a random coefficient model (Swami and Tavlas 1995), where $\beta_{k,ij} = \beta_k + \eta_{k,ij}$, $k = 0, 1, \ldots, 5$, $\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 = 1$, and $\eta$ is a random variable with a zero mean, that is, the weights are different for each participant and each statistic, although the $\beta_{k,ij}$ coefficients cannot be estimated.

The model in Equation (1) is estimated by OLS with 4,711 observations. The estimation results are presented in Table 6. Those corresponding to the whole sample (Total) are in the first column, whereas from the second to the fourth column we report those corresponding to the groups of participants we are considering: users not in the Administration (U), Central Administration (C.A.) and Regional Administration (R.A.), respectively. By doing this we are allowing for some variability in the weights depending on the type of user we consider.

The results in the first column show that this simple and intuitive model (a corrected weighted average) has a substantial explanatory power, almost 80% of the variability of the dependent variable. Moreover, the significance of the coefficient of the constant term corroborates the correction we mentioned when looking at the descriptive analysis with the expected (negative) sign, that is, the weighted average is corrected downwards when configuring the overall assessment.

At the same time, the coefficients of the components (the weights) are statistically different from 0.2, which would be the weight for each component in the case of a simple average, meaning that not all the components have the same influence in assessing overall quality. Note that the sum of the estimated coefficients (weights) without imposing the constraint of the sum being equal to one would be 1.0188, 1.0278, 1.0010, and 1.0110 for the four models in Table 6, respectively, that is, the sums are practically equal to one.

In particular, accuracy and reliability is the dimension with the highest weight (23.2%), whereas relevance has the lowest (15.4%), as we expected from the descriptive analysis of the previous section. There are almost eight percentage points of difference between both weights, which implies accuracy and reliability receive more than 50 per cent more weight than relevance, that is, the differences are not only statistically significant but also practically relevant. This feature of different weights for each dimension is particularly relevant when considering and analysing the trade-offs between the dimensions, an issue that is becoming increasingly important, as mentioned in Ehling and Körner (2007). But this point cannot be interpreted, for instance, as a recommendation to prioritise accuracy

*Table 6. Estimation results for the regression model*

| | Total<br>*Coef. (t-stat.)* | Users not in Admin.<br>*Coef. (t-stat.)* | Central Admin.<br>*Coef. (t-stat.)* | Regional Admin.<br>*Coef. (t-stat.)* |
|---|---|---|---|---|
| Relevance ($\beta_1$) | 0.154 (19.37) | 0.129 (8.59) | 0.124 (8.15) | 0.179 (15.29) |
| Accuracy and Reliability ($\beta_2$) | 0.232 (23.00) | 0.243 (12.61) | 0.201 (10.94) | 0.241 (15.85) |
| Timeliness and Punctuality ($\beta_3$) | 0.182 (21.17) | 0.162 (10.13) | 0.228 (14.76) | 0.171 (12.30) |
| Coherence and Comparability ($\beta_4$) | 0.224 (21.30) | 0.240 (11.50) | 0.251 (12.99) | 0.205 (13.02) |
| Accessibility and Clarity ($\beta_5$) | 0.208 (22.35) | 0.226 (13.18) | 0.196 (9.59) | 0.204 (14.61) |
| Constant ($\beta_0$) | − 0.070 (11.09) | − 0.053 (4.44) | − 0.022 (1.96) | − 0.093 (9.94) |
| Adjusted $R^2$ | 0.787 | 0.811 | 0.811 | 0.749 |
| Sample size | 4711 | 1216 | 1275 | 2220 |
| $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.2$<br>F statistic, p-values in brackets | 12.11 (0.000) | 9.22 (0.000) | 8.09 (0.000) | 3.36 (0.009) |
| $H_0$: $\beta_{k,U} = \beta_{k,CA} = \beta_{k,RA}$<br>F statistic, p-values in brackets | | | | |
| *Relevance* | | | 5.47 (0.004) | |
| *Accuracy and Reliability* | | | 1.70 (0.183) | |
| *Timeliness and Punctuality* | | | 5.31 (0.005) | |
| *Coherence and Comparability* | | | 1.97 (0.140) | |
| *Accessibility and Clarity* | | | 0.73 (0.480) | |
| $H_0$: $\beta_{k,U} = \beta_{k,CA} = \beta_{k,RA}$ k = 1,2,3,4,5<br>F statistic, p-values in brackets | | | 3.35 (0.001) | |

Note: The standard errors used in the calculation of the t-statistics are calculated with a robust estimator to allow for a nonscalar variance-covariance matrix of the errors.

and reliability at the expense of relevance. In fact, relevance is the basis for other quality dimensions and to some extent is a precondition for being a user at all.

The results for the three groups of participants confirm what has been pointed out by several authors (e.g., Ehling and Körner 2007), namely that different users have a different perception of quality and assign a different level of importance to the dimensions. As shown in Table 6, we can reject the null hypothesis of a unique structure of weights (coefficients) for the three groups of participants, as well as the hypothesis that for each group the weights for the different dimensions are the same (0.2).

In particular, the most significant differences between the weights for the different groups correspond to Relevance and Timeliness and Punctuality. Relevance has the lowest coefficient out of all three groups, but it is given significantly higher weight by the Regional Administration than by the Central Administration and users not in the Administration. Having the lowest weight could be a consequence of the fact that, as mentioned by Brackstone (1999), relevance is a dimension which should be considered across the whole output of a statistical office rather than per statistic, as we do in the exercise. The higher weight in the Regional Administration group can be explained by its special interest, as a user, in acquiring more detailed territorial information.

The weight of Timeliness and Punctuality is significantly higher for Central Administration. In fact, this dimension as well as Coherence and Comparability have the higher weights in this group, whereas Accuracy and Reliability occupies the third place in terms of the magnitude of the weights; by contrast, this dimension has the highest weight both for users not in the Administration and Regional Administration. This can be explained by the fact that Central Administration participants also have the role of producers, and this dimension has traditionally been associated with quality and essentially been under the control of producers. Additionally, in recent years and because of (among other things) the economic crisis and the need for data for international comparisons, increasing attention has been paid to issues related to timeliness and punctuality as well as to comparability. This has been translated into a closer association of these two dimensions with quality from the producers' side. The same conclusion is reached if we look at the proportion of coincidences between the score of one dimension and that of the overall assessment of quality. In the case of Timeliness and Punctuality and Comparability and Coherence, the Central Administration presents the highest proportions, whereas in the case of Accuracy and Reliability it has the lowest.

Finally, there are no significant differences between the participant groups in the case of Accessibility and Clarity. In principle, this is also a dimension that should be evaluated for the whole set of statistics produced by a statistical office rather than for one individual statistic, but in our case, since not only the statistical office (INE) is producing official statistics (ministries and other public institutions also do so), the importance of the weight could be capturing some heterogeneity in the way the different producers are making the information accessible and clear.

## 5. Conclusions

Quality has become a key concept in official statistics. At the same time however, this is a concept which can refer to different aspects (product, process) and which can be evaluated

from different viewpoints (users, producers) and with different tools (reports, users' satisfaction surveys, peer reviews).

In this article we have approached the evaluation of the quality of the statistical products from the users' side and considered the characteristics (dimensions or components) of the statistical product as they are described in the European Code of Practice. The evaluation has been performed using the information generated by the structured consultation carried out in 2010 by the SHCS Working Group in Spain when preparing the draft of the proposals and recommendations for the Spanish Multiannual Statistical Programme 2013–16. In this consultation both users not in the Administration and users in the Administration, who are also producers, were included in a purposive sample with the objective of evaluating the official statistics included in the Annual Statistical Programme 2010 by answering a questionnaire for the assessment of the different dimensions of quality.

Additionally, we have included evidence of the importance of the different dimensions of quality in order to assess the overall perceived quality of the statistical products. In order to do so, we have followed a simple and intuitive approach based on regression analysis to estimate the weights of a weighted average of the different components plus a constant term.

The main conclusions of this study referring to the evaluation of the quality of the Spanish official statistics can be summarised as follows:

- The overall assessment of the quality of individual official statistics in Spain is "high" or "very high" for 64.7% of the participants in the structured consultation, and only in 6.3% of the cases the score was "very low" or "low". In fact, for all the different variables analysed the mode is "high" with the exception of relevance ("very high"). Relevance is the dimension with the highest score (above 4), whereas the other four dimensions of quality have very similar scores.
- The assessment of coverage of official statistics at the sectoral level is lower than that of the relevance dimension for individual statistics and that of the overall quality. This indicates that some users' needs are still not fully satisfied in some fields (sectors), showing that there is still a lack of information at the sectoral level.
- Traditional dissemination channels have relatively low scores compared to the web page channel.
- There are significant differences between the scores for the three groups of participants we consider (users not in the Administration, Central Administration and Regional Administration). In general, the highest scores are found for Central Administration and the lowest for users not in the Administration.

The main conclusions regarding the weight of the quality dimensions when assessing overall quality can be summarised as follows:

- A weighted average of the five quality dimensions we have considered (those in the European Statistics Code of Practice) plus a constant term is a fairly good representation of how dimensions are taken into account when making an overall assessment of perceived quality of official statistics. All the dimensions make a significant contribution for all the different groups of participants we have considered in the consultation.

– The contribution of the different dimensions to the overall assessment of the quality of the statistical products is not uniform, that is, the weights are significantly different between dimensions. Accuracy and Reliability is the dimension with the highest weight in the model estimated for the whole sample, although this ranking and the value of the weights differ depending on the type of participants considered. In future work we plan to analyse whether there are significant differences in the weights across groups of statistics (e.g., by sector, short term vs. structural).
– Relevance is the dimension seen to have the highest score but is given the lowest weight in judging overall quality as perceived by users. This may reflect the fact that the participants in the structured consultation assume that the statistics they evaluate are relevant, so they do not focus too much on relevance in determining overall quality.

## 6. References

Brackstone, G. 1999. "Managing Data Quality in a Statistical Agency." *Survey Methodology* 25:139–149.

Brackstone, G. 2001. "How Important is Accuracy?" In *Proceedings of Statistics Canada Symposium 2001, "Achieving Data Quality in a Statistical Agency: A Methodological Perspective"*. Available at: http://www.statcan.gc.ca/pub/11-522-x/2001001/session24/6311-eng.pdf (accessed February 4, 2013).

Castles, I. 1991. "Responding to Users' Needs." *Journal of the Royal Statistical Society (Series A)* 154:6–10.

Ehling, M. and T. Körner. 2007. *Handbook on Data Quality Assessment Methods and Tools*. European Commission, Eurostat.

European Statistical System 2001. "Quality Declaration of the European Statistical System." Statistical Programme Committee, September 20, 2001. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/DECLARATIONS.pdf (accessed May 20, 2014).

European Statistical System 2011. "European Statistics Code of Practice." European Statistical System Committee, September 28, 2011. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-32-11-955 (accessed May 20, 2014).

Eurostat 2002. "Quality in the European Statistical System. The Way Forward". Luxembourg: Office for Official Publications of the European Communities.

Eurostat 2003. "Methodological Documents – Definition of Quality in Statistics." Sixth meeting of the Working Group "Assessment of Quality in Statistics", Luxembourg, Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/ess%20quality%20definition.pdf (accessed May 20, 2014).

Fellegi, I.P. 1991. "Maintaining Public Confidence in Official Statistics." *Journal of the Royal Statistical Society (Series A)* 154:1–6.

Giovannini, E., J. Oliveira Martins, and M. Gamba. 2009. "Statistics, Knowledge and Governance." *Statistika* 6:471–490.

Kruskal, J.B. and M. Wish. 1978. "Multidimensional Scaling." *Sage University Paper series on Quantitative Applications in the Social Sciences* 07-011. Beverly Hills and London: Sage Publications.

Nikolaidis, I. 2012. "Users' Perceptions of Statistics for Improvement Actions." Paper presented at the European Conference on Quality in Official Statistics, Athens, May 2012.

Statistics Canada 2009. *Corporate Business Plan. 2009/10 to 2011/12*. Statistics Canada.

Swami, P.A.V.B. and G.S. Tavlas. 1995. "Random Coefficient Models: Theory and Applications." *Journal of Economic Surveys* 9:165–196.

United Nations Economic Commission for Europe 1992. *Fundamental Principles of Official Statistics*. Available at: http://www.unece.org/stats/archive/docs.fp.e.html (accessed May 20, 2014).

United Nations Statistical Commission 2006. *Report of the Secretary-General on Principles Governing International Statistical Activities. Report on the Thirty-Seventh Session (March 6–10, 2006)*. Available at: http://unstats.un.org/unsd/statcom/doc06/2006-13e-Principles.pdf (accessed May 20, 2014).

Vale, S. 2010. "Statistical Data Quality in the UNECE. 2010 Version." Available at: http://unstats.un.org/unsd/dnss/docs-nqaf/UNECE-Quality%20Improvement%20Programme%202010.pdf (accessed May 20, 2014).

Working Party of the Royal Statistical Society 1991. "Official Statistics: Counting with Confidence." *Journal of the Royal Statistical Society (Series A)* 154:23–44.

Zilhao, M.J. and M. Ribeiro. 2012. "Measuring Client's Satisfaction – the Integrated Management System of the Post-Service Satisfaction Survey at Statistics Portugal." Paper presented at the European Conference on Quality in Official Statistics, Athens, May 2012.

Received October 2012
Revised February 2014
Accepted March 2014

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 563–565, http://dx.doi.org/10.2478/JOS-2014-0035

# Book Review

*Peter-Paul de Wolf*[1]

**Jörg Drechsler.** *Synthetic Datasets for Statistical Disclosure Control, Theory and Implementation*. 2013. NY: Springer, ISBN 9-781461-403258, $99USD.

Nowadays, (national) statistical institutes increasingly feel pressure from the research community to provide datasets with high utility. Researchers want to be able to perform their analyses on the rich datasets that are available at those statistical institutes. The use of tabular data, traditionally provided by statistical institutes, often no longer suffices for their research purposes. They prefer to have microdata available at their own desk to which they can then apply their analyses.

Even though national statistical institutes are often quite willing to meet the wishes of the researchers to some extent, their national statistical laws are often a complicating factor. Among other things, these laws oblige national statistical institutes to safeguard the confidentiality of their data providers' responses. Keeping the response confidential also helps to build the trust that is needed between statistical institutes and their respondents. Only when institutes are trustworthy are respondents willing to provide detailed and sensitive information.

This book, based on a PhD thesis, deals with a method of producing detailed microdata, maintaining utility whilst respecting the confidentiality issues just mentioned. The book deals specifically with a method for producing synthetic datasets, using multiple imputation techniques.

In Chapter 2, the author discusses the history of the use of multiply imputed datasets within the statistical disclosure control setting. Multiple imputation is an approach wherein multiple datasets are created, each with newly imputed missing values. Multiple imputation retains the advantages of imputation, while allowing the uncertainty due to imputation to be directly assessed. Multiply imputed datasets are also used in the setting of nonresponse. At present the method is used more often in the United States than in Europe in both of these settings. Chapter 3 acts as a background chapter on multiple imputation techniques. This is very convenient as it allows readers to follow the rest of the book with ease.

In Chapter 4, a specific dataset is discussed that is used throughout the book as the major example: the German IAB establishment panel. This panel is based on the German employment register and is used to produce official statistics. It is considered one of the most important business surveys in Germany and is very popular among researchers. Chapter 5 is used to show how multiple imputation can be used to address nonresponse, and at the end of the chapter it is explicitly applied to the German IAB establishment

[1] Department of Process Development, IT and Methodology, Statistics Netherlands. Email: pp.dewolf@cbs.nl.

panel. It is shown how estimates can be constructed based on multiple datasets, each being a separate instance of applying the same imputation technique to the original dataset.

Chapter 6 deals with the fully synthetic dataset approach. This means that for a sample, all variables of all nonsampled units in the sample frame are imputed. Then multiple samples are drawn from that fully imputed sample frame. As a special case, one might even go further, fully synthesizing all units in the sample frame and "only" using the original sample to construct a proper imputation model. That way, no original "real" data is left in any of the sampled units.

At present no agency has adopted the fully synthetic approach. One version adopted called partially synthetic datasets is discussed in Chapter 7: here, only some of the variables are replaced using multiple imputation techniques. Although the author speculates that the variables being replaced could be sensitive variables as well as key identifiers, the disclosure risk measures he deals with in this chapter are only related to identification disclosure. Identification disclosure refers to the situation that a single record in a dataset can be linked to a known individual, that is, that individual can be identified in the dataset. These risk measures are thus only influenced by replacing values of key identifiers. At the end of this chapter, the author discusses some pros and cons of fully synthetic datasets versus partially synthetic datasets. The main conclusion is that fully synthetic data sets are harder to produce, but reduce the disclosure risk more effectively. On the other hand, partially synthetic datasets often have a higher utility because fully synthetic datasets are completely determined by the imputation model, whereas partially synthetic datasets still contain 'original' (nonsensitive) data. Moreover, partially synthetic datasets are usually easier to produce.

The first seven chapters deal with multiple imputation in relation to disclosure control. However, imputation techniques are also used to correct for nonresponse. In Chapter 8, the author discusses a way to combine multiple imputation techniques to correct for nonresponse and to limit the disclosure risk at the same time. This is applied to partially synthetic datasets only. The major part of this chapter is devoted to the IAB establishment panel example. At the end, an interesting issue is touched upon. It is stated that "Since the intruder never knows if her match is correct . . . the data are well protected from these kinds of attacks." This poses the question of how well a dataset is protected by uncertainty about the information supposedly disclosed. Indeed, in some cases "disclosing" untrue information might do more harm than disclosing true information about an individual unit.

In the case of multiple imputation, multiple datasets ($m$) are being released. The larger the $m$, the higher the utility of the dataset (the additional variance introduced by the imputation decreases with the number of released datasets), but at the same time the higher the disclosure risk. In Chapter 9, a two-stage imputation process is discussed to deal with the trade-off between utility and risk. Finally, in Chapter 10, some arguments are given to promote the use of multiple imputation techniques in deriving synthetic datasets, over 'standard' SDC techniques, such as local suppression, global recoding or top-coding. These arguments try to deal with the scepticism about the method, the tendency to stick with 'known' methods and the reluctance to use new methods before they are implemented in known statistical software.

This book gives a good overview of recent developments within the area of multiple imputation as a technique of deriving synthetic datasets. It is not an easy book, however.

The notation throughout the book is not always consistent and contains some minor typos. Specifically, Chapters 8 and 9 have some formulas that are very much alike, but are difficult to compare because slightly different notation is used. This makes reading the book and using it for reference somewhat difficult. On the other hand, the use of a single example throughout all chapters (the IAB establishment panel) is very beneficial. It shows the effects of the different methods on the same "real-life" dataset.

In the area of statistical disclosure control, the term "transparency" has received a lot of attention again recently. This term is related to the advantages and disadvantages of revealing information about the statistical disclosure control methods used to produce safe data. In this book, it is evident that transparency is crucial to improve the use of synthetic datasets. Obviously, not all information of the imputation model needs to be released along with the synthetic datasets. However, some information, for example which variables are included in the imputation model, can be used by a researcher to determine whether his or her analysis is still likely to be valid.

Transparency should also apply to the methods themselves. This book is a good example of providing insight into the methods that can be used to produce datasets that are useful to researchers while at the same time limiting the disclosure risk.

*The views expressed in this review are those of the author and do not necessarily reflect the policies of Statistics Netherlands.*

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 567–570, http://dx.doi.org/10.2478/JOS-2014-0036

# Book Review

*Whitney Kirzinger*[1]

**Lawrence Hubert and Howard Wainer.** *A Statistical Guide for the Ethically Perplexed*. 2013 New York: CRC Press, ISBN 978-1-4398-7368-7, 565 pp., $49.95.

This text draws attention to a topic that often seems to be neglected or at least carelessly regarded: the intersection of statistics and ethics. The work is modeled on a manuscript written by a medieval Jewish philosopher as an attempt to harmonize his philosophical views with Jewish law. Similarly, this text represents Hubert and Wainer's attempt to balance statistics and standards of ethical practice. The authors achieve this by providing interesting and relevant real-life examples covering a variety of topics, including the legal burden of proof, the use of statistics in the medical field, the ethics of data collection and data sleuthing, and the use (or misuse) of statistics in Supreme Court cases.

The work is composed of three sections. Part I is structured based on general statistical concepts. The authors discuss statistical tools, formulas, and theorems in an engaging and straightforward manner, using short stories and vignettes that even students beginning the study of statistics will enjoy reading. After introducing the subject in the first two chapters, Part I begins with Chapter 3, a discussion of probability theory and Bayes' theorem. This chapter is very useful for statistics students, as it describes Bayes' theorem several different ways: using mathematical formulas, $2 \times 2$ contingency tables, and in simple nonscientific language. The reader will appreciate the clear examples that illustrate the practical implications of Bayes' theorem. Readers from all educational backgrounds will certainly be able to learn from the misunderstandings and misapplications of Bayes' theorem highlighted in this section. The emphasis is on the conclusion that even experienced statisticians can and historically have fallen prey to these common mistakes. The authors underscore one's duty to think critically about generating, interpreting, and reporting probabilities. Many examples of misused statistics are taken from well-known legal cases, with one interesting example being a problematic conditional probability used by a Harvard law professor who advised the O.J. Simpson defense team. The chapter on probability theory also elaborates on sensitivity, specificity, and positive predictive value in the context of breast cancer screening.

Chapter 4 focuses on application areas touching on concepts such as causation, relative risk, odds ratios, cohort studies and again provides great context and easy reading for the beginning statistician. The chapter also reviews simple experiments and sample space in the context of engaging topics such as spread betting, parimutuel betting, gaming, risk, and point shaving in college basketball. The chapter concludes by examining, with compelling

[1] Data Analysis and Quality Assurance Branch, Division of Health Interview Statistics, National Center for Health Statistics, 3311 Toledo Road, Room 2329, Hyattsville, Maryland 20782. Email: wkirzinger@cdc.gov

attention, the consequences of framing data used to make decisions, the role psychology of risk behavior can play in decision making, and the importance of being able to assess the quality of the information used when making decisions.

Next, Chapter 5 is a cautionary tale on the topic of correlation, as it specifically focuses on correlational fallacies. The first warning is against providing a correlation value without showing the associated scatterplot. Several arguments are given to demonstrate the need for the graphical representation of correlation data. The authors effectively balance a review of basic concepts (biases, such as confirmation bias and detection bias) with information that may be new and interesting to readers at any stage of their statistical education (terminology such as 'apophenia', seeing patterns or connections in random data and 'pareidolia', when random stimuli are perceived as significant, such as in when one sees the Virgin Mary in a grilled cheese sandwich).

The next chapter also has a highly cautionary tone. The authors discuss how the phenomenon of regression toward the mean can lead to invalid reasoning and offer a quote by John von Neumann to caution the reader on overfitting observations when developing models: "With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk". Here the authors are making the point that a good fit to observations is not always the best model, as a model should be flexible in order to capture both systematic and unsystematic patterns. The authors refer to a regression analysis used during World War II to predict the accuracy of bombing; this example clearly demonstrates the effects of misinterpreting regression weights and the variability of a model depending on what variables are included.

Chapter 7 begins with a refresher on populations, samples, distributions, and the central limit theorem, and then progresses to a discussion on the "beauty of natural variation" and the importance of having an "appreciation of random processes". The authors posit that most people tend to underestimate the amount of variation that should be present in random data. For example, in the context of sports, it is tempting to try to explain randomness in performance as "pressure" felt by athletes, players being "in a slump", or a team "having chemistry". One important message in this chapter is that often randomness is misunderstood, and bad things can happen when you don't see randomness where you probably should (such as in the Bernie Madoff case, where investments consistently and unrealistically gave 12% returns without any variability) or when cause is attributed when none exists (effectiveness of medical treatment in clinical trials). Another important point made in this chapter is in the section about the pitfalls of software, stating specifically that software can provide lots of output, but that does not necessarily mean the output should be used. This point is useful for statistics students to be aware of when interpreting output. The authors state their view that open-source software is preferable over closed-source software packages, since closed-source software allows analyses to be conducted without an understanding of what is really happening. However, in reality, using open-source software may not be a reasonable expectation for all readers, especially for statistics students.

The last chapter of Part I is dedicated to the field of psychometrics. The authors' discussion of reliability and validity is clear and is supported by simple examples. The example used to explain Cronbach's alpha relates to the question of whether or not

criminal behavior is a central component of psychopathy. It is an interesting example and very effective for enhancing the readers' understanding of this coefficient.

Part II, on data presentation and interpretation, begins with a short chapter that emphasizes the importance of presentation to help uncover the story behind the data. A list of fourteen common mistakes that an analyst should be aware of and avoid when presenting data is given here. Subsequently, in Chapter 10, the authors address two frequent offenses: underreporting data and misreporting data. The authors take up a quote by Rudy Giuliani that reveals a common confusion between mortality and survival, which leads the authors to explain both of these concepts along with relative risk and absolute risk, and to revisit the importance of framing and providing context when presenting data. Finally, the authors advise readers to know the population surveyed and in particular understand who may have been uncounted in that population.

Chapter 11 is a surprisingly fun chapter discussing internal validity and the Bradford-Hill criteria for establishing causality, with the inclusion of an amusing story about R.A. Fisher. Apparently Fisher disagreed with Hill regarding a link between smoking and cancer. To mock Hill, Fisher wrote an elaborate proof linking apple importation with divorce rates. Next, the authors provide a good explanation of standardization along with a caution to "look under the hood" when making conclusions based on aggregated data. The topic of Chapter 13 is meta-analysis, a type of analysis that has become popular in recent years. The authors provide a thorough explanation of what it is, why it became popular, and of course offer some criticisms of the technique. The authors give supporting examples of problems that arise in the interpretation of meta-analyses and warn against unethically motivated meta-analyses. Wrapping up Part II, Chapter 14 effectively walks the reader through a Supreme Court death penalty case to present the troubling topic of "statistical sleuthing".

Part III explores experimental design and data-collection topics, including general background on types of experimental studies, ethical considerations, and the Federal Rules of Evidence (FRE). Chapter 15 provides an excellent discussion of clinical trials, and of course provides specific cautions a researcher should ask him- or herself when dealing with observational studies. The next chapter is a great reference for statistics and/or public health students, as it walks the reader through several major historical milestones in the development of ethical guidelines for human experimentation. The authors do an excellent job here and throughout the text to incorporate current events into their discussions. In this chapter, to which historical events (Nuremberg Code, National Research Act, Declaration of Helsinki, etc.) lend themselves heavily, they successfully bring more contemporary events into the conversation, such as the apology that former President Clinton gave while he was in office for the circumstances surrounding the Tuskegee Syphilis Study, which makes the text relevant and more interesting than just a history lesson and review. Finally, the section ends with a chapter devoted to issues of admissibility of evidence and expert testimony in court cases. Also an entertaining section, the authors discuss "junk science" and recent examples from news sources to which all readers will be able to relate. The discussion of the Freedom of Information Act (1998) and the Data Quality Act (2001) along with their impacts on the data environment are thought-provoking, helping the reader to recognize recent changes to the environment due to the legislation and to

anticipate changes that will continue as more and more data are made available and new ways to look at these data are realized.

Throughout the text, the authors have provided a plethora of additional resources, with excerpts from court cases, appropriate quotes at the beginning of each chapter, and extensive notes at the end of each chapter. I recommend reading the notes, as some notes are unexpectedly quite humorous and some even contain jokes.

Overall, this book stands out as a unique text that combines a review of mathematical theorems and formulas, guidance on how to be sharp when using or interpreting statistics, and the impacts (often negative) that can happen (and have happened) when analyses are conducted carelessly. All of these components come together effectively to raise the reader's awareness of ethics in creating and interpreting statistics, and ultimately help the reader become a more astute and ethically responsible analyst.

VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 571–573, http://dx.doi.org/10.2478/JOS-2014-0037

# Book Review

*Joseph W. Sakshaug*[1]

**Raymond L. Chambers and Robert G. Clark.** *An Introduction to Model-Based Survey Sampling with Applications*. 2012 New York, USA: Oxford University Press Inc., ISBN 978-0-19-856662-5, 265 pp., £41.99.

This book introduces readers to the fundamental concepts of the model-based approach to survey sampling. For those unfamiliar with the model-based approach, it is worthwhile to note that there are distinct differences between this approach and the more commonly taught design-based approach. In the design-based paradigm, survey inferences are obtained from estimators that are based on the assumption that samples are drawn repeatedly from a fixed and finite population. These design-based estimators lack any assumption regarding the structure of the population under study and therefore can be correctly applied in any setting. In contrast, the model-based approach to inference uses estimators that do not rely on repeated sampling properties, but rather take into account the properties of a specific population through the use of auxiliary information. The population is summarized through the use of a model and responses are assumed to be generated by a stochastic process. Model-based approaches to sampling and inference have received adequate attention in the literature. Some notable citations include Valliant (2009), Valliant et al. (2000), Brewer (1963), and Royall (1970). The book *An Introduction to Model-Based Survey Sampling with Applications* by Raymond L. Chambers and Robert G. Clark is a useful addition to this literature and should satisfy readers who are interesting in acquiring the theoretical and practical knowledge needed to carry out the model-based approach.

The book is divided into three parts. The first part (Chapters 1–7) covers the fundamental aspects of model-based survey inference. The second part (Chapters 8–10) introduces model-based survey methods that are robust to incorrectly specified models and outliers. Finally, the third part (Chapters 11–17) covers several modern applications of model-based survey inference.

Chapter 1 provides an introduction to survey sampling and introduces the relevant notation that is used throughout the book. Chapter 2 introduces the model-based approach and highlights important differences between this approach and other approaches to survey sampling. Chapter 3 considers the simplest possible model in which the survey population has no auxiliary variables (or the auxiliary variables are unrelated to the survey variables of interest). Several important questions are addressed, including how large the sample should be and how to carry out the sampling process. Chapter 4 builds on the

[1] Institute for Employment Research, Regensburger Straße 104, 90478 Nuremberg, Germany. Email: joe.sakshaug@iab.de

previous chapter by extending the model to account for stratified populations. Several stratification schemes are considered, including proportional and optimal allocation, equal aggregate size stratification, and multivariate stratification. Chapter 5 covers linear regression models for populations with a single auxiliary variable. Combining regression and stratification is also discussed. Chapter 6 deals with clustered populations in which the sampling process is carried out in two stages. The clustered population model is introduced for different optimal designs, including fixed sample size and fixed cost. Chapter 7 concludes the first part of the book by demonstrating how the population models presented in the earlier chapters are special cases of the general linear population model.

Chapter 8 demonstrates the robustness of the model-based approach. Specifically, the chapter shows that, under robust sampling designs, approximately unbiased inferences can be obtained even when the assumed model is incorrectly specified. Chapter 9 extends the idea of robustness to variance estimation under model misspecification. Chapter 10 describes strategies for making survey sampling techniques robust to extreme observations, or outliers. Nonparametric regression approaches are among those considered. Practical challenges of applying these outlier robust estimators are discussed at the end.

Chapter 11 departs from the linear population model and focuses on inferences based on non-linear population parameters, including population medians, quantiles, and ratios of two population means. Chapter 12 covers variance estimation for these complex statistics. Various techniques are considered, including random groups, balanced repeated replication, jackknife, and bootstrapping. Chapter 13 departs from the assumption used in previous chapters that there is only one survey variable of interest and deals with the reality that most surveys are multipurpose in nature and collect many $Y$ variables. Switching from a univariate to a multivariate $Y$ poses several design and estimation issues which are addressed here. Chapter 14 covers inference for population units belonging to particular domains. Situations in which domain membership is either known or unknown prior to data collection are both considered. Chapter 15 provides a thorough treatment of inference for small areas. Unit-level models for small areas are the focus of this chapter, which covers various methods including synthetic methods, methods based on random area effects, direct prediction methods, and the use of generalized linear mixed models. Chapter 16 presents methods for obtaining model-based inference about distributions and quantiles. These methods are considered under various survey designs, including stratification and clustering. An application of these methods is provided using data from a large-scale business survey. Lastly, Chapter 17 deals with the use of transformations to achieve linearity. A back transformation approach and a model calibration approach are described and empirical results of these approaches are provided.

In summary, the book covers a wide range of topics devoted to model-based survey sampling. I would highly recommend it for students and applied survey statisticians alike. I was particularly impressed with the book's ability to integrate both theoretical and applied concepts. While reading the book, I got a strong sense that it was written with the practitioner in mind. The book covers many practical survey applications, including an extensive chapter on small area estimation which is a topic growing in importance in survey research. Furthermore, the authors do a nice job of accommodating readers who may be less familiar with matrix algebra by waiting until Chapter 7 before introducing this

notation. If there is anything to criticize about the book, it would be the lack of a chapter devoted to longitudinal surveys which pose many design and estimation issues. Such surveys often have a rich amount of auxiliary information collected from prior waves and it would have been interesting to know how to utilize these data using model-based methods. Nevertheless, whether you are brand new to the topic of model-based survey sampling or are a seasoned user of such methods, this book should serve as a useful reference.

**References**

Brewer, K. 1963. "Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process." *Australian Journal of Statistics* 5: 93–105.

Royall, R.M. 1970. "On Finite Population Sampling Theory under Certain Linear Regression Models." *Biometrika* 57(2): 377–387.

Valliant, R. 2009. "Model-Based Prediction of Finite Population Totals." In *Chapter 13 of Handbook of Statistics 29: Sample Surveys: Design, Methods, and Applications*, edited by D. Pfeffermann and C.R. Rao,. Amsterdam: North Holland.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference*. New York: Wiley.