Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 471–476

Letter to the Editor

Letters to the Editor will be confined to discussion of articles which have appeared in the Journal of Official Statistics and of important issues facing the statistical community.

Revisiting the Multipurpose Property of Sampling Weights

The recent article by Professor R.J.A. Little (Little 2012) includes a discussion of alternative basic philosophies of official statistics production. In this letter, we wish to bring to the attention of JOS readers another, related fundamental property of significance for official statistics production, which we believe is related to the matter that Prof. Little discusses.

Use of sampling weights is a feature that probably distinguishes survey sampling most from other statistical disciplines. In survey practice, statisticians have traditionally called for them to satisfy the so-called *multipurpose property* (Särndal 2007), that is, that a single set of sampling weights is used to estimate all population variables in a multipurpose survey.

Another key concept in official statistics production is auxiliary information. It occurs at different stages: the sampling design (Cochran 1977), the construction of estimators (Särndal et al. 1992), the treatment of nonresponse (Särndal and Lundström 2005), the imputation methods (Haziza 2009), to name perhaps the most noteworthy. Auxiliary information in statistical offices is nowadays abundant, available, up-to-date and of good quality for statistical purposes.

We contend that this increasing availability of auxiliary information invites us to consider putting aside the multipurpose property. We reason as follows. From a purely theoretical standpoint, there is no reasoning that supports the multipurpose property. Moreover, adhering to methodological rigour in sampling weights construction, one can easily find reasons not to have a single set of weights. Let us consider, for instance, nonresponse treatment. Reweighting for nonresponse (see e.g., Särndal and Lundström 2005; Bethlehem et al. 2011) is an elaborate technique where either calibrating against benchmark auxiliary information or modelling response propensity (also using auxiliary information) assists in the weight adjustment and bias reduction. Regarding calibrating, to take a specific example, the following statement by Ranalli (2008) is enlightening in this respect:

> The calibration approach of Deville and Särndal (1992) has been referred as to be "model-free" (Särndal 2007), as opposed to regression estimation in which an assisting model has to be specified to conduct estimation. We believe that model-free, in this case, refers to being free from an *explicit* [original italics] modelling procedure. In fact, the results reported here show that calibration, although developed in a purely design based framework, *implicitly assumes a linear relationship between all the survey variables and the auxiliary ones* [our italics].

Thus, if the auxiliary information needed to adequately deal with the nonresponse differs between different variables of interest, why would one not use the correspondingly different sets of sampling weights for each of them? Furthermore, if accepting different sets of sampling weights in a multipurpose survey, why not use more accurate techniques, such as, for instance, model calibration (Wu and Sitter 2001; Wu 2003; Montanari and Ranalli 2003, 2005) in the construction of estimators? Moreover, what if we use model-assisted techniques with non-linear models rendering the concept of sampling weight itself surpassed by a more complex, although possibly more accurate, notion of (non-linear) sampling estimator (Lehtonen and Veijanen 1998)? Accuracy is clearly an argument in favour of having several sets of sampling weights.

In the present multidimensional reading of data quality, not using a single set of weights can also be viewed as a possible cost reduction in terms of sampling sizes: if for a given sample size n and its corresponding cost *c(n)*, lower variances, say, $V_{\text{no multi}} < V_{\text{multi}}$ can be achieved by dropping the multipurpose property, why not think of keeping the same accuracy V_{multi} but reducing the sampling size $n' < n$ and the corresponding cost $c(n') < c(n)$? Another example going in the same direction stems from the use of multiple frames with Hartley-Fuller-Burmeister-type estimators. In this case, cost reduction because of the use of multiple frames is also present, but is often disregarded because of the multipurpose property (Lohr 2009).

In a more general discourse, dropping the multipurpose property can be viewed as a chance to use model-based techniques in the construction of sampling estimators. It gives the statistician the opportunity to resort to the vast field of classical inference statistical techniques *without crossing the red line between the design-based and model-based approaches* (see e.g., Smith 1994 and references therein for a detailed discussion). As prominent examples, let us cite model-assisted estimation (Särndal et al. 1992) and model calibration (Wu and Sitter 2001; Wu 2003; Montanari and Ranalli 2003, 2005): the estimators obtained thereby are (approximately) design-unbiased, being protected against model-breakdowns. Typically, they are also more accurate than those not using these statistical-modelling assisting techniques. But the door is open: why not use more general techniques, for instance, geostatistical techniques or time-series modelling, in the same fashion?

On the other hand, from a practical point of view, we can suggest several reasons supporting the multipurpose property, namely, (i) sampling weights interpreted in a sense of representativity, which apparently reinforces the multipurpose property; (ii) the numerical consistency among all output tables in multipurpose surveys; and (iii) the concerns about transparency in data dissemination.

As we see it, the representativity interpretation and the multipurpose property are strongly reinforcing each other in survey practice: if a sampling weight of a sample unit k is interpreted as the number of population units represented by k, it is natural to have just a single set of sampling weights in a survey; and vice versa, if only a single set of sampling weights is to be accepted in a survey, it is natural to interpret them as a measure of the representativity of each sample unit. In our opinion, the representativity view has already been challenged by consequently adopting the theoretically correct interpretation of a sampling weight ω_{ks} in a linear estimator $\hat{Y}_U = \sum_{k \in s} \omega_{ks} y_k$ as a multiplicative factor of the variable value y_k of unit k in the sample s when estimating the population total $\sum_{k \in U} y_k$.

Table 1. Estimates of population units exhibiting and not exhibiting habit A

Habit A	Male	Female	Total
Present	$\hat{Y}_m^A$	$\hat{Y}_f^A$	$\hat{Y}_m^A + \hat{Y}_f^A$
Absent	$\hat{Y}_m^{\neg A}$	$\hat{Y}_f^{\neg A}$	$\hat{Y}_m^{\neg A} + \hat{Y}_f^{\neg A}$
Total	$\hat{Y}_m^A + \hat{Y}_m^{\neg A}$	$\hat{Y}_f^A + \hat{Y}_f^{\neg A}$	.

Notice that negative weights and weights $\omega_k < 1$ are indisputable in this interpretation. The possibility of having negative sampling weights underlines the essential difference between the design-based and the model-based approaches to inference. Without going into detail beyond the scope of this letter, take as an example the issue whether sampling weights must be used in analysing survey data with heteroskedastic linear regression models or not (see e.g., Little 2004 and references therein). Choosing a model variance $\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_N^2\}$, with $\sigma_k^2 \propto \omega_{ks}$, is clearly impossible in the case of at least one negative sampling weight $\omega_{k*s} < 0$. We believe that this is a direct consequence of the irreconcilable difference between the two approaches (Smith 1994). In our opinion, official statistics must remain on the safe side of design-unbiasedness, although model-assisted. However, the notion of a sampling weight as a measure of the representativity of the associated unit should be exorcised from survey sampling (Kruskal and Mosteller 1979a,b,c, 1980).

More importantly, numerical consistency among all output tables in a multipurpose survey is a concern. It is indeed a very serious concern, giving justification to the multipurpose property: it ensures numerical consistency. Let us consider an example in a health survey where the presence or absence of two habits A and B is measured in the population. Let us accept that different sets of weights $\{\omega_k^A\}$ and $\{\omega_k^B\}$ are used because different auxiliary variables have been used in the calibrating stage. Suppose that the results are demanded broken down by sex. This is usually presented in the form of contingency tables as in Tables 1 and 2.

Here the issue becomes apparent: rarely, under the assumed working hypotheses, will the pairs of marginal estimated sex totals $(\hat{Y}_m^A + \hat{Y}_m^{\neg A}, \hat{Y}_m^B + \hat{Y}_m^{\neg B})$ and $(\hat{Y}_f^A + \hat{Y}_f^{\neg A}, \hat{Y}_f^B + \hat{Y}_f^{\neg B})$ coincide respectively. This is the consistency alluded to above. In the sphere of official statistics, this can be very difficult to accept from the point of view of users of the statistics: how is it possible that we can be faced with different male and female counts as a result of estimating different variables? Should these counts not be the same irrespective of the estimated variable?

Thirdly, transparency in official statistics entails anonymised microdata released to final users in such a way that they can reproduce almost any published estimate. The case of

Table 2. Estimates of population units exhibiting and not exhibiting habit B

Habit B	Male	Female	Total
Present	$\hat{Y}_m^B$	$\hat{Y}_f^B$	$\hat{Y}_m^B + \hat{Y}_f^B$
Absent	$\hat{Y}_m^{\neg B}$	$\hat{Y}_f^{\neg B}$	$\hat{Y}_m^{\neg B} + \hat{Y}_f^{\neg B}$
Total	$\hat{Y}_m^B + \hat{Y}_m^{\neg B}$	$\hat{Y}_f^B + \hat{Y}_f^{\neg B}$	.

several sets of sampling weights, or even of estimators assisted with possibly non-linear models, renders this task much more complex, to the point of even preventing the user from computing any further estimate not contained in published releases. Such a lack of transparency could damage official statistics.

Any attempt to drop the multipurpose property in official statistics production must in our opinion tackle all these questions. Firstly, the question regarding the interpretation of sampling weights has already been settled in the methodological arena, but the idea of representativity must be carefully dealt with when disseminating official statistics. Secondly, the numerical consistency of any planned or unplanned table must be guaranteed. That is, there must exist a methodological solution to the numerically consistent estimation of population quantities arranged in almost any cross-tabulation of variables. This is the case both for those tabulations contained in the survey production plan and for those not included but later called for. In this sense, it seems nowadays advisable to move the focus of the problem of estimation in a finite population from its traditional univariate setting (see e.g., Hanurav 1966) to a more general and realistic definition: given a finite population U of known size N and composed of identifiable units with variable values $\mathbf{y}_k$, the objective is to produce numerically consistent estimates for any planned or unplanned set of tables of population quantities $f_p(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, $p = 1, \ldots, P$. In this regard, let us cite the repeated weighting technique (Kroese et al. 2000). Repeated weighting resorts to an extensive use of calibrating provided "one is willing to abandon the common practice of using one set of [. . .] weights [. . .]" (Boonstra et al. 2003). Thus, important steps have already been taken in this direction, although more work needs to be done to reach a satisfactory solution.

To sum up, dropping the multipurpose property arises as an attractive invitation to use statistical models and more general techniques in assisting the construction of survey estimators within the design-based framework. In official statistics, this would pave the way not only for the stimulation of novel ideas on how to adapt these techniques in the construction of estimators, but also the inclusion of existing methods in the daily production of statistical offices in a general fashion. However, we also believe that in official statistics any step in this direction must guarantee the accessibility and clarity of the published information, which must be released in an understandable, suitable and convenient manner to the final user. In current user-oriented statistical systems, we are convinced that some pedagogical actions regarding the chosen methodology and dissemination policies should be considered in order to take into account users' needs and to guarantee maximum transparency.

References

Bethlehem, J., Cobben, F., and Schouten, B. (2011). Handbook of Nonresponse in Household Surveys. Hoboken, NJ: Wiley.

Boonstra, H.J.H., van der Brakel, J.A., Knottnerus, P., Nieuwenbroek, N.J., and Renssen, R.H. (2003). A Strategy to Obtain Consistency Among Tables of Survey Estimates. Workpackage 7 of DACSEIS project. Available from: http://www.dacseis.de. Accessed November, 5th, 2012.

Cochran, W.G. (1977). Sampling Techniques, (3rd ed.). New York: Wiley.

Deville, J.C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Hanurav, T.V. (1966). Some Aspects of Unified Sampling Theory. Sankhya A, 28, 175–204.

Haziza, D. (2009). Imputation and Inference in the Presence of Missing Data. In Sample Surveys: Design, Methods and Applications, D. Pfefferman and C.R. Rao (eds). Amsterdam: North-Holland.

Kroese, B., Renssen, R.H., and Trijssenaar, M. (2000). Weighting or Imputation: Constructing a Consistent Set of Estimates Based on Data from Different Sources, In Statistics Netherlands (2000), Netherlands Official Statistics, 15, special issue on Integrating administrative registers and household surveys. Voorburg/Heerlen: Statistics Netherlands.

Kruskal, W. and Mosteller, F. (1979a). Representative Sampling. I: Non-scientific Literature. International Statistical Review, 47, 13–24.

Kruskal, W. and Mosteller, F. (1979b). Representative Sampling. II: Scientific Literature, Excluding Statistics. International Statistical Review, 47, 111–127.

Kruskal, W. and Mosteller, F. (1979c). Representative Sampling. III: The Current Statistical Literature. International Statistical Review, 47, 245–265.

Kruskal, W. and Mosteller, F. (1980). Representative Sampling, IV: The History of the Concept in Statistics, 1895–1939. International Statistical Review, 48, 169–195.

Lehtonen, R. and Veijanen, A. (1998). Logistic Generalized Regression Estimators. Survey Methodology, 24, 51–55.

Little, R.J.A. (2004). To Model or not to Model? Competing Modes of Inference for Finite Population Sampling. Journal of the American Statistical Association, 99, 546–556.

Little, R.J.A. (2012). Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics. Journal of Official Statistics, 28, 309–334.

Lohr, S. (2009). Multiple-frame Surveys. In Sample Surveys: Design, Methods and Applications, D. Pfefferman and C.R. Rao (eds). Amsterdam: North-Holland.

Montanari, G.E. and Ranalli, M.G. (2003). On Calibration Methods for the Design-based Finite Population Inferences. Bulletin of the International Statistical Institute, 54th session, vol. LX, contributed papers, book 2, 81–82.

Montanari, G.E. and Ranalli, M.G. (2005). Nonparametric Model-calibration Estimation in Survey Sampling. Journal of the American Statistical Association, 100, 1429–1442.

Ranalli, M.G. (2008). Recent developments in calibration estimation. Proc. XLIV Meeting of the Italian Statistical Society, pages 355–362.

Särndal, C.-E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, 33, 99–119.

Särndal, C.-E. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Chichester: Wiley.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Sampling. New York: Springer.

Smith, T.M.F. (1994). Sample Surveys 1975–1990: An Age of Reconciliation? International Statistical Review, 62, 5–19.

Wu, C. (2003). Optimal Calibration Estimators in Survey Sampling. Biometrika, 90, 937–951.

Wu, C. and Sitter, R.R. (2001). A Model-calibration Approach to Using Complete Auxiliary Information from Survey Data. Journal of the American Statistical Association, 96, 185–193.

D. Salgado[1]
C. Pérez-Arriero[2]
M. Herrador[2]
I. Arbués[1]

[1] National Statistical Institute, D.G. Methodology,
Quality and Information and Communications Technologies,
Paseo de la Castellana, 28071 Madrid, Spain
E-mail: david.salgado.fernandez@ine.es;
ignacio.arbues.lombardia@ine.es

[2] National Statistical Institute, S.G. Sampling and Data Collection,
Paseo de la Castellana, 28071 Madrid, Spain
E-mail: carlos.perez.arriero@ine.es;
monserrat.herrador.cansado@ine.es

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 477–499

Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection

James Wagner[1], *Brady T. West*[1], *Nicole Kirgis*[1], *James M. Lepkowski*[1], *William G. Axinn*[1], *and Shonda Kruger Ndiaye*[1]

In many surveys there is a great deal of uncertainty about assumptions regarding key design parameters. This leads to uncertainty about the cost and error structures of the surveys. Responsive survey designs use indicators of potential survey error to determine when design changes should be made on an ongoing basis during data collection. These changes are meant to minimize total survey error. They are made during the field period as updated estimates of proxy indicators for the various sources of error become available. In this article we illustrate responsive design in a large continuous data collection: the 2006–2010 U.S. National Survey of Family Growth. We describe three paradata-guided interventions designed to improve survey quality: case prioritization, "screener week," and sample balance. Our analyses demonstrate that these interventions systematically alter interviewer behavior, creating beneficial effects on both efficiency and proxy measures of the risk of nonresponse bias, such as variation in subgroup response rates.

Key words: Nonresponse; paradata; responsive design; interviewing.

1. Introduction

Survey data collection is filled with uncertainty. This is particularly true for large, face-to-face surveys that rely on interviewers to make most of the decisions about how to achieve contact with (and cooperation from) sampled units. For these surveys, many aspects of the process can only be quantified with probability statements. Commonly used sampling frames (e.g., address lists) may contain many ineligible units. Often, our ability to predict eligibility is weak. Interviewers vary in their ability to find the best times to call on households to maximize contact rates and in their ability to obtain cooperation once contact has been made. Overall, our ability to predict the likelihood of either contact or cooperation is also often weak. Unfortunately, each of these uncertainties interferes with our ability to control the cost, timeliness, and error properties of survey data. This article illustrates the application of a new generation of methodological tools for addressing these uncertainties.

Pre-specified survey designs are not well suited to highly uncertain settings. Any departure from the expectations of the design may lead to a failure to meet some or all of

[1] University of Michigan, USA, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48105, U.S.A. Corresponding author: James Wagner, telephone: 734-647-5600; facsimile: 734-764-8263. Email: jameswag@isr.umich.edu, Brady T. West, Email: bwest@umich.edu, Nicole Kirgis, Email: nkirgis@isr.umich.edu, James M. Lepkowski, Email: jimlep@isr.umich.edu, William G. Axinn, Email: baxinn@isr.umich.edu and Shonda Kruger-Ndiaye, Email: shondak@isr.umich.edu.

the targeted outcomes. These failures frequently include both cost and error failures (Groves 1989), leading to costs that run higher than budgets or errors that are larger than expected. For example, if more effort to complete interviews is required than initially expected, then fewer interviews may be completed and the sampling error of estimates will increase.

Responsive survey designs attempt to address these issues by gathering information about the survey data collection process and using these data to compute indicators that decrease this uncertainty (Groves and Heeringa 2006). These data are used to make decisions about altering design features *during* the survey field work. Groves and Heeringa define five steps for these responsive designs:

1. Pre-identify a set of design features potentially affecting costs and errors of survey statistics;
2. Identify a set of indicators of the cost and error properties of those features;
3. Monitor those indicators in initial phases of data collection;
4. Alter the active features of the survey in subsequent phases based on cost/error tradeoff decision rules; and
5. Combine data from the separate design phases into a single estimator.

These responsive designs rely upon indicators that are built from the available data. Frequently, sampling frames include auxiliary variables that are only weakly predictive of important outcomes of the survey process, including indicators of response and measures on key survey variables collected from respondents. For this reason, researchers have turned to *paradata*, or survey process data (Couper 1998; Couper and Lyberg 2005), as an additional source of auxiliary data. These data may include records of call attempts; interviewer observations about the neighborhood, sampled unit, or sampled person; and timing data from computerized instruments. Responsive designs incorporating paradata to guide design decisions during the field work have the potential to reduce the costs and errors associated with survey data collection. Survey methodology has made advances in the use of paradata (Kreuter et al. 2010; Durrant et al. 2011), but there is very little published research evaluating responsive design tools.

To advance this area of science, this article reviews several responsive design features of a large, face-to-face demographic survey – the 2006–2010 U.S. National Survey of Family Growth (NSFG). The NSFG is sponsored by the National Center for Health Statistics and was conducted by the University of Michigan's Survey Research Center. The responsive design tools described in this article are built upon paradata that have been tailored to the demographic data collected in the NSFG. They are meant to increase our control over the costs, timeliness, and quality of the collected data. Conceptually, responsive designs can be understood from a total survey error perspective, and include monitoring and control of other error sources. We focus on the use of responsive design principles to control the risk of nonresponse bias as a crucial dimension of total survey error.

In most situations, researchers do not have direct information about nonresponse bias. Surveys that do have "gold standard" data or true values available on selected variables for an entire sample are usually performed for methodological – as opposed to substantive – research purposes. Therefore, in order to control the risk of nonresponse bias in a production environment, proxy indicators of nonresponse bias are needed. For example, the NSFG sample includes multiple subgroups defined by the cross-classification of age, race, ethnicity and gender. A recent review of specialized studies of nonresponse found that the variation in response rates across groups defined by these sorts of demographic variables was not predictive of nonresponse biases (Peytcheva and Groves 2009). In the case of the NSFG, however, these demographic factors are predictive of key survey variables (Martinez et al. 2012). To the extent that these characteristics are predictive of the key statistics measured by the NSFG, large variance in the response rates across these groups is an indicator for potential nonresponse biases (Groves and Heeringa 2006). In another NSFG-specific example, the NSFG asks interviewers to make observations about the sampled persons. These observations are highly correlated with several of the key statistics produced by the survey (Groves et al. 2009; West 2013). These proxy indicators may also be used as indicators for the risk of nonresponse bias. The assumption here is that once we have equalized response rates across subgroups defined by these proxy indicators, the nonresponders and responders within each subgroup will not differ with respect to the survey variables being collected. In other words, we assume that the nonrespondents are "Missing at Random" (Little and Rubin 2002), conditional upon the characteristics used to balance the sample. In this article, we discuss attempts to use such proxy indicators in a responsive design framework to control the risk of nonresponse bias in the NSFG.

In order to make effective use of these proxy indicators, the NSFG design called for centralized direction of data collection effort. We believe that this is a unique feature of the NSFG design. In contrast, most large-scale face-to-face surveys leave the prioritization of effort to the interviewer. The interviewers determine which cases to call and when. Many surveys provide general guidelines to interviewers in this regard. For example, the European Social Survey (ESS) guidelines suggest that a minimum of four calls be placed to each household and that these calls should be spread over different times of day and days of the week, with at least one call in the evening and one on the weekend (Stoop et al. 2010). Others have used prioritization schemes developed prior to fielding the survey in order to increase these sorts of proxy indicators for nonresponse bias (Peytchev et al. 2010). The NSFG is unique in that interviewer behaviors are at times guided by centralized decisions of the managers based on the analysis of paradata. These altered behaviors lead to greater balance on the proxy indicators for nonresponse bias, and we illustrate this result in this article. The special centralized design of the 2006–2010 NSFG gives us a distinctive opportunity to investigate responsive design tools that are intended to alter interviewer behavior in response to incoming paradata during field data collection.

After describing relevant aspects of the NSFG design, we investigate three types of paradata-driven responsive design interventions. The first (Section 3.1) is a set of interventions that was designed to determine our ability to alter interviewer behavior and had specific objectives with relation to the cost and error properties of the data. The second set of interventions (Section 3.2) was aimed at identifying eligible persons earlier in the field period than might have otherwise occurred, thereby procuring data that are informative about the risk of nonresponse bias as quickly as possible in order to enable better control over this error source. The third type of intervention (Section 3.3) uses the variation in subgroup response rates as a proxy indicator for the risk of nonresponse bias. Investigations of these three types of responsive design interventions provide a crucial advance in the tool set for implementing responsive designs and for using such designs to reduce the uncertainty in survey data collection.

2. NSFG Management Framework

The interventions reported here were developed in the context of a survey management framework that used paradata to guide decision making about survey design features. These responsive design interventions were implemented in the NSFG, which collects data from an ongoing, national, cross-sectional, multistage area probability sample of households in the United States (Groves et al. 2009). In each sampled household, interviewers completed a screening interview by collecting a roster of household members. One person aged 15–44 was selected at random from the age-eligible persons within the household. The interviewer then sought a 60–80 min interview from the selected person. The interview involved the administration of a computer-assisted personal interview (CAPI) questionnaire that contained questions on the respondent's sexual and fertility experiences. More sensitive items (e.g., risk behaviors for HIV) were administered using an audio computer-assisted self-interview (ACASI) application on the interviewer's laptop. A token of appreciation ($40) was paid to respondents upon completion of the main interview.

Each year of data collection for the NSFG consisted of four replicate samples yielding 5,500 completed interviews per year on average. Replicate samples were introduced at the beginning of each quarter. The full data collection period for a year lasted 48 weeks (four 12-week quarters), with four weeks for end-of-year holidays and mid-year training of new interviewers. New interviewers were introduced as part of a rotation of the primary sampling units (PSUs) each year. During any given year, the sample consisted of 33 PSUs and about 38 interviewers across them. The American Community Survey (ACS) uses a similar continuous measurement design to produce timely, frequent, and high-quality data in place of the previous United States Census long form (National Research Council 2007).

Unlike many surveys, the NSFG used a two-phase or double sample process to address the problem of nonresponse. Each 12-week quarter was divided into a 10-week period (Phase 1) and a 2-week period (Phase 2). During Phase 1, interviewers were assigned an average of about 120 sample addresses to screen and interview. At the end of ten weeks, some addresses remained outstanding, that is, they had not yet been finalized as an interview, a refusal, a non-sample case, or some other final disposition. A sample of about one-third of the outstanding addresses was selected and sent back to the interviewers. This sample was selected as a stratified random sample of cases. The strata were defined by eligibility status (eligible or unknown) and tertiles of the estimated probability of response. The sampling rate was chosen based on management experience from Cycle 6 of the NSFG. The sampling rate effectively triples the effort on the selected cases (since the interviewers work a constant 30 per week). This sampling rate allowed us to meet targeted response rates while controlling costs. More information on the second phase sample design is available in the NSFG Series 2 report (Lepkowski et al. 2010). The interviewers then had two weeks at the same weekly effort level to complete interviews with as many of the double sample addresses as possible. The NSFG was also able to provide a higher token of appreciation ($80 for adult respondents) during Phase 2. Later, during data processing, the Phase 1 and 2 samples were combined in the final survey data set. Weighted response rates were computed to account for the additional interviews obtained from the Phase 2 respondents. Additional details about the design and operations of the Continuous NSFG, including detailed descriptions of paradata collected, can be found in Groves et al. (2009).

The NSFG used a management decision model to guide interventions in a responsive design framework. The model has three input elements that management can manipulate (Effort, Materials, and Quality of Materials), and three broadly defined outcomes (Interviews, Cost, and Sample Characteristics). All inputs and outcomes are monitored through the processing and analysis of paradata. *Effort* refers to survey features such as number of calls, whether in total or within a time frame (e.g., per day); proportion of hours worked during "peak" calling times; number of interviewers working on the study; and hours worked by interviewers. *Materials* include active cases remaining to be screened in the field data collection; cases with identified eligible persons who have yet to be interviewed; or the number of cases not attempted as of a fixed date in the data collection. *Quality of Materials* includes such measures as the proportion of remaining cases that have ever resisted an interview attempt through refusal or other actions indicating a reluctance to participate; the proportion of active cases in a locked building; or the mean of the estimated response propensities for each active case.

Three primary outcomes were of interest to NSFG managers. *Interviews* were measured by such outcomes as the number of interviews completed by day or response rates by day, week, or other time period. *Cost* was measured by hours required to complete an interview or expenditure to travel to a sample location. *Sample characteristics* included measures of how well the set of respondents matched the characteristics of the sample (for example age, sex, race, ethnicity, or interviewer observations about relevant household characteristics), and whether estimates from the observed data converged after a specified number of calls. The overall production model asserts that the number and cost of interviews as well as the characteristics of the sample are a function of the field effort applied and the current state of the active sample (materials and the quality of the materials). This model was applied to the dynamic process of daily data collection.

The elements in the production model were monitored through a "dashboard" consisting of graphs portraying the status of various measures for each of these elements (see Groves et al. 2009). The graphs were updated daily throughout the data collection period. The dashboard served as a central feature in the management process, allowing for monitoring of all elements in the model and guiding management decisions about how and when to intervene.

3. Three Paradata-Driven Interventions

The 2006–2010 NSFG implemented three different types of management interventions in the responsive design framework: *case prioritization*, *screener week*, and *sample balance*. Each of the three types of interventions had different objectives. The *case prioritization* intervention was aimed at checking whether the central office could influence field outcomes by requesting that particular cases be prioritized by the interviewers. If this prioritization proved to be successful, then the second objective was to determine what impact these case prioritizations could have on the composition of the final set of respondents. *Screener week* sought to shift the emphasis of field work in such a way that eligible persons (and proxy indicators of nonresponse bias for those persons) would be identified as early as possible. Since the screening interview also generates valuable data about the demographic characteristics of sampled persons, screener week improved our ability to balance the sample. The "*sample balance"* intervention sought to minimize the risk of nonresponse bias by endeavoring to have the set of respondents match the characteristics of the original sample (including nonresponders) along key dimensions, such as race, ethnicity, sex, and age. We describe each of these interventions in detail and provide examples of their implementation in the following subsections.

3.1. Case Prioritization: Paradata-Guided Randomized Experiments

The idea of embedding randomized experiments in an ongoing survey data collection is not new. Possible reductions in survey errors from adaptively embedding randomized experiments in survey designs have been discussed previously by Fienberg and Tanur (1988, 1989). The first set of interventions that we describe here involved assigning a random subset of active cases with specific characteristics to receive higher priority from the interviewers. NSFG managers targeted these cases for intervention in response to trends in selected elements of the production model that indicated possibly increased risks of survey errors. This type of intervention was replicated 16 times on different quarterly samples.

The case prioritization interventions involved late-quarter targeting of specific types of sampled housing units or persons (if already screened) to increase the number of calls to these specific groups. The first objective of these experiments was to determine whether interviewers would respond to a request to prioritize particular cases. While one can assume that interviewers will do what is requested of them, we knew of no research examining the outcomes of such requests in a field data collection. It was hoped that if the calls were increased, then response rates for the targeted cases would rise, relative to those of other cases. In this section, we focus our analysis on determining whether these types of interventions can have an impact on effort and, subsequently, on response rates for the targeted subgroups. If these interventions are successful, they may be an important tool in reducing interviewer variance and controlling the composition of the set of respondents. In subsequent sections, we will consider how these interventions might be used to improve survey outcomes relative to the risk of nonresponse bias.

Each of the experiments also had a secondary objective related to reduction of survey errors. Table 1 lists all 16 of the randomized experiments and describes the secondary objectives for targeting each of the specified subgroups. In some cases, the objective was to improve overall response rates. In other cases, the objective was to evaluate the utility of data available on the sampling frame. In still other cases, the objective was to bring the distribution of the characteristics of the respondents closer to the distribution of the characteristics of the original sample.

All 16 of these interventions were randomized experiments in which one half of the target cases was assigned to the intervention and one half remained as a control group.

Table 1. *16 randomized interventions, 2006–2010 Continuous NSFG*

				Sample size	
Intervention Type[a]	Description	Objective	Length (Days)	Intervention	Control
EXT1	Active screener addresses matched with Experian data indicating household eligibility (at least one person age 15–44 in household)	Evaluate the utility of commercially available data. Evaluate whether prioritizing likely eligible persons leads to better sample balance.	11	759	755
EXT2	Active screener addresses matched with Experian data indicating household not eligible (no person age 15–44 in household)		11	637	624
EXT3	Active screener addresses with no Experian match (indeterminate household eligibility)		11	430	434
INT1	Active screener addresses with high predicted probability of eligibility (based on NSFG paradata)	Determine whether prioritizing likely eligible persons leads to better sample balance	13	204	165
INT2	Active main addresses with high predicted probability of response (based on NSFG paradata), no children, and high predicted probability of eligibility (based on NSFG paradata)		14	115	109[b]
INT3	Active screener addresses with high predicted probability of response (based on NSFG paradata), no children, and high predicted probability of eligibility (based on NSFG paradata)		14	146	146
INT4	Active main addresses with high base weights and large or medium predicted probabilities of response (based on NSFG paradata)	Determine whether it is possible to improve response rates by prioritizing cases with relatively high weights.	8	100	88[b]
INT5	Active screener addresses with high base weights and larger or medium predicted probabilities of response (based on NSFG paradata)		8	133	133

Table 1. Continued

Intervention Type[a]	Description	Objective	Length (Days)	Sample size	
				Intervention	Control
DS1	Active main addresses in double sample with large or medium base weights	Determine whether it is possible to prioritize cases during the second phase.	11	46	46
DS2	Active screener addresses in double sample with large or medium base weights		11	26	25
DS3	Active main addresses in double sample with large base weights		10	28	28
DS4	Active screener addresses in double sample with large base weights		10	20	20
SB1	Active main addresses with no children under 15 years of age on household roster	Determine whether it is possible to improve sample balance through prioritization.	15	232	188[b]
SB2	Active main addresses with no children under 15 years of age by interviewer observation		8	167	315
SB3	Active main addresses with older (age 20–44) non-Black and non-Hispanic males		13	103	85
SB4	Active main addresses with older (age 20–44) Hispanic males		11	69	62

[a] EXT = subgroup defined by external data; INT = subgroup defined by internal paradata used to estimate predicted probabilities of response; DS = Phase 2 subgroup defined by stratification and weight paradata; SB = sample balance subgroup.

[b] Subset of control cases that were also part of simultaneous non-randomized sample balance intervention (see Section 3.3) deleted from comparison.

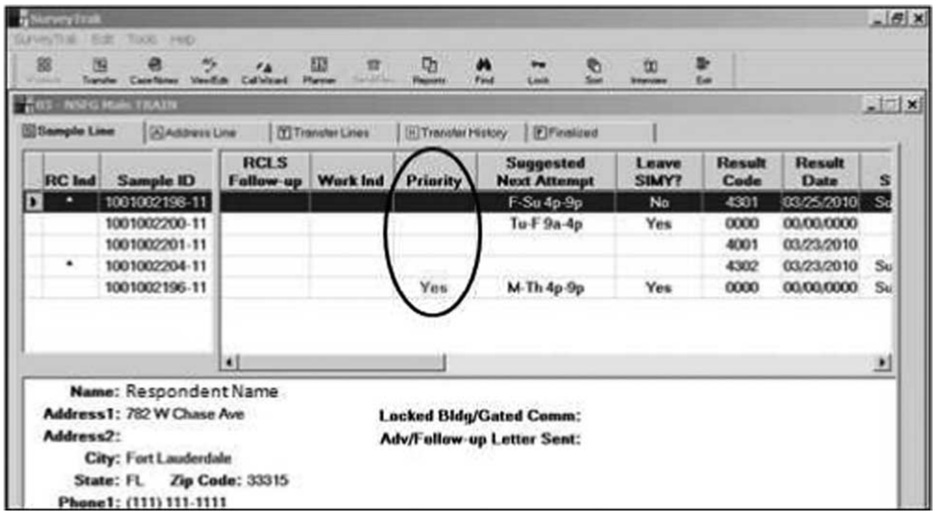
The prioritized cases were “flagged” in the interviewers’ view of the sample management system. Figure 1 shows how these flags appeared to the interviewer. Interviewers were asked to prioritize the “flagged” cases and apply more effort to these cases. Instructions about the interventions were communicated to interviewers in a weekly telephone call (or “team meeting”) and by email. Subsequent analyses examined effort on intervention and control addresses to determine if a null hypothesis of no difference in number of calls or response rates between the two groups of cases could be rejected.

Since these interventions occurred later in some quarters and also targeted different types of cases (given secondary objectives), sample sizes in intervention and control groups were sometimes small. There was limited power to detect even modest differences in response rates in many of these randomized experiments. Rather than focus on individual experiments that rejected the null hypothesis of no difference, we summarize findings across experiments. Across the 16 interventions, the null hypotheses for the number of calls or response rates might be expected to be rejected (using a 5% level of significance) in less than one intervention by chance alone.

For each randomized intervention, there are two questions posed:

1. Do interviewers do what we ask of them (that is, do they increase the number of calls to high priority cases)?
2. Does the intervention increase the response rate among the target high priority cases?

Table 1 summarizes the characteristics of the 16 randomized interventions. Intervention periods ranged from eight to 15 days, and sample sizes from 20 to 759 per intervention or control group. Subgroups subject to intervention varied on a number of characteristics; specifically, we distinguish between four types of case prioritization interventions. Three interventions were primarily based on external (EXT) commercial data purchased to determine whether household eligibility could be reliably predicted for addresses from the external source before the screening interview was completed. Five were based on internal



RC Ind	Sample ID	RCLS Follow-up	Work Ind	Priority	Suggested Next Attempt	Leave SIMY?	Result Code	Result Date	S
•	1001002198-11				F-Su 4p-9p	No	4301	03/25/2010	Su
	1001002200-11				Tu-F 9a-4p	Yes	0000	00/00/0000	
	1001002201-11						4001	03/23/2010	
•	1001002204-11						4302	03/23/2010	Su
	1001002196-11			Yes	M-Th 4p-9p	Yes	0000	00/00/0000	Su

Name: Respondent Name
Address1: 782 W Chase Ave
Address2:
City: Fort Lauderdale
State: FL Zip Code: 33315
Phone1: (111) 111-1111

Locked Bldg/Gated Comm:
Adv/Follow-up Letter Sent:

Fig. 1. Screen shot of an active sample line “flagged” as high priority in the sample management system, 2006–2010 Continuous National Survey of Family Growth

(INT) NSFG paradata used to predict either propensity to respond on a given day of the quarter or eligibility status. These predictions were based on logistic regression models fitted to addresses or households for which response status was known (responded or not) or household eligibility was known (eligible or not). Predictors included contact information recorded by interviewers at each household contact, interviewer observations about sample block or sample address characteristics, or interviewer judgments about individuals living in the household. Interventions of this type targeted addresses with high or medium predicted probabilities of response, addresses with high base weights in an effort to improve response rates, or high predicted probabilities that an address had one or more eligible persons residing there. These interventions helped to increase the yield of the sample, which was an important objective. Other interventions and design features were aimed at minimizing nonresponse bias. Four randomized interventions (INT2, INT3, INT4, and INT5) involved combinations of sample selection criteria. The subgroups for these four interventions were all based, though, on internal models driven by paradata, and are thus classified as the internal type.

Four interventions were conducted on the Phase 2 or double sample (DS) selected addresses. Cases with a high selection weight or a high probability of response were prioritized during the second phase.

Four additional interventions were randomized experiments to assess whether sample balance (SB) on key subgroups could be restored by intervention on high priority addresses. In addition to a sample balance intervention on Hispanic males ages 20–44 years, interventions were conducted on main addresses judged by interviewers to have no children under 15, with no children on the household roster from the screening interview, and non-Black and non-Hispanic males ages 20–44 years, groups that were observed to have lower response rates in particular quarters. The interviewer judgment about the presence of young children was one of several interviewer observations collected to provide NSFG managers with auxiliary information enabling comparisons of responding and nonresponding households. Groves et al. (2009) provide a more detailed description of these interviewer observations, and West (2013) examines the accuracy of the observations and shows that the observations are correlated with both response propensity and several key variables collected in the NSFG interview.

We consider first whether flagging high priority addresses changed interviewer behavior. Figure 2 presents bar charts of the mean cumulative calls per address for both the intervention and control groups of addresses at the conclusion of each of the 16 interventions. Significant differences in mean cumulative calls at the $P < 0.05$ level based on independent samples t-tests are highlighted. The means in Figure 2 consistently show the intervention addresses receiving more calls than the control addresses. Approximately half (seven) of the experiments resulted in statistically significant two-sample hypothesis test results.

The interventions clearly had a consistent impact on interviewer calling efforts. The next question was whether the increased effort led to corresponding increases in response rates for intervention relative to control addresses. Figure 3 presents comparisons of final response rates (according to the AAPOR RR1 definition) at the end of each intervention for the intervention and control groups. Significant differences in final response rates with $P < 0.05$ for a χ^2 test of independence (where distributions on a binary indicator of

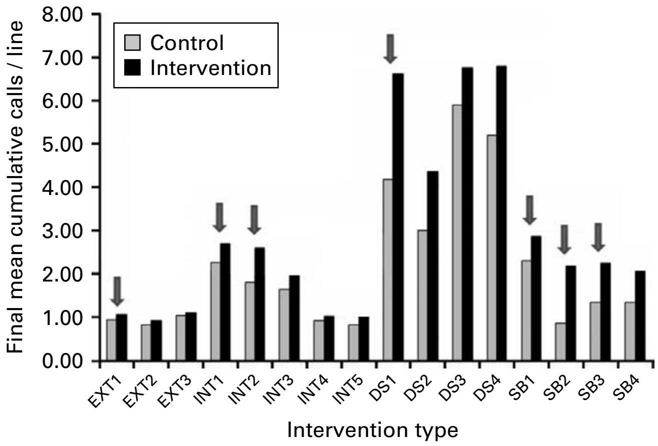


Fig. 2. Mean cumulative calls at the end of an intervention (arrows indicate significance at $\alpha = 0.05$ for independent samples t-tests) for intervention and control groups in 16 randomized trials, 2006–2010 Continuous National Survey of Family Growth

response were compared between the intervention and control groups of addresses) were found for only two of the 16 interventions (screener addresses predicted to have high eligibility and main addresses with no children under age 15 years in the household from the roster data). Response rates were generally found to be higher in the intervention groups, but there were four experiments with slightly higher response rates in the control group. Thus, there is some evidence that increased calling efforts tended to result in increases in response rates, although statistically significant increases occurred in only two of the interventions. Across all 16 interventions, there was a weak positive association between increased calling effort and increased response rates.

Two of the four interventions with higher response rates for control cases are interventions that were implemented during the double sample (DS) period. A third double sample intervention (DS3) resulted in equal final response rates in the two groups. During the second-phase period of the NSFG's double sampling operation, more attention was

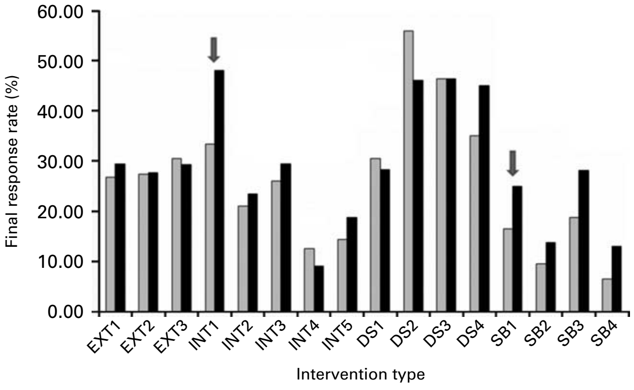


Fig. 3. Response rates (arrows indicate significant differences at $\alpha = 0.05$ in χ^2 tests of independence) for intervention and control groups in 16 randomized interventions, 2006–2010 Continuous National Survey of Family Growth

being paid to all active addresses. If these three double sample interventions (DS1, DS3, and DS4) are removed, there is much clearer evidence of a positive association between increase in effort and increase in response rates. These results for the double sample interventions indicate that intervening during an already intensive effort in a double sample period will not necessarily increase response rates. Because interviewers have a greatly reduced workload (approximately 1/3 of their assigned cases that have not been finalized are retained during the second phase), it may be that all cases are already being called more frequently than during Phase 1, and the additional calls on prioritized cases do not lead to additional contacts and interviews.

As a final evaluation of the randomized interventions, we present a more detailed analysis of the effectiveness of one of the 16 "internal" interventions, INT5. During each of the interventions, interviewers established appointments with active cases, and interviews were then completed after the end of the intervention period (which was chosen arbitrarily by NSFG managers). The question of interest is whether higher effort levels continued for intervention addresses after the end of the intervention period, and whether there is an increase in completed interviews relative to control cases. We suspected that higher calling rates would continue for intervention cases, because more calls should yield more contact with household members, more appointments, and interviewer visit patterns guided by more information about when household members are more likely to be at home.

We chose INT5 for this analysis for three reasons. First, this intervention had a balanced design, with a relatively large sample of 133 addresses in each arm of the experiment. Second, anecdotal reports from interviewers indicated that this intervention, although it did not lead to a significant difference in response rates, did lead to an increase in the number of appointments for the group receiving the intervention. We hypothesized that this appointment-setting work may have led to increased response rates after the intervention concluded. Third, because the experimental group did not receive higher numbers of calls or have higher response rates in this intervention, we wanted to see if this was an artifact of our arbitrarily ending the analysis of the treatment effect with the end of the prioritization.

A total of 266 active addresses without a screening interview, with larger base sampling weights (the largest tercile of the distribution of weights) and higher estimated response propensities predicted by the paradata (upper one-third of all active addresses), were selected for the INT5 intervention. One half (133) of these addresses were assigned to the intervention group, and the rest were assigned to the control group (where they received standard effort from the interviewers). There was a clear long-run benefit of the intervention on calling behavior. After the "end date" of this intervention (29 August 2007), at which time there was a slightly higher number of calls in the intervention group, the gap between the groups continued to increase, eventually leading to roughly 0.5 calls per address more on average than control cases. This result indicates that intervention addresses did receive higher calling effort during the intervention periods, and that the higher call effort continued with more calls being placed to intervention cases until the end of the quarter.

When we examined the cumulative response rate for each group in INT5, the largest gap in response rates between the two groups occurred when the intervention was originally stopped on 29 August 2007. After this date, the gap between the two groups remained similar, with response rates increasing at the same rate in both groups, and the intervention group continuing to have a higher response rate until the end of the quarter. This constant gap may have been a function of the continued increase in calls to these cases after the intervention was stopped. In sum, these case prioritization experiments demonstrated that we have the ability to alter field data collection efforts from a central office. This capability should aid the reduction of interviewer variability while improving the balance of selected characteristics of the set of interviewed cases relative to the full sample. The latter may result from identifying eligible cases more quickly or by improving sample balance (see Section 3.3). Finally, we note the importance of continued experimentation with these techniques for discovering unintended consequences. For example, in the case of the interventions applied to the second phase samples, we found that the interventions were not effective, as interviewers were essentially prioritizing all of their remaining cases.

3.2. Screener Week: Shifting Effort to Incomplete Screener Addresses

From our experience with the implementation of NSFG Cycle 6 (Groves et al. 2005), the management team had observed that interviewers varied in how they scheduled work. Interviewers typically scheduled main interviews even when assignments included a large proportion of incomplete screener addresses. In the last weeks of Cycle 6, there remained data collection screener addresses with a limited number of calls and no completed screener interview. These indicators pointed to an interviewer preference for completing main interviews over screening households for eligible persons.

This apparent preference created two issues for the continuous design employed in the 2006–2010 NSFG. First, because main interviews could not be completed until screener interviews were completed, interviewers had limited time to complete main interviews with cases that were screened later in the process. This hampered our ability to attain high response rates in a study with a relatively short field period each quarter (12 weeks). Second, the screening interview generates important auxiliary data for further responsive design decisions. Information about the age, race, ethnicity, and sex of the selected person as well as an interviewer judgment about whether the selected person is in a sexually active relationship with a person of the opposite sex are used in subsequent interventions to improve the balance of the interviewed cases relative to the full sample along these dimensions (see Section 3.3 for a full description).

In the Continuous NSFG, project management sought to divert interviewer effort to screener addresses at an earlier point in the data collection. An intervention strategy was sought that would increase effort to call at any remaining previously not-contacted screener sample addresses, resolve access impediment issues that blocked contact attempts, and ultimately produce more screener interviews (regardless of whether age-eligible persons were present).

The field management strategy in week 5 of the first quarter was to instruct interviewers to keep all current firm main interview appointments made previously, to set main interview appointments at screener interview completion with selected eligible respondents during week 5 if a later time was not available, and to schedule main interview appointments with sample persons not present at the completion of the screener interview *after* week 5. Field management then emphasized the importance of making calls on screener addresses during this week. The instructions were given in regularly scheduled telephone conference calls and in email correspondence.

Field management monitored screener calling and interview progress by using daily electronically-submitted interviewer call records before, during, and after week 5. There was an increase in screener calls and an increase in the ratio of screener to main calls during week 5 of year 1, quarter 1 (Y1Q1). Field management instituted screener week in week 5 (days 29 to 35) in Y1Q1, and in each subsequent quarter until the conclusion of data collection in 2010. There was one exception – in Y2Q2, screener week was implemented in week 4.

Graphs of daily and seven-day moving averages of completed screener and main interviews, such as those shown in Figure 4, were examined throughout each quarter. The upper black lines in Figure 4 track the daily and seven-day moving average number of screener interviews. In later quarters after the first, field management compared current quarter results to a previous quarter, to a previous quarter one year earlier, or a yearly average across quarters from a previous year. The lower grey lines similarly track the corresponding daily and seven-day moving average number of main interviews.

The number of screener interviews in the first three weeks of a quarter (Figure 4 presents data from Y4Q1) was between 80 and 100. The count gradually declined to less than 20 per day at the end of Phase 1 each quarter. There were relatively steady main interview counts per day of around 20 after the first three or four weeks of each quarter. The upper black lines in Figure 4 show (where vertical lines separate the weeks) a rise in the number of screener interviews in week 5, and little change in the number of main interviews.

The number of calls to active screeners and the screener to main call ratio increased in each quarter during screener week. While the size and consistency of the increases in each

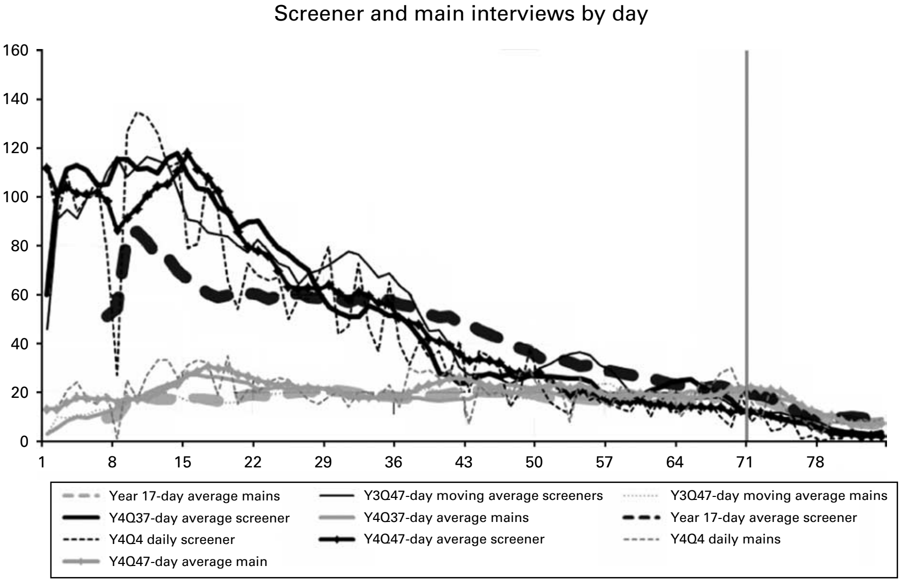


Fig. 4. Number of and seven-day moving average screener and main interviews by day for year 1 and data collection quarters Y3Q2, Y4Q1, and Y4Q2, 2006–2010 Continuous National Survey of Family Growth

screener week suggested a change in interviewer behavior, there was no experimental validation of this result. In an effort to evaluate further whether "screener week" had an impact on the volume of screener calls, two statistical models were fit to the paradata, and hypothesis tests about model parameters were conducted.

In the first model, the dependent variable was the daily number of screener calls for weeks 3 through 7 (days 15 to 49) in each of the 16 quarters – that is, the days before (days 15–28), during (days 29–35), and after (days 36–49) screener week. This included two weeks before screener week, the screener week, and two weeks after the screener week. There was one exception. InY2Q2, screener week was initiated in week 4, and for that quarter, weeks 2 through 6 are included in the analysis. There were 559 days across 16 quarters in the analysis (Y2Q2 only included 34 days, because the screener week intervention lasted only six days in that quarter).

The number of screener calls was regressed on the day number, an indicator of whether the day was in screener week, and the quarter number. Interactions among day, the screener week indicator, and the quarter were also included in the model. A three-way screener week by day by quarter interaction suggests a complex interviewer response in which screener week call levels were irregular during screener week and across quarters. Two-way screener week by quarter and day by quarter interactions would indicate whether screener call levels differ across quarters and across days within quarters. A two-way day by screener week interaction indicates whether there was a different number of screener calls across days in screener week. The screener week by quarter interaction was expected to be statistically significant, because there was observed variation in the number of screener calls during screener week across quarters. The day by screener week interaction was also expected to be significant, because in each screener week there was a rising number of screener calls from the beginning to the end of the week. Table 2 and Figure 5a and 5b summarize the model and the results of tests of null hypotheses about model parameters for the number of screener calls. Figure 5a presents predicted screener call levels by day for each quarter obtained from a reduced model that used only the statistically significant coefficients to compute the predicted values. That is, Figure 5a presents a "smoothed" image of the daily screener call levels as estimated from the reduced model.

Table 2. Analysis of factors affecting the number of calls per day made before, during, and after screener week, 2006–2010 Continuous National Survey of Family Growth

Factor	F-Statistic	Numerator DF	Denominator DF	*P*-value
Day of field period	159.20	1	525	< 0.0001
Screener week	7.44	1	525	0.0066
Quarter	2.13	15	525	0.0079
Screener week × day	12.04	1	525	0.0006
Screener week × quarter	NS	–	–	–
Day × quarter	1.73	15	525	0.0425
Screener week × day × quarter	NS	–	–	–

NS = Not significant. Model $R^2 = 0.345$. The F-statistics test the hypothesis that the factor coefficients are different from zero in the presence of the other factors in the model.

There were statistically significant interactions between day and screener week and day and quarter, as expected. After removing the parameters associated with the other factors which could not be distinguished from zero, the remaining five factors explained 34.5% of the variance in daily screener calls. The significant interactions indicate that the number of

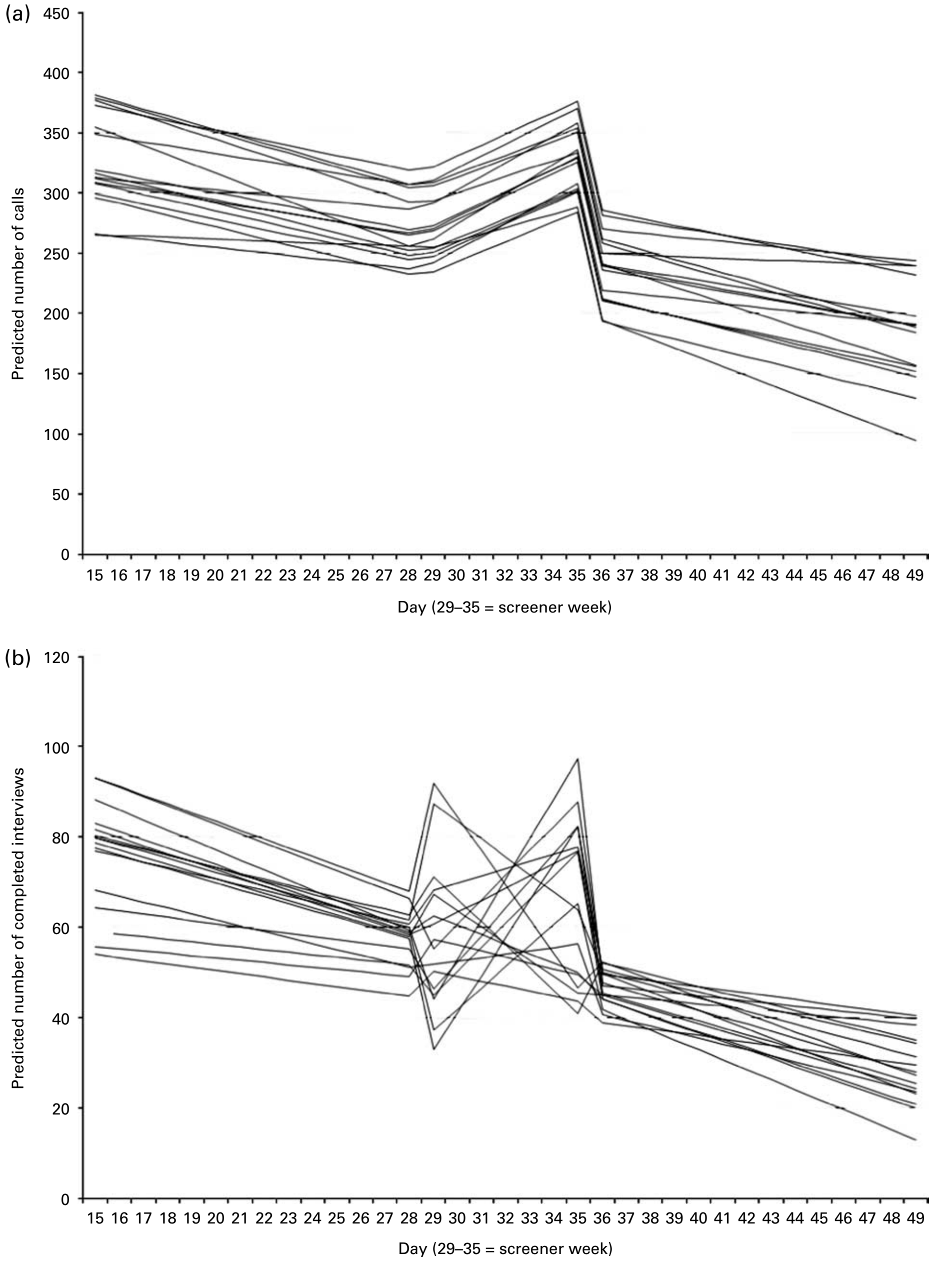


Fig. 5. Number of daily calls (Figure 5a) and number of daily completed screeners (Figure 5b) predicted under models with only statistically significant coefficients for weeks 3–7 (screener week days 29–35) for the 16 quarters of data collection, 2006–2010 Continuous National Survey of Family Growth

screener calls each day changes during screener week, and that the number of screener calls on average was different across quarters.

These findings are confirmed in Figure 5a. The figure shows that the predicted number of calls declines, except during screener week (days 29–35). There is an increasing predicted number of calls made per day across screener week. This increase occurred consistently across all quarters. The increasing number of screener calls during week 5 reverses a negative trend in screener calls per day before, and after, screener week. There is also evidence of variation in calling behavior across quarters, with some quarters having more calls than others over the five-week period. The effects of screener week and changes in the screener calling trends during screener week were, however, consistent across the 16 quarters.

The second model had identical predictors, but the dependent variable was changed to the daily number of completed screener interviews. Table 3 summarizes the test statistics for the second model. There is a significant three-way interaction between day, screener week indicator, and quarter, suggesting that changes in the number of interviews occurred across day within screener week, and that the day by screener week trend was not the same across quarters. The consistent increases in the number of screener calls across days in screener week were not repeated across quarters for the number of completed screener interviews.

Figure 5b shows "smoothed" predicted counts of screener interviews per day based on the fitted regression model including the three-way interaction. Figure 5b is a striking contrast to Figure 5a. The general trend of decreasing numbers of screener interviews before, and again after, screener week, is interrupted by a complex rise and fall of completed screeners during screener week in each quarter. The expected rate of completed screener interviews per day did not consistently increase during screener week across the 16 quarters. Consistent increases in screener calls across days of screener week did not produce consistently increasing numbers of completed screener interviews. There were initial decreases in screener interviews followed by increases during one half of the screener weeks, while in the other weeks there were sharp to modest increases early in screener week followed by decreases. Across all 16 quarters, during screener week there was an average effect of increased numbers of completed screeners, but the rates of

Table 3. Analysis of factors affecting the number of completed screener interviews per day before, during, and after screener week, 2006–2010 Continuous National Survey of Family Growth

Factor	F-Statistic	Numerator DF	Denominator DF	*P*-value
Day of field period	309.26	1	495	< 0.0001
Screener week	4.82	1	495	0.0286
Quarter	3.57	15	495	< 0.0001
Screener week × day	7.94	1	495	0.0050
Screener week × quarter	1.78	15	495	0.0351
Day × quarter	2.81	15	495	0.0003
Screener week × day × quarter	1.86	15	495	0.0245

Model $R^2 = 0.463$. The F-statistics test the hypothesis that the factor coefficents are different from zero in the presence of the other factors in the model.

completed screeners were not consistent across quarters. The reasons for this inconsistent effect may have to do with changes in the interviewing staff, variation between samples, or possible seasonal effects.

Although not experimentally implemented, we would argue that the emphasis on early screening helped to improve response rates. Logically, the screening interview needs to be completed before the main interview. The sooner that this task is completed, the more opportunity there is to complete the main interview. The screening week intervention was implemented during days 29 to 35 each quarter. Empirically, about 93% of the cases are interviewed within 49 days after being screened as eligible. Only 89.5% of cases are interviewed within 42 days after being screened as eligible. Identifying eligible persons as early as possible will therefore increase the likelihood of completing an interview. It was our experience from a prior Cycle of the NSFG (and other large-scale surveys using screening) that interviewers prefer to complete main interviews. They may delay screening, thus decreasing the time available to complete interviews with newly identified eligible persons. In addition, the rapid screening of households enabled us to use paradata from the household screening to create a proxy indicator for nonresponse bias that guided the types of interventions described in the next section.

3.3. Sample Balance: Targeting Subgroups in Order to Reduce Variation in Subgroup Response Rates

The third type of intervention, *sample balance*, was designed to reduce the risk of nonresponse bias. Since the survey variables for nonresponders are not known, a proxy indicator for nonresponse bias was needed. The proxy indicator chosen for this purpose was variation in subgroup response rates. NSFG management monitored the response rates of 12 individual subgroups and the coefficient of variation of these subgroup response rates on a daily basis. The variation in subgroup response rates reflects how closely the set of interviewed cases matches the sampled cases on the key characteristics used to define the subgroups – in this case, age, race, ethnicity and sex. In this sense, this indicator is very similar to the R-Indicator (Schouten et al. 2009). This type of intervention sought to bring the composition of the set of interviewed cases closer to the composition of the full sample by prioritizing cases from subgroups that were responding at lower rates. The key characteristics used for this purpose were age, race-ethnicity, sex, and presence of children under the age of 15 in the household (each collected during the screener interview), as well as presence of children under the age of 15 in the household (from interviewer observation). Of course, we cannot be certain that this approach actually reduces bias.

The distribution of the daily response rates by subgroups varied some over the years and quarters. This variation could be due to changes in the composition of the samples each quarter and changes in the interviewing staff each year. In many quarters, one subgroup showed lower numbers of interviews and lower response rates: Hispanic males ages 20–44. Figure 6 is an actual dashboard display monitored by NSFG management. It shows response rates for days 1 to 70 (the first 10 weeks) of Y4Q2. The denominator for each subgroup changes daily, as new cases are identified through the screening process. For instance, on the first day of Y4Q2, one Hispanic male 15–19 years of age was identified and interviewed. Therefore, the response rate for this subgroup is 100% on day 1, and goes

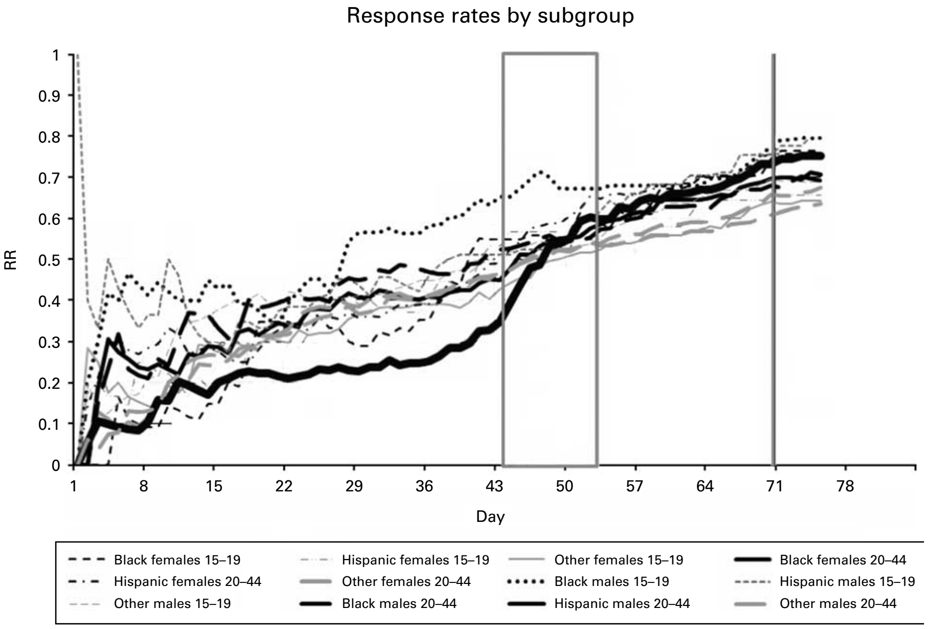


Fig. 6. Daily cumulative response rates for twelve subgroups defined by gender, race-ethnicity, and age, with the intervention period for Hispanic males 20–44 years of age in days 44–55 highlighted, 2006–2010 Continuous National Survey of Family Growth

down the next day as new cases are identified. As sample sizes increase, differences in response rate stabilize. In the case of Y4Q2, through week 6, Hispanic males 20–44 years of age had lower response rates.

In response to observed trends in a given quarter, field management developed an intervention to restore balance in the composition of the interviewed cases. At different points in each quarter, all outstanding addresses known to contain selected persons in a low response rate subgroup were identified. At the start of a sample balance field intervention, field management marked these addresses as high priority in the central sample management system. During nightly uploads of data, interviewers also downloaded the updated priority data from the sample management system.

In several quarters when sample balance interventions were conducted, the high priority designation was randomly assigned to one half the target subgroup addresses. The results of these randomized experiments are discussed in Section 3.1. Here only the non-randomized interventions are examined.

The high priority cases in randomized and non-randomized sample balance interventions were marked in laptop address lists with a high priority indicator (see Figure 1). Field management subsequently monitored daily response rates and numbers of interviews to observe if the priority assignment yielded the desired effect. Since some Hispanic males may require bilingual interviewers, field managers also assigned traveling interviewers with bilingual capabilities to segments containing addresses in this target subgroup.

Figure 6 also shows the effect of the "Hispanic male 20–44 years of age" intervention on response rates. The intervention began on day 44 of this quarter, with high priority case flags assigned in the sample management system to all addresses with a selected person from the targeted subgroup. There is a clear increase in response rates for this subgroup over the next week. This Y4Q2 intervention yielded, at the end of ten weeks of data collection, a response rate for Hispanic males 20–44 years of age that was similar to that for the other eleven subgroups. The intervention, therefore, had the beneficial effect of decreasing variation in response rates among these six subgroups. The variation in subgroup response rates is a process indicator monitored by NSFG managers to assess balance in the data set. This beneficial effect also translates to a reduction in the variation in nonresponse adjustment weights, and reduced sampling variance of weighted estimates.

The age-race-ethnicity-sex subgroups were not the only ones monitored and for which sample balance interventions, randomized and non-randomized, were attempted. For example, during Y4Q3, field managers noticed, while reviewing the dashboard, lagging response rates among sample households with children under the age of 15 (identified with screening data). Since the presence of young children is correlated with many of the key outcome statistics produced by the NSFG (West 2013), this indicator was also used as a proxy indicator for nonresponse bias. Establishing balance on this proxy indicator is meant to reduce the risk of nonresponse bias and mitigate the inflation of variance estimates due to the variability of nonresponse adjustment weights. Just as for Hispanic males ages 20–44 years, field management "flagged" high priority addresses for subgroups such as households without children less than 15 years of age (from the screening interview) to receive increased effort from the interviewers. Field management also sent email reminders advising interviewers that high priority addresses required extra effort on their part.

In sum, the interviewers followed centralized directions of how to prioritize their sample. This centralized prioritization can be used to improve the composition of the final set of respondents by increasing the response rates of groups that are "underrepresented" by the response process.

4. Summary and Conclusions

This article presents case studies of responsive design interventions generated from active monitoring of paradata. Three types of paradata-driven management interventions were examined: one applied to subgroups identified through a variety of internal and external paradata (*case prioritization*), one applied to all interviewers on a very broad level (*screener week*), and one applied to a selection of addresses with known key subgroup members (*sample balance*). Each illustrates important dimensions of the tools of responsive design, including the ability to use paradata to systematically alter interviewer behaviors during field work and the consequences of those behavioral changes for the nature of the survey data collected.

For case prioritization interventions, we found that interviewers will respond to centralized requests that set priorities on cases from key subgroups that are underresponding. The first analysis examined 16 different randomized interventions applied to addresses selected from groups defined by a variety of paradata. Interviewers followed intervention guidelines, making more calls on the experimental intervention addresses than on the control addresses. The intervention addresses also tended to achieve larger increases in response rates than the control cases during the intervention period.

The second intervention successfully increased the effort on active screener cases. This led to earlier identification of eligible sample persons and the collection of key information used later to assess sample balance. To model the impact of the second intervention, the week of the screener intervention was contrasted to two prior weeks and two following weeks. Rates of interviewer calling significantly increased in a consistent manner across quarters during the screener week, indicating that the intervention did indeed influence interviewer behavior. The increased rates of calling, however, did not consistently lead to increased numbers of interviews during the same period. Once again, responsive design tools can be effective at altering interviewer behavior as desired, but tests across a broader range of interventions will be required to determine the most effective tools.

The third set of interventions was based on a proxy indicator for the risk of nonresponse bias – variation in subgroup response rates. Intervention on cases from "under-represented" subgroups not only affected interviewer behaviors, but in this important case also increased the subgroup response rate and reduced the variation of the response rates among key subgroups. This type of targeted intervention was successful at improving the balance of cases interviewed across subgroups defined by age, race-ethnicity, sex and other characteristics important in predicting survey outcome variables.

The overall conclusion that can be drawn from these findings is that interviewers were attentive to and accepted the centralized intervention strategies in the NSFG, despite not being told the reason for increased effort on certain addresses in most interventions. Interviewers were notified electronically and via conference calls that certain addresses would be high priority and to place emphasis on these lines as they planned their work.

A *sine qua non* of responsive design is, therefore, the ability of the central office staff to instruct the field interviewers to change their focus from one task to another. The three case studies in this article show that real-time interventions can lead to changes in key indicators of survey quality. All interventions were successful at altering interviewer behaviors, but not all interventions were successful at altering survey outcomes. Continuous examination of the practice of responsive design and investigation across a broader set of interventions is necessary to identify the types of interventions that further improve survey costs and reduce survey errors.

The techniques demonstrated in this article can be used by survey organizations to control progress toward key quality indicators (Kirgis and Lepkowski forthcoming). These techniques require the development of reporting mechanisms that allow managers to review progress on a frequent basis. Managers may decide to intervene based on the information in these reports. If, for example, important subgroups are responding at a lower rate, managers may wish to redirect interviewer effort toward cases in these low-responding subgroups. In order to re-prioritize field interviewer effort toward specific cases, managers must have the means to do so – for example, the use of "flags" in interviewer sample management systems. In this way, survey managers can control progress toward key indicators.

Given what we have learned in this investigation, the highest priority for new research in this area is to understand the circumstances under which centralized prioritization will lead to increased effort. We experienced variation in outcomes across the 16 interventions.

Understanding the sources of this variation may help researchers design more effective interventions. What factors mitigate the effectiveness of these experimental treatments? Is it a factor that varies across interviewers, or other factors that vary across samples? Or is it interactions between features of the design? For example, it appears that when interviewers have small workloads where all cases are receiving high priority (as in the NSFG second phase), centralized prioritization will be less effective. In addition, the consequences of using proxy indicators for nonresponse bias need to be evaluated. Understanding when this practice produces the desired results may require methodological "gold standard" studies designed specifically to investigate this question. There is certainly more work to be done in the development of these proxy indicators. In the case of the NSFG, demographic variables such as age, sex, race, and ethnicity are predictive of the key survey measures (Martinez et al. 2012). This may not be true for every study. A recent study by Peytcheva and Groves (2009) found that the types of demographic variables used to define some of our interventions were not predictive of nonresponse bias in the 23 specialized studies that they examined. More work is needed to develop "tailored" paradata suited for predicting the key survey variables of each particular study. Finally, although our focus was on the risk of nonresponse bias, other sources of error need to be included in the planning and execution of responsive designs. The tools outlined in this article are a valuable first step toward a "total survey error" perspective for responsive designs.

5. References

Couper, M.P. (1998). Measuring Survey Quality in a CASIC Environment. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Dallas, TX.

Couper, M.P. and L. Lyberg (2005). The Use of Paradata in Survey Research. Proceedings of the International Statistical Institute Meetings.

Durrant, G.B., D'Arrigo, J., and Steele, F. (2011). Using Paradata to Predict Best Times of Contact, Conditioning on Household and Interviewer Influences. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174, 1029–1049.

Fienberg, S.E. and Tanur, J.M. (1988). From the Inside Out and the Outside In: Combining Experimental and Sampling Structures. The Canadian Journal of Statistics, 19, 135–151.

Fienberg, S.E. and Tanur, J.M. (1989). Combining Cognitive and Statistical Approaches to Survey Design. Science, 243, 1017–1022.

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

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Nonresponse and Costs. Journal of the Royal Statistical Society, Series A, 169, 439–457.

Groves, R.M., Benson, G., Mosher, W.D., Rosenbaum, J., Granda, P., Axinn, W., Lepkowski, J.M., and Chandra, A. (2005). Plan and Operation of Cycle 6 of the National Survey of Family Growth. Vital and Health Statistics, Series 1, No. 42. Hyattsville, MD: National Center for Health Statistics (Available from http://www.cdc.gov/nchs/data/series/sr_01/sr01_042.pdf, accessed October 11, 2011).

Groves, R.M., Mosher, W.D., Lepkowski, J., and Kirgis, N.G. (2009). Planning and Development of the Continuous National Survey of Family Growth. National Center for Health Statistics. Vital Health Statistics, Series 1, No. 48. Hyattsville, MD: National Center for Health Statistics (Available from http://www.cdc.gov/nchs/data/series/sr_01/sr01_048.pdf, accessed October 11, 2011).

Kirgis, N. and Lepkowski, J.M. (forthcoming). Design and Management Strategies for Paradata-Driven Responsive Design. Improving Surveys with Paradata: Analytic Use of Process Information, Frauke Kreuter (ed.).

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys. Journal of the Royal Statistical Society: Series A (Statistics in Society), 17, 389–407.

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M., and Van Hoewyk, J. (2010). The 2006–2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey, National Center for Health Statistics, 2(150).

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. Hoboken, N.J.: Wiley.

Martinez, G., Daniels, K., and Chandra, A. (2012). Fertility of men and women aged 15–44 years in the United States: National Survey of Family Growth, 2006–2010. National health statistics reports; no. 51. Hyattsville, MD: National Center for Health Statistics (Available from http://www.cdc.gov/nchs/data/nhsr/nhsr051.pdf, accessed August 12, 2012).

National Research Council (2007). Using the American Community Survey: Benefits and Challenges. Panel on the Functionality and Usability of Data from the American Community Survey. Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Constance F. Citro and Graham Kalton (eds). Washington, D.C. The National Academies Press.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010). Reduction of Nonresponse Bias in Surveys through Case Prioritization. Survey Research Methods, 4, 21–29.

Peytcheva, E. and Groves, R.M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193–201.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the Representativeness of Response. Survey Methodology, 35, 101–113.

Stoop, I.A.L., Billiet, J., Koch, A., and Fitzgerald, R. (2010). Improving Survey Response: Lessons Learned from the European Social Survey. Chichester, West Sussex, U.K. Hoboken, N.J.: Wiley.

West, B.T. (2013). An Examination of the Quality and Utility of Interviewer Observations of Household Characteristics in the National Survey of Family Growth. Journal of the Royal Statistical Society, Series A (forthcoming).

Received September 2011
Revised March 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, p. 501

Editorial Note

The value of the peer review process for advancing theory and practice in science is well recognised and widely acknowledged as an important feature of a sound scientific evaluation process. One part of the process, often taking place ‘behind the scenes’, is the discussion between the actors involved – authors, expert reviewers, journal editors – on merits, possible shortcomings, and ways to improve the submitted contribution. This discussion in effect directly influences what a journal publishes and what it does not publish, and in the long run paves the way of scientific progress within the field.

Opening up this process somewhat, JOS in this issue publishes an article that addresses a complex phenomenon – the comparison of methods for evaluation of survey questions. This specific area still lacks an established, standard approach. Therefore, together with the article by Yan et al., we publish two discussions, one by Willem Saris and one by Jennifer Madans and Paul Beatty. In doing so, we hope to stimulate the discussion and identification of areas in need of further scientific attention by openly presenting the existing issues and reasoning behind the different approaches to, in this case, comparison of survey question evaluation methods.

Editors-in-Chief

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 503–529

Evaluating Survey Questions: A Comparison of Methods

Ting Yan[1], *Frauke Kreuter*[2], *and Roger Tourangeau*[3]

This study compares five techniques to evaluate survey questions — expert reviews, cognitive interviews, quantitative measures of reliability and validity, and error rates from latent class models. It is the first such comparison that includes both quantitative and qualitative methods. We examined several sets of items, each consisting of three questions intended to measure the same underlying construct. We found low consistency across the methods in how they rank ordered the items within each set. Still, there was considerable agreement between the expert ratings and the latent class method and between the cognitive interviews and the validity estimates. Overall, the methods yield different and sometimes contradictory conclusions with regard to the 15 items pretested. The findings raise the issue of whether results from different testing methods should agree.

Key words: Cognitive interviews; expert reviews; latent class analysis; measurement error; question pretests; reliability; validity.

1. Introduction

Survey researchers have a variety of techniques at their disposal for evaluating survey questions (see Presser et al. 2004b). These range from cognitive interviews (e.g., Willis 2005), to the conventional pretests recommended in many questionnaire design texts (e.g., Converse and Presser 1986), to behavior coding of various types (Maynard et al. 2002; van der Zouwen and Smit 2004), to question wording experiments (e.g., Fowler 2004), to the application of statistical procedures, such as latent class analysis (e.g., Biemer 2004) or structural equation modeling (Saris and Gallhofer 2007), that provide quantitative estimates of the level of error in specific items. These different evaluation techniques do not bear a close family resemblance. Although they all share the general goal of helping question writers to evaluate survey questions, they differ in their underlying assumptions,

[1] NORC at the University of Chicago, 1155 E. 60th Street, Chicago IL 60637, U.S.A.
Email: tingyan@umich.edu
[2] Joint Program in Survey Methodology, University of Maryland, College Park, 1218Q LeFrak Hall, Maryland, U.S.A. Institute for Employment Research/LMU, Nuremberg/Munich, Germany. Email: fkreuter@umd.edu
[3] Joint Program in Survey Methodology, University of Maryland, 1218Q LeFrak Hall, Maryland, U.S.A. Survey Research Center, University of Michigan, Ann Arbor, U.S.A. Email: RogerTourangeau@Westat.com

Acknowledgments: An earlier version of this article was presented at the 2008 Meeting of the American Association for Public Opinion Research. The work reported here is supported by the National Science Foundation under Grants SES 0550002 and 0549916 to the authors. We would like to thank the Methodology, Measurement, and Statistics (MMS) Program and Dr. Cheryl Eavey for their support. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Cat Tourangeau for her able research assistance; Norman Bradburn, Ashley Landreth, Stanley Presser, and Nora Cate Schaeffer for their expert reviews of the questions; and Willem Saris for providing SQP predictions for item quality. In addition, we thank Paul Beatty and Willem Saris for valuable comments on earlier versions of the manuscript.

the data collection methods they use, the types of problems they identify, the practical requirements for carrying them out, the type of results they generate, and so on.

To illustrate the differences across techniques, consider cognitive interviewing and the use of latent class modeling methods for evaluating survey items. Cognitive interviewing is a practice derived from the protocol analyses used in the work of Simon and his collaborators. Loftus (1984) first pointed out the potential relevance of Simon's work to the testing of survey items more than 25 years ago. The key assumptions of protocol analysis and its latter-day survey descendant, cognitive interviewing, are that the cognitive processes involved in answering survey questions leave traces in working memory (often intermediate products of the process of formulating an answer) and that respondents can verbalize these traces with minimal distortion (see Ericsson and Simon 1980). Cognitive interviews added several techniques to the think-aloud methods introduced by Simon and his colleagues, especially the use of specially designed probes, or follow-up questions; responses to these probes are also thought to provide important clues about how respondents come up with their answers and about potential problems with those processes. Some researchers see later developments of cognitive interviews as a departure from the original paradigm proposed by Ericsson and Simon, and argue that the use of probes has largely supplanted the use of think-aloud methods in cognitive interviewing (see, for example, Schaeffer and Presser 2003, p. 82; see also Beatty and Willis 2007; Gerber 1999). Cognitive interviews are rarely subjected to formal analyses; instead, the questionnaire testing personnel, often staff with advanced degrees, draw conclusions about the questions from their impressions of the verbal reports produced by respondents during the cognitive interviews (see Willis 2005 for a thorough discussion of cognitive interviewing).

At the other end of the continuum stands the application of quantitative methods, such as latent class modeling, to assess problems in survey questions. In a series of papers, Biemer and his colleagues (Biemer 2004; Biemer and Wiesen 2002; Biemer and Witt 1996) have used latent class models to estimate error rates in survey items designed to assess such categorical constructs as whether a person is employed or not. Latent class analysis is sometimes described as the categorical analogue to factor analysis (e.g., McCutcheon 1987, p. 7). It is used to model the relationships among a set of observed categorical variables that are indicators of two or more latent categories (e.g., whether one is truly employed or unemployed). In contrast to cognitive interviews, latent class analysis is a statistical technique that yields quantitative estimates. It uses maximum likelihood methods to estimate parameters that represent the prevalence of the latent classes and the probabilities of the different observed responses to the items conditional on membership in one of the latent classes.

Given the large number of different evaluation methods and the large differences between them, it is an important theoretical question whether the different methods *should* yield converging conclusions and, if not, whether they should be used alone or in combination with each other. In practice, the choice between techniques is often dictated by considerations of cost and schedule, and it is an important practical question whether clear conclusions will result even if different methods are adopted.

The answers to the questions of whether converging conclusions should be expected and how to cope with diverging conclusions about specific items depend in part on how researchers conceive of the purpose of the different evaluation methods. Much work on question evaluation and pretesting tends to treat question problems as a binary characteristic – the question either has a problem or it does not. Question evaluation methods are used to identify the problems with an item and group questions into two categories – those with problems that require the item to be revised, and those without such problems. Under this conceptualization, all of the question evaluation methods flag some items as problematic and others as non-problematic, even though the methods may differ in which items they place in each category. Of course, questions may have problems that differ in seriousness, but ultimately questions are grouped into those that require revision and those that do not. Different question evaluation methods are, then, compared on their success in identifying question problems and correctly placing items into one of the two categories. Presser and Blair's (1994) study is a classic example of such a conceptualization. Implicit in such work is the assumption that if *any* method reveals a problem with an item, that problem should be addressed. That assumption has been challenged recently by Conrad and Blair (2004; see also Conrad and Blair 2009), who argue that the "problems" found in cognitive interviews may well be false alarms.

Recently, the field of question evaluation and pretesting has seen a shift towards a more general conceptualization of question problems and goals of question evaluation methods (e.g., Miller 2009). Survey questions measure the construct they are supposed to measure more or less well (Saris and Gallhofer 2007). Thus, it is possible to conceive of question problems as a matter of degree, and the purpose of question evaluation methods is to determine the degree of fit between question and construct (Miller 2009).

There is limited empirical work comparing different question evaluation methods, especially work comparing qualitative methods (like cognitive interviews and expert reviews) with quantitative methods (like measurements of reliability and validity). The few prior studies that have been done seem to suggest that the consistency between the different methods is not very high, even at the level of classifying items as having problems or not (see Presser and Blair 1994; Rothgeb et al. 2001; and Willis et al. 1999, for examples). Table 1 provides a summary of the major studies comparing question evaluation techniques.

It is apparent from Table 1 that large disagreements across methods exist about which items have problems or which problems they have. There are several possible reasons for discrepant results across evaluation methods. The different methods may identify different types of problems. For instance, Presser and Blair (1994) found that interviewer debriefings are likely to pick up problems with administering the questions in the field, whereas cognitive interviews are likely to detect comprehension problems. The two methods may yield complementary sets of real problems. In addition, the methods may not be all that reliable. Partly, this unreliability may reflect differences in what the researchers count as problems and in how they conduct different types of evaluations. Several studies have examined whether multiple implementations of the "same" method yield similar conclusions about a set of items; the results suggest that unreliability within a method is often high (e.g., DeMaio and Landreth 2004; Presser and Blair 1994; Willis et al. 1999). Finally, another reason for disagreement across methods is that some of the evaluation methods may not yield valid results (cf. Presser et al. 2004a; on the potential for invalid conclusions, see Conrad and Blair 2004; 2009).

Table 1. *Studies comparing question evaluation methods*

Paper	Methods tested	Criteria	Conclusions
Fowler and Roman (1992)	1. Focus groups 2. Cognitive interviews 3. Conventional pretest 4. Interviewer ratings of items 5. Behavior coding	• Number of problems found • Type of problem found	1. Focus groups and cognitive interviews provide complementary information 2. Results from two sets of cognitive interviews (done by separate organizations) are similar 3. Interviewer debriefing identifies more problems than interviewer ratings and ratings identify more problems than behavior coding 4. All five methods provide useful information
Presser and Blair (1994)	1. Conventional pretests 2. Behavior coding 3. Cognitive interviews 4. Expert panels	• Number of problems found • Type of problem found • Consistency across trials with the same method	1. Conventional pretests and behavior coding found the most interviewer problems 2. Expert panels and cognitive interviews found the most analysis problems 3. Expert panels and behavior coding were more consistent across trials and found more types of problems 4. Behavior coding was most reliable but provided no information about the cause of a problem, did not find analysis problems, and did not distinguish between respondent-semantic and respondent-task problems 5. Expert panels were most cost-effective 6. Most common problems were respondent-semantic
Willis, Schechter, and Whitaker (1999)	1. Cognitive interviewing (done by interviewers at two organizations) 2. Expert review 3. Behavior coding	• Number of problems found • Consistency within and across methods regarding the presence of a problem (measured by the correlation across methods and organizations between the percent of the time items were classified as having a problem)	1. Expert review found the most problems 2. The correlation between behavior coding trials was highest (.79), followed closely by the correlation between the cognitive interviews done by two organizations (.68)

Table 1. Continued

Paper	Methods tested	Criteria	Conclusions
		• Type of problems	3. Across methods of pretesting and organizations, most problems were coded as comprehension/communication; there was a high rate of agreement in the use of sub-codes within this category across techniques
Rothgeb, Willis and Forsyth (2001)	Three organizations each used three methods to test three questionnaires	• Number of problems found	1. Formal cognitive appraisal (QAS) found most problems but encouraged a low threshold for problem identification
	1. Informal expert review 2. Formal cognitive appraisal 3. Cognitive interviewing	• Agreement across methods based on summary score for each item (summary scores ranged from 0 to 9 based on whether the item was flagged as a problem item by each technique and each organization)	2. Informal expert review and cognitive interviewing found similar numbers of problems, but found different items problematic 3. Results across organizations were more similar than across techniques: Moderate agreement across organizations in summary scores (r's range from .34 to .38) 4. Communication and comprehension problems were identified most often by all three techniques
Forsyth, Rothgeb and Willis (2004)	1. Informal expert review	• Conducted randomized experiment in a RDD survey that compared the original items pretested in 2001 study with revised items designed to fix problems found in the pretest	1. Items classified as high in interviewer problems during pretesting also had many problems in the field (according to behavior coding and interviewer ratings)
(Note: This study is a follow-up to Rothgeb et al. 2001)	2. Formal cognitive appraisal (QAS) 3. Cognitive interviewing	• Classified items as low, moderate, or high in respondent and interviewer problems, based on behavior coding data and interviewer ratings	2. Items classified as high in respondent problems during pretesting also has many problems in the field. 3. Items classified as having recall and sensitivity problems during pretesting had higher nonresponse rates in the field.

Table 1. Continued

Paper	Methods tested	Criteria	Conclusions
			4. The revised items in the experimental questionnaire produced nonsignificant reductions in item nonresponse and problems found via behavior coding, but a significant reduction in respondent problems (as rated by the interviewers); however, interviewers rated revised items as having more interviewer problems.
DeMaio and Landreth (2004)	1. Three cognitive interview methods (three different "packages" of procedures carried out by three teams of researchers at three different organizations) 2. Expert review	• Number of problems identified • Type of problem identified • Technique that identified the problem • Frequency of agreement between organizations/methods	1. The different methods of cognitive interviewing identified different numbers and types of problems 2. Cognitive interviewing teams found fewer problem questions than expert reviews, but all three organizations found problems with most questions for which two or more experts agreed there was a specific problem 3. The problems identified by the cognitive interviewing teams were also generally found by the experts 4. Different teams used different types of probes 5. Cognitive interviews done on revised questionnaires found that only one team's questionnaire had fewer problems than the original
Jansen and Hak (2005)	1. Three-Step Test Interview (cognitive interviews with concurrent think-alouds followed by probes and respondent debriefing) 2. Expert review	• Number of problems found • Places in questionnaire where problems were found • Type of problem found	1. Three-step test interview identified more problems than expert reviews 2. Three-step test-interview identified unexpected problems stemming from non-standard drinking patterns and from local norms regarding drinking alcohol

A limitation on the comparison studies summarized in Table 1 is that it is rarely clearly evident whether the problems identified by a given technique actually reduce the validity or accuracy of the answers in surveys. As Groves and his colleagues note: "[The assumption is that] questions that are easily understood and that produce few other cognitive problems for the respondents introduce less measurement error than questions that are hard to understand or that are difficult to answer for some other reason" (Groves et al. 2009, p. 259). As a result, the problems detected by the qualitative methods should in theory be related to quantitative measures of response validity. Question problems identified by the qualitative methods could also be attributed to lower reliability of survey items if the problems are not systematic in their effects (for example, some respondents misinterpret the questions in one way while other respondents interpret the questions in another way).

Most of the studies in Table 1 compare several qualitative techniques to each other; this is unfortunate since the ultimate standards by which items should be judged are quantitative – whether the items yield accurate and reliable information. The study described here attempts to fill this gap in the literature. We compare results from both qualitative and quantitative assessments of a set of items, including estimates of item validity and reliability, and assess how well the conclusions from qualitative methods for question evaluation stack up against the conclusions from more direct quantitative estimates of validity and reliability.

As one reviewer noted, some question "problems" may not lead to response error but interrupt the flow of the interview. Both types of problems are typically addressed in the evaluation and pretesting process. We completely agree with this view, and for the remainder of this paper, we use the term "problem" to refer to suspected or purported problems identified by a given evaluation method without implying that these "problems" actually reduce the value of the data. Still, we believe that question evaluations are mainly done to ensure that the data that are ultimately collected are valid and accurate and that the main value of question evaluation methods is in improving data quality rather than improving the flow of the questions.

2. Comparing Five Evaluation Methods

The five methods we compare include two qualitative methods (expert reviews and cognitive interviews) and three quantitative methods (measures of validity and reliability and estimated error rates from latent class analysis). We chose expert reviews and cognitive interviews because they are popular methods for evaluating survey questions. We included latent class analysis because of its ability to estimate error rates without an external gold standard. And last but not least, we included validity and reliability because these are the ultimate standards a good item should meet.

We begin by describing each of these methods and reviewing the prior studies that have examined them; then in the Section 3, we describe how we compared them.

2.1. Expert Reviews

One relatively quick and inexpensive method for evaluating draft survey questions is to have experts in questionnaire design review them for problems. Not surprisingly, expert reviews have become a common practice in questionnaire development (Forsyth and Lessler 1991). As Willis et al. (1999) point out, expert reviews can be conducted individually or in group sessions. In addition, the experts can rely exclusively on their own judgments, making informal assessments that typically yield open-ended comments about the survey items to be evaluated, or they can be guided by formal appraisal systems that provide a detailed set of potential problem codes.

Four studies have examined the effectiveness of expert reviews, and they differ somewhat in their findings (see Table 1 for details). Two of the studies found that expert reviews identified more problems than other methods, such as cognitive interviews (Presser and Blair 1994; Willis et al. 1999), but Rothgeb and her colleagues (2001) reported that expert reviews identified roughly the same number of problems with questions as cognitive interviews, and that the two methods identified different questions as problematic. Finally, Jansen and Hak (2005) report that their three-step cognitive testing procedure found more problems than an expert review. The three-step variant on cognitive interviewing developed by Jansen and Hak (2005) begins with a concurrent think-aloud, follows that with probing the attempt to clarify observed during the think-aloud portion of the interview, and concludes with a debriefing interview to explore the respondent's problems in answering the questions. In these studies, there is no independent evidence that the "problems" identified by the experts or those found in cognitive interviews are, in fact, problems for the respondents in the survey. Expert reviews are especially likely to identify problems related to data analysis and question comprehension (Presser and Blair 1994; Rothgeb et al. 2001). In addition to turning up lots of potential problems, expert reviews are less expensive than cognitive interviews or behavior coding (Presser and Blair 1994).

2.2. *Cognitive Interviewing*

As we noted earlier, cognitive interviewing relies on verbalizations by respondents to identify problems with the questions. Even though cognitive interviewing has become popular among survey practitioners, there is little consensus about the exact procedures that cognitive interviewing encompasses or even about the definition of cognitive interviewing (Beatty and Willis 2007). Beatty and Willis (2007) offer a useful definition; cognitive interviewing is "the administration of draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends" (p. 288). They also noted that cognitive interviews have been carried out in various ways. Some cognitive interviewers use think-alouds (either concurrent or retrospective), but others rely mainly on probes (either scripted or generated on the fly by the interviewers) intended to shed light on potential problems in the response process.

The evidence regarding the effectiveness of cognitive interviewing is inconsistent (see Table 1). Some studies have found that cognitive interviews detect fewer problems than expert reviews (Jansen and Hak 2005; Presser and Blair 1994; Willis et al. 1999), but Rothgeb and colleagues (2001) found that the two methods identified about the same number of problems. Cognitive interviews may find more problems than behavior coding

(Presser and Blair 1994), or the opposite may be true (Willis et al. 1999). In addition, Presser and Blair (1994) found that cognitive interviews identified more problems than conventional pretesting. Rothgeb and colleagues (2001) showed that cognitive interviews detected fewer problems than the formal appraisal method.

Willis and Schechter (1997) carried out several experiments testing whether predictions based on cognitive interviewing results were borne out in the field, and concluded that the predictions were largely confirmed. Other studies show that cognitive interviewing produces reasonable consistency across organizations at least in the number of problems identified (Rothgeb et al. 2001; Willis et al. 1999), and Fowler and Roman (1992) claim there is reasonable agreement across two sets of cognitive interviews done by different organizations but do not attempt to assess the level of agreement quantitatively. The results of Presser and Blair (1994) are less reassuring; they argue that cognitive interviews were less consistent across trials than expert reviews or behavior coding in the number of problems identified and in the distribution of problems by type.

2.3. Reliability and Validity

Expert reviews and cognitive interviews generally produce only qualitative information, typically in the form of judgments (either by the experts or the cognitive interviewers) about whether an item has a problem and, if so, what kind of problem. Still, most survey researchers would agree that the ultimate test a survey question must meet is whether it produces consistent and accurate answers — that is, whether the question yields reliable and valid data. These quantitative standards are rarely employed to pretest or evaluate survey questions because they require the collection of special data. For example, the reliability of an item can be assessed by asking the same question a second time in a reinterview, but this entails carrying out reinterviews. Or validity might be assessed by comparing survey responses to some gold standard, such as administrative records, but that requires obtaining the records data and matching them to the survey responses.

The most common strategy for estimating the reliability of survey items is to look at correlations between responses to the same questions asked at two different time points, a few weeks apart (e.g., O'Muircheartaigh 1991). This method of assessing reliability assumes that the errors at the two time points are uncorrelated. As Saris and Gallhofer (2007, pp. 190–192) note, the correlation between the same item (say, y_1) administered on two occasions (y_{11} and y_{12}) is not a pure measure of reliability, but is the product of the reliabilities of the item at time 1 (r_{11}) and time 2 (r_{12}) and the correlation between the true scores over time (s):

$$\begin{aligned}\rho(y_{11}, y_{12}) &= r_{11} s r_{12} \\ &= r_1^2 s\end{aligned} \tag{1}$$

The equation simplifies if we assume that the reliability of the item remains the same across the two occasions; the result is shown in the second line of Equation 1 above. Since the stability over time (s) is a characteristic of the true score rather than of the items, it follows that ranking a set of items (y_1, y_2, y_3) that measure the same construct by their correlations with themselves over two occasions is identical to ranking them by their reliability. The major drawback of estimating reliability through over time correlations is the possibility of correlated errors in the test and retest due to learning or memory effects. Because we administered the items in different surveys conducted several weeks apart, we believe that any learning or memory effects are likely to have had only minimal impact on our ranking of the items by their test-retest reliability.

A simple approach for assessing the validity of survey items is to measure the correlations between each of the items to be evaluated and other questions to which they ought, in theory, to be related. Again, this is not a pure measure of validity (see Saris and Gallhofer 2007, p. 193). The correlation between an item of interest (y_1) and some other variable (x) is the product of the reliability (r_1) of y_1, its validity (v_1), and the true correlation (ρ) between the underlying constructs measured by x and y_1:

$$\rho(y_1, x) = r_1 v_1 \rho \tag{2}$$

However, as Equation 2 shows, because ρ is a property of the underlying constructs, ranking a set of items tapping the same construct by their correlation with some other variable is equivalent to ranking them by their overall accuracy – that is, by the product of the reliability (which reflects only random measurement error) and the validity (which reflects only systematic error).

Alternative measures of validity and reliability can be obtained using the SQP program of Saris and Gallhofer (Saris and Gallhofer 2007). Based on a meta-analysis of 87 multitrait-multimethod (or MTMM) experiments, the SQP program produces estimates of reliability, validity, and quality (a product of reliability and validity). Reliability is defined as one minus the random error variance over the total variance, and quality is defined as the proportion of the observed variance explained by the latent construct (Saris and Gallhofer 2007).

2.4. Latent Class Analysis (LCA)

As we already noted, latent class analysis is a statistical procedure that has been used to identify survey questions with high levels of measurement error. Proponents of the use of LCA in questionnaire development argue that it does not require error-free gold standards. Instead, it takes advantage of multiple indicators of the same construct and models the relationship between an unobserved latent variable (a.k.a., the construct) and the multiple observed indicators. The indicators are not assumed to be error-free. However, the errors associated with the indicators have to be independent conditional on the latent variable. This assumption – the local independence assumption – is almost always made in applications of LCA models. When this is satisfied, LCA produces unbiased estimates of the unconditional probabilities of membership in each of the latent classes (e.g., $P(c = 1)$

Table 2. Key parameters in latent class models

	Latent class	
Observed value	c = 1	c = 2
$u_1 = 1$	$P(u_1 = 1 \mid c = 1)$	$P(u_1 = 1 \mid c = 2)$
$u_1 = 2$	$P(u_1 = 2 \mid c = 1)$	$P(u_1 = 2 \mid c = 2)$
Unconditional probabilities	$P(\mathrm{c} = 1)$	$P(\mathrm{c} = 2)$

in Table 2 below). These unconditional probabilities represent the prevalence of each class in the population. LCA also produces estimates of the probability of each observed response conditional on membership in each latent class. For example, in a two-class model like the one in Table 2, the probability that a binary item u_1 is equal to 1 conditional on being in the first latent class ($c = 1$) is $p_{1|1} = P(u_1 = 1|c = 1)$, and the probability that this particular item is equal to 2 conditional on being in the first latent class is $p_{2|1} = P(u_1 = 2|c = 1)$.

Two of the conditional probabilities in Table 2 represent error rates. These are the probabilities of a false positive ($P(u_1 = 1|c = 2)$) and false negative response ($P(u_1 = 2|c = 1)$) to the question, given membership in latent class c. A high false positive or false negative probability signals a problem with a particular item. The primary purpose of applying LCA to the evaluation of survey questions is to identify questions that elicit error-prone responses – that is, questions with high rates of false positives or false negatives. When the local independence assumption is not satisfied (e.g., when the responses to three items measuring the same underlying construct are correlated even within the latent classes), then the LCA estimates of the unconditional and conditional probabilities may be erroneous.

Biemer and his colleagues have carried out several studies that use LCA to identify flawed survey questions and to explore the causes of the problems with these items (Biemer 2004; Biemer and Wiesen 2002). For example, Biemer and Wiesen (2002) examined three indicators used to classify respondents regarding their marijuana use and used LCA estimates to pinpoint why the multi-item composite indicator disagreed with the other two indicators. The LCA results indicated that the problem was the large false positive rate in the multi-item indicator (Biemer and Wiesen 2002).

A recent paper by Kreuter, Yan, and Tourangeau (2008) attempted to assess the accuracy of the conclusions from such applications of LCA. Kreuter and her colleagues conducted a survey of alumni from the University of Maryland that included several questions about their academic records at the university. They compared the survey answers to university records. They also fit LCA models to the survey responses and found that the LCA approach generally produced qualitative results that agreed with those from the comparison with the records data; the item that the LCA model singled out as having the largest estimated misclassification rate was also the one with the largest disagreement with the university records according to a traditional "gold standard" analysis. However, the quantitative estimates of the error rates from the LCA models often differed substantially from the error rates found in comparisons to the records data.

3. Research Design and Methods

In this study, we carried out two large-scale web surveys that allow us to measure the reliability of the answers for some of our items across two interviews (see Equation 1) and the construct validity of the items by examining the relation of each item to other questions in the same survey (as in Equation 2). We examined a total of fifteen items, five triplets consisting of items intended to measure the same construct. All fifteen items were assessed by four experts, tested in cognitive interviews, and investigated by latent class modeling. Six of the items were administered as part of a two-wave web survey that allowed us to measure both the reliability and construct validity of the items; the nine remaining items were administered in a one-time web survey, and we used the data from this survey to estimate the construct validity for these items.

3.1. Questions

The five triplets concerned a range of constructs — evaluations of one's neighbors, reading habits, concerns about one's diet, doctor visits in the past year, and feelings about skim milk.

One member of each of the triplets administered as part of a two-wave web survey was deliberately "damaged," that is, it was written so as to have more serious problems than the other two items in the triplet. For example, the neighborhood triplet asks respondents to evaluate their neighbors:

1a. How much do you agree or disagree with this statement? People around here are willing to help their neighbors. (Strongly agree, Agree, Disagree, Strongly Disagree)
1b. In general, how do you feel about people in your neighborhood?
  0.1. They are very willing to help their neighbors.
  0.2. They are somewhat willing to help their neighbors.
  0.3. They are not too willing to help their neighbors.
  0.4. They are not at all willing to help their neighbors.
1c. How much do you agree or disagree with this statement? People around here are willing to help other people. (Strongly agree, Agree, Disagree, Strongly Disagree)

The third item was written to be vaguer and therefore worse than the other two items. All fifteen items making up the five triplets are included in Appendix 1.

3.2. Expert Reviews

We asked four experts in questionnaire design to assess all fifteen items. Two of the experts were authors of standard texts on questionnaire design; the third has written several papers on survey questions and taught classes on questionnaire design; and the fourth was an experienced staff member of the unit charged with testing questions at one of the major statistical agencies in the United States.

We told the experts that we were doing a methodological study that involved different methods of evaluating survey questions but did not give more specific information about the aims of the study. We asked them to say whether each item had serious problems (and, if it did, to describe the problems) and also to rate each item on a five-point scale. The scale values ranged from "This is a very good item" (=1) to "This is a very bad item" (=5). We used the average of the four ratings of each item to rank order the items.

3.3. Cognitive Interviews

All fifteen of the items were tested in interviews carried out by five experienced cognitive interviewers at the Survey Research Center (SRC) at the University of Michigan. Three versions of the questionnaire were tested, each containing one item from each of the five triplets plus some additional filler items. Respondents were randomly assigned to get one version of the questionnaire. A total of 15 cognitive interviews were done on each version.

The respondents were adults (18 years old or older) recruited from the Ann Arbor area and paid $40 for participating. (Respondents were also reimbursed for their parking expenses.) The respondents included 22 females and 23 males. Sixteen were 18 to 34 years old; 15 were 35 to 49 years old; and 14 were 50 years or older. Thirty of the respondents were white; ten were African-American; and five characterized themselves as "Other." Fourteen had a high school diploma or GED; 25 had at least some college; and six had more than a four-year college degree.

The interviews took place at SRC's offices and were recorded. An observer also watched each interview though a one-way mirror. The cognitive interviewers asked the respondents to think aloud as they formulated their answers, administered pre-scripted "generic" probes (such as "How did you arrive at your answer?" or "How easy or difficult was it for you to come up with your answers?"; see Levenstein et al. 2007 for a discussion of such probes), and followed up with additional probing ("What are you thinking?" or "Can you say a little more?") to clarify what the respondents said or how they had arrived at an answer. (Our cognitive interviews thus included both concurrent probes and immediate retrospective probes.) After the respondent completed each item, both the interviewer and the observer checked a box indicating whether he or she thought the respondent had experienced a problem in answering the question. The interviewer and observer also indicated the nature of the problems they observed (that is, whether the problem involved difficulties with comprehension, retrieval, judgment or estimation, reporting, or some combination of these). We counted a respondent as having had a problem with an item if both the interviewer and the observer indicated the presence of a problem.

3.4. Web Surveys: Reliability, Validity, and LCA Error Rates

3.4.1. Web Survey Data Collection

The two first triplets (see the neighborhood triplet — items 1a-1c — and the triplet of book items — 2a-2c — in Appendix I) were administered as part of two web surveys that were conducted about five weeks apart. The six questions were spread throughout the questionnaires in the two surveys. Respondents who completed the first web survey were invited to take part in the second one. They were not told that the second survey had any relationship to the first. The second survey was the subject of an experiment described in detail by Tourangeau et al. (2009). Briefly, the invitation to the second survey and the splash page (i.e., the first web screen shown to respondents once they logged on) for that survey systematically varied the description of the topic and sponsor of the survey. (Neither of the experimental variables affected the items we examine here.)

A total of 3,000 respondents completed the first survey. Half of the respondents came from Survey Sampling Inc.'s (SSI) Survey Spot frame, and the other half were members of the e-Rewards web panel. Both are opt-in panels whose members had signed up online to receive survey invitations via e-mail. The response rate (AAPOR 1; see American Association for Public Opinion Research 2008) for the first wave of the survey was 4.1% among the SSI members and 14.8% among the e-Rewards members. A total of 2,020 respondents completed the second wave of the survey. The response rate (AAPOR 1) for the second wave was 61.1% for the SSI members and 73.7% for the e-Rewards panel.

The first wave of the survey was conducted from January 25, 2007, to February 1, 2007; the second wave, from March 2, 2007, to March 19, 2007.

The response rates for this survey, particularly for the first wave, were quite low, and neither panel from which the respondents were drawn is a probability sample of the general population. As a result, Tourangeau and his colleagues (Tourangeau et al. 2009) attempted to measure the effects of any selection and nonresponse biases on the representativeness of the responding panel members. They compared the respondents from each wave of the survey to figures from the American Community Survey (ACS) on sex, age, race, Hispanic background, and educational attainment. In both waves, the web respondents did not depart markedly from the ACS figures on age, race, or Hispanic background. The web samples did underrepresent persons who were 18 to 29 years old (members of this group made up 14 percent of the wave 1 sample and 11 percent of the wave 2 sample, versus 21 percent of the population according to the ACS) and overrepresented those who were 60 years and older (28 percent and 32 percent in waves 1 and 2 of our survey, versus 22 percent in the ACS). The web samples also overrepresented college graduates (50 percent and 52 percent in the two waves, versus 25 percent in the ACS) and underrepresented those with less than a high school education (1 percent in both waves, versus 14 percent in the ACS). Of course, there could still be biases in the results we present, unless the data are missing at random (MAR), conditional on these variables.

The items making up the final three triplets – the diet triplet (items 5a-5c), the doctor visits triplet (items 6a-6c), and the skim milk triplet (items 7a-7c; see Appendix I) – were administered as part of a one-time web survey completed by 2,410 respondents. Half of these respondents came from the SSI Survey Spot panel, and the other half were from the Authentic Response web panel. The response rate (AAPOR 1) was 1.9% among the SSI members and 16.5% among the members of the Authentic Response panel. The survey was carried out from September 2 to September 23, 2008.

Again, because the web sample was a non-probability sample and the response rate was low, we compared the demographic makeup of the respondents in our second study sample to that of the American Community Survey. The results were similar to those for our earlier web survey. The web respondents in the second study also tended to be more highly educated and older than the U.S. adult population as a whole; in addition, they were more likely to be white (89 percent versus 77 percent in ACS) and less likely to be Hispanic (4 percent of our web respondents versus 13 percent in the ACS) than the U.S. general population. Again, this does not demonstrate an absence of bias in the results we present.

The nine target questions were spread throughout the questionnaire in the second web survey, with one item from each triplet coming at the beginning of the survey, one coming in the middle, and one coming at the end.

3.4.2. Reliability and Validity

Because we intended to apply latent class models to each target item, we first recoded the responses to all fifteen target items to yield dichotomies. For example, with item 1b (the second item in the neighborhood triplet, see Appendix I), we combined the first two response options and the last two. The results presented below in Tables 3 through 5 do not differ markedly if we do not dichotomize the items offering more than two response options, but treat them as scales instead.

We computed reliabilities for the neighborhood and book triplets (the first six items in Appendix I). Our reliability estimate was the correlation between responses to the same question in the two waves (after recoding the answers to yield dichotomies). This is the same approach summarized earlier in Equation 1. Similarly, the validity coefficients were the correlations between the dichotomized responses to the items in each triplet with some other item in the questionnaire. For example, for the three neighborhood items (items 1a, 1b, and 1c above), we correlated dichotomized responses (in the initial interview) with answers to the first item in the wave 1 questionnaire, which asked for an overall assessment of the respondent's neighborhood (see Appendix I for detailed wordings of all the questions examined in this article). This is the same approach described earlier (see Equation 2).

3.4.3. LCA Error Rates

We fit latent class models (like the one summarized in Table 2 above) to the data from the three items in each triplet, using the Mplus software (Muthén and Muthén 1998–2007). We dichotomized each item prior to fitting the latent class models. For each triplet, we fit a model with two latent classes and estimated the false positive and false negative rates for each of the three items presented in Appendix II. In ranking the items in each triplet, we used the sum of the two error rates for each item and labeled it as 'misclassification rate' in Tables 3 and 4.

4. Results

Tables 3 and 4 present the main results from the study. Table 3 displays the summary statistics for the six items included in the two-wave web survey. It shows the mean ratings of the experts for each item (with higher ratings indicating a worse item), the proportion of cognitive interviews in which the item was found to have a problem, the misclassification rates from the latent class modeling, and the validity and reliability coefficients for each of the items. Table 4 displays similar summary statistics for the nine items included in the second web study. Because the second web study was a single-wave survey, we could not compute reliability estimates for the nine items in that survey.

Both tables also provide ranking of the items within each triplet and standard errors for the main statistics. For the statistics derived from the web survey data (that is, the reliability and validity coefficients and the error rates from the latent class models), we used the "random groups" approach to calculate the standard errors for the statistics themselves as well as for the differences between pairs of statistics (see Wolter 1985, ch. 2, for a detailed description of the random groups technique). We randomly subdivided the sample into 100 replicates and used the variation in the statistic of interest across replicates to estimate the standard error:

$$SE(\hat{\theta}) = \left[\frac{1}{k}\sum\frac{(\hat{\theta}_i - \bar{\theta})^2}{(k-1)}\right]^{1/2} \qquad (3)$$

where $\hat{\theta}_i$ is a statistic (such as a reliability coefficient) computed from replicate i and $\bar{\theta}$ is the mean of that statistic across all 100 replicates. We also used the random groups

Table 3. Indicators of item quality (and ranks), by item — Study 1

	Expert reviews		Cognitive interviews		LCA model error rates			Validity			Reliability		
	Mean rating (higher is worse)	SE	% with problems	SE	Full sample estimate	Mean across replicates	SE	Full sample estimate	Mean across replicates	SE	Full sample estimate	Mean across replicates	SE
Neighborhood items													
Item 1a	4.25 (1)	.48	26.7 (1)	11.8	.092 (1)	.088	.015	.318 (2)	.313	.028	.449 (2)	.449	.030
Item 1b	4.50 (1)	.29	21.4 (1)	11.4	.189 (3)	.158	.021	.341 (1)	.345	.030	.566 (1)	.599	.034
Item 1c	4.25 (1)	.25	40.0 (2)	13.1	.183 (2)	.145	.018	.322 (2)	.317	.026	.549 (1)	.550	.031
Book items													
Item 2a	3.75 (1)	.63	46.7 (1)	13.3	.203 (2)	.196	.011	.227 (1)	.219	.026	.680 (1)	.672	.019
Item 2b	3.50 (1)	.65	50.0 (1)	13.9	.013 (1)	.016	.003	.226 (1)	.215	.023	.717 (1)	.706	.017
Item 2c	2.75 (1)	.85	46.7 (1)	13.3	.067 (1)	.060	.007	.231 (1)	.219	.024	.725 (1)	.724	.018

Table 4. Indicators of item quality (and ranks), by item — Study 2

	Expert reviews		Cognitive interviews		LCA model error rates			Validity		
	Mean rating (higher is worse)	SE	% with problems	SE	Full sample estimate	Mean across replicates	SE	Full sample estimate	Mean across replicates	SE
Diet items										
Item 5a	4.00 (1)	.48	60.0 (3)	13.1	.298 (1)	.304	.025	−.282 (2)	−.274	.023
Item 5b	4.50 (1)	.29	0.0 (1)	0.0	.468 (3)	.400	.028	−.404 (1)	−.405	.021
Item 5c	4.25 (1)	.25	13.3 (2)	9.1	.386 (2)	.349	.017	−.354 (1)	−.358	.020
Doctor visit items										
Item 6a	2.75 (1)	.48	46.7 (2)	13.3	.046 (1)	.038	.010	−.408 (1)	−.407	.011
Item 6b	3.00 (1)	.41	13.3 (1)	9.1	.042 (1)	.037	.005	−.419 (1)	−.412	.013
Item 6c	5.00 (2)	.00	46.7 (2)	13.3	.039 (1)	.035	.010	−.399 (2)	−.395	.012
Skim milk items										
Item 7a	4.25 (2)	.25	20.0 (1)	10.7	.262 (3)	.246	.015	−.207 (1)	−.215	.018
Item 7b	2.25 (1)	.63	57.1 (2)	13.7	.038 (1)	.043	.007	−.194 (1)	−.208	.018
Item 7c	3.50 (2)	.65	60.0 (2)	13.1	.061 (2)	.060	.008	−.172 (2)	−.184	.018

technique to estimate differences between pairs of statistics (e.g., between the reliabilities of items 1a and 1b). Because each evaluation method yields results on different metrics, we rank order the questions based on their performance on each method. These ranks ignore "small" differences, which we defined somewhat arbitrarily as differences of one standard error or less. These ranks are displayed in Tables 3 and 4 in parentheses.

For the neighborhood items (items 1a, 1b, 1c in Table 3), the validities of the items are quite similar, but item 1a seems to have the lowest reliability. The experts seem to agree with these quantitative results; they rated the items as not very different from each other and saw all three items as problematic. The latent class model picks out item 1a as having the *lowest* misclassification rate of the three items; that item was also the least reliable item. Cognitive interviewing was the only method that picked out the damaged item (item 1c) as worse than the other two items.

All three items in the book triplet (items 2a, 2b, 2c in Table 3) had similar estimated validities and also similar estimated reliabilities. Cognitive interviews and expert reviews do not find much difference between the three items in this triplet. The LCA model identifies item 2a as having the highest misclassification rate among the three items. None of the five methods picked out the damaged item (item 2c) as worse than the other two items.

For the diet items (items 5b, 5b, and 5c in Table 4), both the validity analysis and the cognitive interviews indicate that items 5a is the weakest item among the three, whereas the LCA picks it out as the best member of the set.

For the doctor visit items (items 6a, 6b, and 6c in Table 4), the experts agree with the validity analysis in finding 6c the weakest item in this triplet. The LCA method, however, did not seem to find much difference between them. Cognitive interviews produced the opposite conclusions, identifying item 6b as the best item.

Expert reviews and the LCA method both ranked item 7b as the best in this triplet on skim milk (items 7a, 7b, and 7c in Table 4). By contrast, cognitive interviews and the validity measure favored item 7a over the other two.

So far, we have considered only how the different methods rank order the items within each triplet; this corresponds with how a questionnaire designer might make a decision about the items in a given triplet. Table 5 presents a quantitative assessment of the agreement across methods; the table shows the matrix of correlations among the mean expert ratings, the proportion of cognitive interviews in which both the interviewer and observer thought the item exhibited problems, the misclassification rates from the LCA models, and the estimates of quality obtained from SQP predictions provided by Dr. Willem Saris. (We drop the reliability estimates from this analysis since they are available only for six of the items.) It is reasonable to compare the validity estimates used in Tables 3 and 4 within triplets, but across triplets the comparisons are confounded with strength of the underlying relationship between the construct tapped by our three items and the construct we are trying to predict (see Equation 2, presented earlier, where this relationship is represented by ρ). We therefore include the correlations of the other methods with a statistic we call the validity ratio in Table 5. The validity ratio is just the ratio between the validity estimate for a given item within a triplet and the lowest validity estimate for the items in that triplet. This ratio renders the correlations across triplets more comparable by removing the effect of ρ. Italicized entries in the table take the opposite of

Table 5. Correlations (and number of items) among quantitative indicators of item quality

				Validity analysis		
	Expert review	Cognitive interviews	Latent class model	Validity estimate	Validity ratio	Quality
Expert rating	–	− _.408_ (15)	.526 (15)*	_.326_ (15)	_.230_ (15)	_.608 (15)_*
Cognitive interview		–	− _.570_ (15)*	− .560 (15)*	− .715 (15)*	− .070 (15)
Latent class model			–	_.201_ (15)	_.757_ (15)*	_.369 (15)_
Validity analysis						.063 (15)

Note: $*$ indicates the $P < .05$ (two-tailed). The indicator from the expert review was the mean rating of the item across the four experts; for the cognitive interviews, it was the proportion of interviews in which both coders judged the item to have a problem; for the latent class analysis, it was the misclassification rate; for the validity analysis, the validity estimate refers to the correlation of the item with a conceptually related item as used in Tables 3 and 4; validity ratio is the ratio between the validity estimate for a given item within a triplet and the lowest validity estimate for the items in that triplet; and, for the quality measure, it was the prediction from the SQP program (provided by Dr. Willem Saris). Italics indicate that the entry takes the opposite of the direction expected.

the direction expected. Just to be clear, we expected the validity estimates, the item reliabilities, and the quality measure from the SQL model to be positively correlated with each other. These measures are all quantitative measures of item quality, with higher numbers indicating a "better" item. Similarly, we expected the expert ratings, the proportion of cognitive interviews finding a problem with an item, and the misclassification rates to be positively correlated with each other, since they all measure the degree to which an item has problems. Finally, the measures in the first group should correlate negatively with those in the second group.

As Table 5 makes clear, the correlations are not very high and several of them go in the wrong direction. The indicators seem to fall into two groups. The expert ratings show good agreement with the LCA misclassification rates. The correlation between the mean of the expert ratings and the misclassification rates from the LCA models was significant ($r = .526$, $p < .05$, based on $n = 15$ items). The cognitive interviews and the validity analyses also produce converging conclusions. The correlation between the proportion of interviews in which a problem was found with an item and the validity coefficient for the item was significant and, as expected, negative (a higher rate of problems found in the cognitive interviews was associated with lower validity estimates; $r = -.560$, $p < .05$); this correlation increases to $-.715$ when we use our validity ratio statistic in place of the original validity estimates.

There are two other significant correlations in the table and both are in the wrong direction. The correlation between the LCA misclassification rates and the proportions of cognitive interviews in which a problem was observed with an item was significant but negative ($r = -.570$, $p < .05$) – the higher the proportion of cognitive interviews revealing problems with the item, the lower the misclassification rate according to the LCA models. The LCA error rates also are significantly correlated (in the wrong direction) with our validity ratio statistic. The correlation between expert ratings and the quality measure was significant but positive ($r = .608$, $p < .05$) – the higher the experts' ratings (and the worse the items), the higher the predicted quality according to SQP.

5. Conclusions and Discussion

This article examined a variety of question evaluation methods. As the studies reviewed in Table 1 might suggest, the methods generated different results, giving inconsistent, even contradictory, conclusions about the items in a triplet. As shown in Table 5, even though we find considerable agreement with the expert ratings and the LCA results and the cognitive interview results and the validity analysis, most of the correlations among the indicators generated by each method take the opposite of the direction expected (see Table 5).

Why are the results not more consistent across different methods? One possibility is that the methods do not all give valid indications of problems with the items. All of the methods make assumptions, and these assumptions may often be violated in practice. In an earlier paper examining the use of LCA models to evaluate survey items, Kreuter et al. (2008) found that the LCA models often gave good qualitative results (e.g., correctly identifying the worst item among a set of items designed to measure the same construct) but were substantially off in their quantitative estimates of the error rates. LCA models make strong assumptions and their results seem to be sensitive to the violation of those assumptions (e.g., Spencer 2008). The data here suggest they are not a substitute for direct estimates regarding item validity. Of course, the validity estimates we present are hardly perfect or assumption-free either; as we noted, they reflect both the properties of the items and the strength of the underlying relationship between the relevant constructs.

The more qualitative methods may be especially prone to yielding unreliable or invalid conclusions. As Presser and Blair (1994) first demonstrated, multiple rounds of expert reviews and cognitive interviews often yield diverging conclusions. More recently, Conrad and Blair (2004) have found that cognitive interviews may be prone to false positives in question evaluation, evidenced by the high percentage of items found to have problems (see also Levenstein et al. 2007 and Conrad and Blair 2009). Our results indicate some convergence between the cognitive interview results and the validity estimates. This was true even though the consistency across cognitive interviewers was quite low. Three of the cognitive interviewers did seven or more cognitive interviews and, for these three, we calculated the proportion of interviews in which a problem was found with each item. The correlations in these proportions across the fifteen items ranged from only .143 to .326. (The convergence across experts was a little higher; the median correlation in the expert ratings was .360). The relatively low agreement across cognitive interviewers and across experts may put a low ceiling on their convergence with quantitative measures of item performance such as the validity and reliability measures used here.

Another possible reason for the low consistency across methods is the low agreement among question evaluation methods about the *nature* of the problem. We calculated the proportion of the experts who saw each item as presenting a comprehension problem, a recall problem, or a problem with judgment or reporting, and we correlated these proportions with the proportion of cognitive interviews in which the interviewer indicated there was a problem of the same type. (Problems in judgment and reporting were relatively rare, which is why we combined those categories.) The correlations were −.09 and −.33 for comprehension and judgment/reporting problems; the correlation was .86 for recall problems. This picture does not change much if we look at the proportion of the time the observers of the cognitive interviews indicated that there was a problem of a given type; the correlations are very similar (.03 for comprehension problems, −.47 for judgment or reporting programs, and .80 for recall problems).

Thus, one potential source of the conflicting conclusions about an item is that the different question evaluation methods focus on different aspects of the questions and different types of problems. As one reviewer pointed out, the experts and the cognitive methods tend to concentrate more on how well the underlying constructs are measured and somewhat less on the response scales. By contrast, the latent class methods focus on the probabilities of errors and marginal distributions of responses whereas the quality measures from the SQP predictions emphasize purely on the effects of the form of the questions and the response scales.

Whatever the reason for the diverging results across question evaluation methods, until we have a clearer sense of which methods yield the most valid results, it will be unwise to rely on any one method for evaluating survey questions (cf. Presser et al. 2004a). Most textbooks advocate applying more than one evaluation method in testing survey questions, and our results indicate that a multi-method approach to question evaluation may be the best course for the foreseeable future. The natural next steps for this research are to understand how to reduce the inconsistencies and to investigate how to best combine different evaluation methods while capitalizing on the strengths of each. We believe that there is no substitute for the traditional psychometric indicators and we recommend that more questionnaire evaluation studies include validity and reliability measures. This may be expensive, but there seems to be no low-cost qualitative substitute for these indicators of item quality. We believe that the methods used to evaluate survey questions should have a firmer scientific basis and, in our view, more studies with credible estimates of the validity and reliability of the items are needed if we are ever to understand how much confidence we can place on the different qualitative methods currently used to evaluate survey questions.

Appendix I: Items Used in the Study

Items included in two-wave web survey

Neighborhood Triplet

1a. How much do you agree or disagree with this statement? People around here are willing to help their neighbors. (Strongly agree, Agree, Disagree, Strongly Disagree)

1b. In general, how do you feel about people in your neighborhood?

1. They are very willing to help their neighbors.
2. They are somewhat willing to help their neighbors.
3. They are not too willing to help their neighbors.
4. They are not at all willing to help their neighbors.

1c. How much do you agree or disagree with this statement? People around here are willing to help other people. (Strongly agree, Agree, Disagree, Strongly Disagree)

Book Triplet

2a. Which, if any, of the following have you done in the past 12 months? . . . Read more than five books? (Yes, No)

2b. During the past year, how many books did you read?

2c. During the past year, about how many books, either hardcover or paperback, including graphic novels, did you read either all or part of the way through?

Question used in validity estimates for Neighborhood triplet (1a, 1b, and 1c)

3. The first few questions are about some general issues. First, how would you rate your neighborhood as a place to live? (Poor, Fair, Good, Very Good, Excellent)

Question used in validity estimates for Book triplet (2a, 2b, and 2c)

4. What is the highest level of education you’ve completed? (Grades 1 through 8, Less than High School Graduate, High School Graduate, Some college/Associates’ degree, College graduate, Master’s degree, Doctoral/Professional degree)

Items included in final web survey

Diet Triplet

5a. On a scale of 0 to 9, where 0 is not concerned at all and 9 is strongly concerned, how concerned are you about your diet? (Nine-point scale, with labeled endpoints)

5b. Would you say that you care strongly about your diet, you care somewhat about your diet, you care a little about your diet, or you don't care at all about your diet? (Strongly, Somewhat, A little, Not at all)

5c. Do you worry about what you eat or do you not worry about it? (Worry about what I eat; Do not worry about what I eat)

Doctor Visit Triplet

6a. The next item is about doctor visits — visits to a physician or someone under the supervision of a physician, such as a nurse practitioner or physician's assistant for medical care. During the last 12 months — that is, since [INSERT CURRENT MONTH] of 2007 — how many times have you visited a doctor? (Open-ended answer)

6b. Over the last 12 months, how many times have you seen a doctor or someone supervised by a doctor for medical care? (Open-ended answer)

6c. How many times have you seen a doctor over the past year? (0 times; −2 times; 3–4 times; 5–6 times; 7 or more times)

Skim Milk Triplet

7a. Please indicate how you feel about the following foods . . . . Apples; Whole milk; Skim milk; Oranges (These items appeared in a grid, with a ten-point response scale; the end points of the scale were labeled "Like Very Much" and "Dislike Very Much")

7b. How much would you say you like or dislike skim milk? (Like very much; Like somewhat; Neither like nor dislike; Dislike somewhat; Dislike very much)

7c. How much would you say you agree or disagree with the statement "I like skim milk." (Agree strongly; Agree somewhat; Neither agree nor disagree; Disagree somewhat; Disagree strongly)

Question used in validity estimates for the Diet triplet (5a, 5b, 5c) and Skim Milk triplet (7a, 7b, 7c)

8. Indicate how much you favor or oppose each of the following statements . . . . "Maintaining healthy diet" (Strongly oppose, Somewhat oppose, Neither favor nor oppose, Somewhat favor, Strongly favor)

Question used in validity estimates for the Doctor Visit triplet (6a, 6b, and 6c)

9. How many different PRESCRIPTION DRUGS are you currently taking? (None, 1, 2, 3, 4, 5 or more)

Appendix II: False Positive and False Negative Rates, by Triplet and Item

	Neighborhood items (Triplet 1)		Book items (Triplet 2)		Diet items (Triplet 3)		Doctor visit items (Triplet 4)		Skim milk items (Triplet 5)	
	False positive	False negative	False positive	False negative	False positive	False negative	False positive	False negative	False positive	False negative
Item a	0.052	0.040	0.176	0.027	0.244	0.054	0.037	0.028	0.114	0.148
Item b	0.184	0.005	0.002	0.011	0.450	0.018	0.011	0.041	0.013	0.025
Item c	0.152	0.031	0.055	0.012	0.021	0.365	0.026	0.051	0.028	0.033

6. References

The American Association for Public Opinion Research (2008). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, (5th edition). Lenexa, Kansas: AAPOR.

Beatty, P.C. and Willis, G.B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. Public Opinion Quarterly, 71, 287–311.

Biemer, P.P. (2004). Modeling Measurement Error to Identify Flawed Questions. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 225–246.

Biemer, P.P. and Wiesen, C. (2002). Measurement Error Evaluation of Self-Reported Drug Use: A Latent Class Analysis of the US National Household Survey on Drug Abuse. Journal of the Royal Statistical Society, Series A, 165, 97–119.

Biemer, P.P. and Witt, M. (1996). Estimation of Measurement Bias in Self-Reports of Drug Use with Applications to the National Household Survey on Drug Abuse. Journal of Official Statistics, 12, 275–300.

Conrad, F.G. and Blair, J. (2004). Aspects of Data Quality in Cognitive Interviews: The Case of Verbal Reports. In Questionnaire Development, Evaluation and Testing Methods, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 67–88.

Conrad, F.G. and Blair, J. (2009). Sources of Error in Cognitive Interviews. Public Opinion Quarterly, 73, 32–55.

Converse, J.M. and Presser, S. (1986). Survey Questions: Handcrafting the Standardized Questionnaire. Beverly Hills, CA: Sage.

DeMaio, T. and Landreth, A. (2004). Do Different Cognitive Interview Techniques Produce Different Results? In Methods for Testing and Evaluating Survey Questionnaires. S. Presser et al. (eds), pp. 891-08. Hoboken, NJ: John Wiley and Sons.

Ericsson, K.A. and Simon, H.A. (1980). Verbal Reports as Data. Psychological Review, 87, 215–257.

Forsyth, B.H. and Lessler, J.L. (1991). Cognitive Laboratory Methods: A Taxonomy. In Measurement Errors in Surveys, P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds). New York: John Wiley, 393–418.

Forsyth, B., Rothgeb, J., and Willis, G. (2004). Does Questionnaire Pretesting Make a Difference? An Empirical Test Using a Field Survey Experiment. In Questionnaire Development, Evaluation, and Testing, S. Presser, et al. (Eds.), pp. 525-546. Hoboken, NJ: John Wiley and Sons.

Fowler, F.J. (2004). The Case for More Split-Sample Experiments in Developing Survey Instruments. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 173–188.

Fowler, F.J. and Roman, A.M. (1992). A Study of Approaches to Survey Question Evaluation, Final Report for U.S. Bureau of the Census, Boston: Center for Survey Research.

Gerber, E. (1999). The View from Anthropology: Ethnography and the Cognitive Interview. In Cognition and Survey Research, M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds). New York: Wiley, 217–234.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2009). Survey Methodology. New York: Wiley.

Jansen, H. and Hak, T. (2005). The Productivity of the Three-Step Test-Interview (TSTI) Compared to an Expert Review of a Self-Administered Questionnaire on Alcohol Consumption. Journal of Official Statistics, 21, 103–120.

Kreuter, F., Yan, T., and Tourangeau, R. (2008). Good Item or Bad – Can Latent Class Analysis Tell? The Utility of Latent Class Analysis for the Evaluation of Survey Questions. Journal of the Royal Statistical Society, Series A, 171, 723–738.

Levenstein, R., Conrad, F., Blair, J., Tourangeau, R., and Maitland, A. (2007). The Effect of Probe Type on Cognitive Interview Results: A Signal Detection Analysis. In Proceedings of the Section on Survey Methods, 2007. Alexandria, VA: American Statistical Association, 3850–3855.

Loftus, E. (1984). Protocol Analysis of Responses to Survey Recall Questions. In Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines, T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau (eds). Washington, DC: National Academy Press.

Maynard, D.W., Houtkoop-Steenstra, H., Schaeffer, N.C., and van der Zouwen, J. (2002). Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview. New York: John Wiley and Sons.

McCutcheon, A.L. (1987). Latent Class Analysis. Newbury Park, CA: Sage.

Miller, K. (2009). Cognitive Interviewing. Paper presented at the Question Evaluation Methods Workshop at the National Center for Health Statistics.

Muthén, L.K. and Muthén, B.O. (1998–2007). Mplus User's Guide, (Fifth Edition). Los Angeles, CA: Muthén & Muthén.

O'Muircheartaigh, C. (1991). Simple Response Variance: Estimation and Determinants. In Measurement Error in Surveys, P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds). New York: John Wiley, 551–574.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J., and Singer, E. (2004a). Methods for Testing and Evaluating Survey Questions. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 1–22.

Presser S., Rothgeb J., Couper M.P., Lessler J.T., Martin E., Martin J., Singer E. (eds) (2004b). Methods for Testing and Evaluating Survey Questionnaires. New York: John Wiley.

Presser, S. and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? Sociological Methodology, 24, 73–104.

Rothgeb, J., Willis, G., and Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results. Proceedings of the Section on Survey Methods (2001). Alexandria, VA: American Statistical Association.

Saris, W.E. and Gallhofer, I.N. (2007). Design, Evaluation, and Analysis of Questionnaires for Survey Research. New York: John Wiley.

Schaeffer, N.C. and Presser, S. (2003). The Science of Asking Questions. Annual Review of Sociology, 29, 65–88.

Spencer, B.D. (2008). When Do Latent Class Models Overstate Accuracy for Binary Classifiers? Unpublished manuscript.

Tourangeau, R., Groves, R., Kennedy, C., and Yan, T. (2009). The Presentation of the Survey, Nonresponse, and Measurement Error. Journal of Official Statistics, 25, 299–321.

van der Zouwen, J. and Smit, J.H. (2004). Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Question-Answer Sequences: A Diagnostic Approach. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley, 109–130.

Willis, G.B. (2005). Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA: Sage.

Willis, G.B. and Schechter, S. (1997). Evaluation of Congitive Interviewing Techniques: Do the Results Generalize to the Field? Bulletin de Methodologie Sociologique, 55, 40–66.

Willis, G.B., Schechter, S., and Whitaker, K. (1999). A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What do They Tell US? In Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association. 28–37.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Received March 2010
Revised March 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 531–535

Discussion

Jennifer H. Madans[1] *and Paul C. Beatty*[2]

Over the last several decades, questionnaire evaluation has become an increasingly prominent component of methodological work aimed at maximizing the quality of survey data. Question evaluation methods are among the most important tools survey methodologists have for describing and improving data quality, but these methods generally require a great deal of investigator interpretation, which makes them difficult to use. The statistical methods available to evaluate sample errors are straightforward when compared to the methods used to evaluate survey questions. Questionnaire evaluation methods also range from the most qualitative to the most quantitative, and require a wide range of expertise.

The questionnaire evaluation methods reviewed in the article by Yan, Kreuter, and Tourangeau, considered together, show the breadth of efforts in this area. As they note, the methods vary widely in their assumptions, implementation, and nature of the data that they produce–and indeed, very different sets of knowledge and skills would be needed to utilize them. For example, expert review presumably requires extensive knowledge of questionnaire design literature, experience crafting questions, and ability to make qualitative judgments; latent class analysis and structural equation models require little of these, but require sophisticated and specific quantitative skills. Given the broad differences in these approaches, and the fact that few survey methodologists are likely to be proficient in all of them, attempts to compare them and evaluate their respective contributions are most welcome. Such comparisons have the potential not only to expand and improve the application of these methods, but also to serve as an impetus for further methodological research. Yan, Kreuter and Tourangeau provide a very useful overview of prominent questionnaire evaluation methods, but the article also illustrates just how difficult it is to compare these methods. The way that specific techniques are used, and the way that results are summarized and combined, can greatly affect any conclusions about the quality of the questions and the data that are produced. Question evaluation methods provide crucial information on data quality, both to improve the quality of data collections and to inform users of existing data. However, they can only be effective if used in a way that provides credible evidence that itself can be evaluated by data users and question developers. This requires that the methods are described and used in a manner that is as transparent as possible.

Comparative methodological evaluation studies such as this one run the risk of oversimplifying the purpose of the methods. In this study and others, there is an implicit

[1] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20,782, U.S.A. Email: jhm4@cdc.gov
[2] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20,782, U.S.A. Email: pbb5@cdc.gov

dichotomy between questions that work, and those that have problems that need to be addressed. Presumably, the effectiveness of a method is linked to its ability to identify these problems; methods that identify more problems (or at least, more genuine problems) are considered to be more effective. On the face of things, this does not seem like a particularly controversial set of assumptions. Clearly, questionnaire evaluation often identifies various question flaws, such as ambiguous terms, inappropriate response categories, or overly challenging response tasks, and questions are rewritten to eliminate the problems.

But we suggest that the process of questionnaire evaluation is often more complicated than finding and fixing problems, a paradigm which suggests that there is an "ideal" way to ask a question. The art of question design involves obtaining information about complex concepts through a very limited interaction with a respondent, using questions that the respondent might or might not be paying close attention to. As a result, every question has some degree of imperfection–for example, ambiguity in the way some concept is described. A revision may reduce this ambiguity through adding clarifying details, but these details may add confusion for respondents who would not have had trouble with the original question. Similarly, a term may be problematic for some respondents, and an alternative might be simpler for them, but lack specificity that others need. Questionnaire evaluations should identify the strengths and weaknesses of particular ways of asking a question, and helps conscientious social scientists understand the tradeoffs involved in the various alternatives they could select. Question evaluation should move toward this paradigm to optimize the chances that the appropriate information will be captured. As it is not possible to design questions that mean the same thing to all respondents, or to tap the exact concepts that the researcher desires, evaluation techniques must not only identify problems, but must also provide information to users about what the question means to respondents. Hopefully this will maximize the likelihood that the question will obtain the information desired by its author.

Unfortunately, attempts to quantify this sort of contribution generally fail to capture the nuances of how evaluation methods help to make questionnaire design decisions. Understandably, researchers conducting such evaluations rely on what they can actually measure, such as counts of problems. But quantifying problems is not a very useful metric. For one thing, it requires an operational definition of a problem. Counts also generally assume that problems are of equal weight–but clearly some problems are minor imperfections, while others threaten the usefulness of any data generated by the question. Perhaps more importantly, quantifying problems fails to capture the level of insight produced by various methods. Yan, Kreuter and Tourangeau themselves note that researchers commonly suggest that the main value of cognitive interviewing is that it produces qualitative insight into the fit between the question and the concept it is trying to measure (cf., Beatty and Willis 2007; Miller 2011). Yet many methodological studies, including the current one, evaluate cognitive interviewing in terms of whether it flags the presence of "a problem."

In our view, reducing the output of qualitative methods such as cognitive interviewing in this manner is not only artificial, but undervalues their potential contributions. Similarly, expert reviews may provide rich assessments of which characteristics of questions are likely to lead to particular errors, and here, such insights are only summarized as simple quality ratings. Admittedly, it is hard to quantify the insights gleaned from such methods in a way that makes them amenable to comparative research. Still, it is problematic to criticize the value of these approaches when the contributions have been reduced to a few variables that do not really represent their contributions to the questionnaire design process.

Evaluating methods that produce qualitative insights is difficult for other reasons as well. Cognitive interviewing, in particular, is practiced in a variety of forms and with varying degrees of expertise. For example, some variants place strong emphasis on "thinking aloud" with minimal interviewer intervention, while others rely heavily upon probing – sometimes prescripted, sometimes determined based on interviewer content. But inevitably, methodological studies must define the practice of cognitive interviewing in a particular way, and the evaluation can only really address the way that it is conducted at that time. In other words, results of an evaluation don't generalize to "the method" – only the way the method was carried out in the particular study. More generally, it is difficult to perform evaluation of qualitative methods because comparisons require that the methods be standardized to some degree – otherwise, it is impossible to specify what exactly is being evaluated. However, this is problematic if one believes that a key strength of the method lies in its ability to adapt to issues that emerge in an interview in ways that would be difficult to predict in advance – in other words, its non-standardization. By standardizing the method, the researcher has compromised its strength and introduced a high degree of artificiality. While the authors are transparent about the approach and assumptions taken in the cognitive interviews within their study, it is very difficult to say anything conclusive about the overall value of "cognitive interviewing" as a method because the method, researchers and particular questions can all be confounded in challenging ways. Hopefully, the development and adoption of best practices for conducting cognitive interviews and for analyzing and reporting results will greatly facilitate question evaluation.

For any evaluation method to be effective, it should produce *measurably better questions*. Insights that do not actually contribute to that goal may be interesting, but are ultimately irrelevant. Evidence regarding the quality of these insights should be generated through carefully designed studies that use appropriate techniques. Question validity is often used as the gold standard for comparing the results of various evaluation methods. Theoretically, methods that produce more *valid* questions would be demonstrably better than alternatives.

The problem is that in practice, true validity is unknown, and attempts to quantify it have numerous problems of their own. In many cases there really is no "gold standard" for comparison – and even if there is, obtaining it is often either difficult or expensive. Furthermore, while latent class analysis is useful for some things, it does not truly measure validity. An alternative is to use the more limited concept of *construct validity*, in which researchers examine correlations with items that should theoretically be related to the question. Unfortunately, such validity assessments are only as good as the external comparators used, which might not be tapping the intended concept. More importantly, being correlated with another item is not the same as actually measuring what is intended. Statements that questions have been "validated" are powerful, but must be used with great caution. Measures of validity need to be improved, and evaluations of validity should report findings in a way that the criteria used to measure validity are clearly defined.

For all of these reasons, we do not find it particularly surprising that the methods evaluated in this study did not produce the same results. The differences are partially attributable to the fact that methods naturally create different sorts of insights, which cannot be easily compared. They are also partially attributable to the fact that questions are not easy to rank in terms of quality, nor readily categorized as “good” or “flawed” – realistically, most questions are imperfect and multifaceted, better for some purposes than others. Furthermore, they are partially attributable to the fact that standardizing methods, and abstracting results into scoring measures, alters both the methods and the results that they produce. Although the authors stop short of concluding that any of the methods are “better” in an absolute sense, they do suggest that qualitative methods have more to prove than quantitative measures of reliability and validity. We suggest that such conclusions are in part based on assumptions about the comparability of measures that are difficult to support. In fact, we wonder whether the question “which approach is best?” is really the right one to ask. It is an advantage that the methods provide different types of information, as this provides richer evaluation.

Instead, it is very important to ask “what does each method contribute?” and “under which circumstances is each method likely to be useful?” Yan, Kreuter and Tourangeau’s article offers a helpful response to the first question through a solid review of the variety of evaluation methods currently available. Their analysis did not really address the second question, but could have through a different approach: rather than attempting to determine overall measures of the quality of each method, and thereby suggesting varying degrees of methodological value, they could have started with the assumption that each method was *likely* to produce different results. From there, it would be possible to examine the nature of evidence from each method, how each are used to draw conclusions, and what sorts of decisions are actually made as a result of each. Such an analysis would not need to assume that all of the methods produced results of equal quality – in fact, it could still conclude that methods produced results of limited worth, at least within the current study. But it would probably not lend itself to conclusions about the relative value of each method. Then again, we find such conclusions to be limited, for the various reasons described above. There will always be interest in combining findings using different methods, and in learning about questionnaire design in general from all methods. As findings will be very dependent on the methods used, those who undertake question evaluations need to be explicit about how tests are done and how evidence from the tests is summarized and evaluated. A lack of information on question behavior is major threat to data quality, but acting on information that does not accurately convey what is known is even more dangerous.

Whether subsequent researchers build from the approach taken by Yan, Kreuter and Tourangeau, or instead decide to pursue different strategies, we think it is useful for their analysis to be part of a larger discussion. Hopefully our reservations with their approach and findings serve a similar purpose and will be seen in that light. Question evaluation remains a vital component of survey quality, and it is clear that we do not yet know enough about the contributions of the various approaches that are available. It is certainly undesirable for methodologists to work without more knowledge about what these approaches do and do not accomplish, although it is also undesirable to draw unwarranted conclusions about their merits. As we have seen, comparative research is difficult, and we have much more work to do before definitive statements can be made about what each method produces and when it should be used.

References

Beatty, P. and Willis, G.B. (2007). The Practice of Cognitive Interviewing. Public Opinion Quarterly, 71, 287–311.

Miller, K. (2011). Cognitive Interviewing. In Question Evaluation Methods, J. Madans, K. Miller, A. Maitland, and G. Willis (eds). Hoboken, NJ: John Wiley & Sons.

Received October 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 537–551

Discussion
Evaluation Procedures for Survey Questions

Willem E. Saris[1]

In this article, different criteria for the choice of an evaluation procedure for survey questions are discussed. Firstly, we mention a practical criterion: the amount of data collection the procedures require. Secondly, we suggest the distinction between personal judgments and model-based evaluations of questions. Thirdly, we suggest that it would be attractive if the procedure could evaluate the following aspects of the questions: 1. The relationship between the concept to be measured and the question specified; 2. The effects of the form of the question on the quality of the question with respect to: a. the complexity of the formulation, b. the precision, c. possible method effects, d. many other characteristics; 3. The social desirability of some of the response categories. Besides that, it would be desirable if the procedure could indicate the effect of respondents lack of the knowledge about the topic on their answers. We compare 13 procedures for the evaluation of questions with respect to these criteria and will derive some conclusions from this overview.

1. Introduction

In their article, Yan, Kreuter and Tourangeau mention a number of papers which compare the results of different evaluation procedures for survey questions: Fowler and Roman (1992), Presser and Blair (1994), Willis, Schechter and Whitaker (2000), Rothgeb, Willis and Forsyth (2001, 2004), DeMaio and Landreth (2004), and Jansen and Hak (2005). In these papers, the following evaluation procedures are mentioned: expert panels, focus groups, cognitive interviews, behavioral coding, three-step procedure of Jansen and Hak, standard pretests with debriefing, Quaid, SQP, latent variable models like test-retest, factor analysis and LCA, quasi-simplex design and model, MTMM design and model.

We would like to add to this list “the three step procedure” developed by Saris and Gallhofer (2007), “scaling procedures” developed by many people (see, for example Torgerson 1958), and item response theory (see, for example Hambleton et al. 1991).

We are not aware of papers discussing the criteria that could be used to select procedures for the evaluation of survey questions. Therefore, in the following pages we would like to suggest such criteria.

The first criterion we would like to suggest is a practical one: what one has to do to be able to use the different procedures. In this context, we distinguish between approaches that can be used without any data collection, procedures which require a small data set and

[1] Universitat Pompeu Fabra, Research and Expertise Centre for Survey Methodology, Passeig Pujades, 1, 08003, Barcelona, Spain. Barcelona. Email: w.saris@telefonica.net

Acknowledgment: I am very grateful for the useful comments of my colleagues of RECSM on an earlier version of this article.

those that require a more or less complete survey. It is clear that this criterion will play a role in the choice of an evaluation procedure.

As a second criterion to choose between the different procedures for question evaluation, we would like to mention whether the procedure is based on personal judgments or on model-based evaluations. We think that this criterion should also play a role in the choice of procedure.

Finally, we would like to suggest as a criterion the possible aspects of questions that are evaluated by the different procedures. In this context, we think about the following aspects of the quality of questions: 1. The relationship between the concept to be measured and the question specified; 2. The effects of the form of the question on the quality of the question with respect to a. the complexity of the formulation, b. the precision, c. possible method effects, d. many other characteristics; 3. The social desirability of some of the response categories. Besides that, it would be desirable if the procedure could evaluate questions with respect to the fourth criterion: the effect of respondents lack of knowledge about the topic on their answers. The use of the last criterion will lead to the suggestion to use combinations of different procedures in the evaluation of questions, because they evaluate different quality aspects of questions.

First, we will classify the different procedures with respect to the first two criteria. Thereafter, we will discuss what quality aspects the different procedures evaluate, and finally, we will describe which quality criteria can be evaluated with the different evaluation procedures. Based on this overview, we will finally draw some conclusions.

2. Two Basic Characteristics of Evaluation Procedures

In Table 1 we have classified the different procedures with respect to the amount of data needed for the evaluation (practical) and the evaluation procedures used.

It is, of course, very attractive if no new data have to be collected for the evaluation of the questionnaire. By new data we mean that one has to collect responses for the questions one would like to evaluate. There are a few procedures which satisfy this criterion. That does not mean that no new information is collected. In some cases, one has to ask experts

Table 1. The classification of 13 question evaluation procedures with respect to two procedural characteristics

Practical criterion	Evaluation procedure	
For quality prediction	Personal judgment	Model based
Without new data	Expert panels Focus groups Three step procedure (Saris and Gallhofer)	Quaid SQP Scaling methods
With few new data	Cognitive interviews Behavioral coding Tree step procedure (Jansen and Hak)	Scaling methods Behavioral coding
With a large pilot or full study	Debriefing of pilots	Latent variable models Quasi-simplex design/model MTMM design/model

about their judgments. In other cases, one has to code characteristics of the questions to obtain information about the quality of the questions.

There are also procedures which do not need a full study of the questionnaire, only a limited data collection for the evaluation of the questions. This is typically the case for cognitive interviewing using the think-aloud procedure, behavior coding, or some scaling procedures.

Finally, there are procedures that require a rather large data collection, such as most model-based procedures mentioned in Table 1, but also the standard procedure of debriefing interviewers after a pilot study.

It will be clear that, in principle, approaches that do not require new data are more attractive than procedures which require a new data collection before the official fieldwork. However, it should also be clear that this cannot be the only criterion.

Another very attractive criterion is whether the procedure is based on personal judgments of experts, interviewers, or respondents, or on model-based evidence collected in a special study or collected in the past. All procedures presented in the left column of Table 1 are based in some way or another on personal judgment, while the procedures on the right are model-based, collected on the spot, or evidence built up in the past. The scaling methods can be based on prior empirical studies or new empirical studies.

The model-based procedures will be more reliable if studies are well done. The results of such studies will not depend on the judgment of the researcher, and so repetition of applications of such studies will lead to approximately the same results. This is not necessarily the case when the procedure is based on personal judgments. With the change of the judges one may get different results. This is, for example, one of the problems that is mentioned in the study of Yan et al.

Combining the two criteria, one would say that the procedures on the top right side seem very attractive because they do not need the collection of new data and are based on existing evidence. This conclusion, however, would be overly hasty because the attraction of the procedures also depends on what aspects of the quality of questions are evaluated by the approach. This issue will, therefore, be discussed in the next section.

3. The Quality Aspects Evaluated by the Different Procedures

In our opinion it would be attractive if the evaluation procedures could evaluate the following aspects of the questions: 1. The relationship between the concept to be measured and the question specified; 2. The effects of the form of the question on the quality of the question with respect to: a. the complexity of the formulation, b. the precision, c. possible method effects, d. many other characteristics; 3. The social desirability of some of the response categories; 4. The lack of knowledge about the issue.

3.1. The Relationship Between the Concept to Be Measured and the Question Specified

Although the issue of validity of questions has been mentioned in all methodology books, one of the most ignored issues in survey research is the relationship between the concept to be measured and the questions specified. In this context, Blalock (1968) and others make a distinction between concepts by postulation and concepts by intuition. For concepts by intuition, questions can be formulated for which it is obvious that they measure the concept of interest. For example, there is no doubt that the question "How satisfied are you with your job?" measures "Job satisfaction". However one can also measure job satisfaction by asking about the satisfaction with different aspects of a job like the salary, social contact, spare time etc. In that case, the concept "Job satisfaction" becomes a concept by postulation, because we define the concept by a combination of the satisfaction with respect to the different aspects of the job. Here, the concept by postulation is defined by the combination of different concepts by intuition.

In the case of a concept by postulation, one has to evaluate the quality of the measurement of the concept on the basis of the relationship between the indicators for the concepts by intuition and the quality of the questions for these indicators. In the case of a concept by intuition, the evaluation of the question is much simpler, because one only has to evaluate an obvious question for the concept. Nevertheless, even this simple task is often not performed well. One can very easily provide many examples of cases where people specify what they want to mention, but specify questions which do measure something different. Two examples from research follow here.

In our first example, the researchers suggested measuring the opinion about the "policy of income equality". In order to measure this concept, the same researchers suggested using the question:

> "*To what extent do you agree with the statement: The government should take care that people get a job?*"

This question does not measure income equality, but an opinion about a "policy concerning full employment".

The second example comes from another study where the idea is to measure the concept "interest in work". In that study, the researchers suggest asking:

> "*How frequently did you think last month that you are interested in your work?*"

In this question, it is assumed that people who are more interested in their work think more often that their work is interesting. That does not have to be true. Why don't they ask directly "how interested are you in your work"?

The problem is that the relationship between the variable to be measured and the responses to the question can be very weak, because of the effect of other variables on the responses.

We think that it would be attractive if procedures for the evaluation of questions could detect such differences in the operationalization. The problem is, however, that often the researchers do not even specify what they want to measure, but immediately specify the questions. In that case evaluation is not possible.

3.2. The Effects of the Form of the Question on the Quality of the Question

Besides the validity of a question, one should consider the consequences of the form of the question for the quality of the measure. There are many alternatives for evaluating the same question. The most common aspect evaluated by survey researchers is whether the questions are too complicated for the respondents. Besides that, one has to consider the precision of the scale and the effect of the specific method chosen. There are, however, many more aspects of the question which have consequences for quality, such as the presence of an introduction, labeling of the scale, the nonresponse option etc. Saris and Gallhofer (2007) distinguish more than 50 form characteristics of a single question. We cannot discuss them in detail. Here we will mention only the main factors starting with the complexity of the formulation.

a. The Complexity of the Formulation

The complexity of a question has to do with the unnecessary complexity of the formulation. Typical examples are: unnecessary linguistic complications such as superfluous lengthy words and sentences, or complex sentences using of subordinate clauses or complex grammatical forms. Such complexities, if not necessary, can cause confusion in the mind of the respondent and lead to uncertainty, which can cause random fluctuation in the answers.

b. Precision of the Measurement

With respect to precision, we have to make a distinction between measures for concepts by postulation operationalized using several indicators and measures for concepts by intuition which can be operationalized by a single question. In the former case, the quality depends indirectly on the quality of several questions, while the precision of a single question depends on the precision of the scale that is used, besides other characteristics. A large variety of scales is in use. Most common are 2-, 3-, 5-, 7- and 11-point scales. However, there are also procedures available using continuous scales, like magnitude estimation or line production or so-called visual analog scales.

c. The Effect of the Method Used

A lot of attention has been paid in psychological literature to the problem of “common method variance”. This CMV is a consequence of the fact that people may react in a specific way to a specific method consistently across questions. In that case, a correlation will occur between these variables. This correlation, caused by the reaction of the respondents to the method used, has no substantive meaning. In this context, the method can be the mode of data collection but it also can be a type of scale or another characteristic. If such a systematic effect exists, this may not only cause CMV but also invalidity in the responses, because the responses are not only affected by the opinion or attitude to be measured, but also by the reaction to the method used.

d. Other Form Characteristics

Besides these basic form characteristics, there are many other aspects of the form of a question which can have an effect. To mention some: presence of an introduction, or an instruction, or a show card, the labeling of the response alternatives, direction of the alternatives, etc. There are many specific studies that evaluate some of these characteristics (Schuman and Presser 1981, Andrews 1984, Scherpenzeel 1995, Tourangeau et al. 2000, Alwin 2007, Saris and Gallhofer 2007).

3.3. The Social Desirability of Some of the Response Categories

Social desirability also is a common concern of survey researchers. If respondents are affected in their choice of an answer category by the social desirability of the categories, this will lead to lack of validity because a different variable has an effect on the responses than the variable one would like to measure.

3.4. Lack of Knowledge of Respondents About the Topic

In many cases, questions are asked about topics which the respondent has never thought about. This means that the respondent creates an answer on the spot (Zaller 1992, Tourangeau et al. 2000). The respondent can do so on the basis of related information that is available in his/her mind. This automatic process will be based on the information which is most salient at that moment for the respondent. Therefore Zaller suggests that the responses of the same person can vary from one moment to the other. This expresses itself in a large random variation in the responses (see also Converse 1964).

4. Evaluation of the Different Procedures

In this section we want to describe the different procedures and the kind of results one can obtain with them.

4.1. Expert Panels

It is very common in survey research to ask colleagues to evaluate questions or even whole questionnaires. The researcher can ask the expert to give the evaluations without any structure, but he/she can also provide a formal appraisal system. In case of an evaluation without an appraisal system, the experts may make comments about the validity of the question, some form effects like complexity, the precision of the scale, and possible social desirability problems and knowledge problems, but they most likely will not give a detailed discussion of many possible characteristics of the questions and their consequences. In general, different people will provide comments on different aspects. This can be seen as an advantage of this procedure because in this way the information becomes more complete. On the other hand, one can also wonder about the significance of the remarks if some experts detect some problems while others do not see these problems.

The use of a formal appraisal system can avoid both problems, and one can get as detailed information as one would like. However, it is unlikely that an expert has sufficient knowledge of the consequences of the different choices to also give an evaluation of the effects on the quality of the question, let alone with respect to the effects of the combination of all these choices.

4.2. Focus Groups

In general, focus groups are used to determine how potential respondents interpret specific concepts which are used in a questionnaire. In this way one tries to check the validity of the questions for the concepts they want to measure. In focus groups, one can also detect that some questions are too complex or that the people have no knowledge of the topic in question. What this procedure cannot provide is information about the positive or negative effects of specific choices with respect to the form of the questions.

4.3. The Three-step Procedure of Saris and Gallhofer

Saris and Gallhofer (2007) developed a procedure to design survey questions of which they claim that it guarantees that the question measures what the researcher wants to measure. So this procedure is completely directed at the validity of the measures.

The first step in the process is the decision whether the variable one wants to measure is a concept by intuition or a concept by postulation. If it is the former, one can immediately proceed to the next step. If it is a concept by postulation, one has first to define the concept in concepts by intuition. This is, of course, a theoretical step which can only be evaluated by the researcher and the research community.

The second step is the specification of a statement for the chosen concepts by intuition. For this step, Saris and Gallhofer have specified production rules. One first has to decide what the concept is that one wants to measure: an evaluation, a feeling, a norm, a policy, a preference, or another concept, and what the object is. Having done so, the production rules can be used to generate assertions for the concept of interest. These production rules are based on linguistic knowledge (Koning and van der Voort 1997, Harris 1978, Givon 1984, Weber 1993, Graesser et al. 1994, Huddleston 1994, Ginzburg 1996, and Groenendijk and Stokhof 1997).

In the third step, the assertions can be transformed into requests for answers as they call it, because not all so-called questions in survey research are real questions. One can also use imperatives or assertions. Characteristic of all forms is that they require an answer.

The guarantee of validity in this approach comes from the procedures developed for steps two and three. Step one is a theoretical step.

While this three-step procedure is a production system, one can also use it to evaluate the quality of questions by comparing the existing question with the results expected when the three-step procedure was used, or by looking to see if the question specified has the characteristics that were expected for the concept of interest.

The limitation of this procedure is that it concentrates completely on the validity of the measures and no other aspect. So for more complete evaluations of questions, this procedure has to be combined with other methods.

4.4. Cognitive Interviews

The most common procedure of cognitive interviewing is that one asks potential respondents to think aloud while answering the questions. An alternative is that one asks the respondent to tell how he/she came to his/her answer after the answer was given. Whatever procedure is chosen, this procedure aims at detecting whether the respondent interprets the concepts in the question in the correct way, and therefore this procedure aims again at the evaluation of the validity of the questions. However, like in the focus group approach, one can also see whether the respondents have the knowledge to answer the question or whether the question is formulated in too difficult a manner. Furthermore, in this case one will not get much information about the form effects.

4.5. Behavioral Coding

Behavioral coding is another way to achieve the same information. In this case, the communication between the respondent and the interviewer is recorded and later checked for indications of misunderstandings by the respondent to a question, which should show themselves in discussion with the interviewer about the meaning of the question. This procedure can also be used to detect wrong behavior of the interviewer, but that is less relevant here.

4.6. Three-step Procedure of Jansen and Hak

This is a combination of different forms of cognitive interviewing, starting with a think-aloud step, followed by probing to clarify the understanding of the process and later a normal debriefing. Given that the basis is cognitive interviewing, we expect that this procedure also mainly provides information about the validity of questions and possibly also about lack of knowledge and the complexity of the formulation.

4.7. Standard Pretests With Debriefing

In large and important surveys, it is rather common to pretest the questionnaire before the official data collection in order to check whether there are any problems. The check on problems is mostly done by asking the interviewers about the problems they have encountered while interviewing. Because the interviewer is mainly concerned about the communication with the respondent, the information one gets from the interviewers is similar to that obtained by behavioral coding, i.e., the misunderstandings about the meaning of questions, complexity of the questions, and lack of knowledge about the issue at stake.

4.8. Quaid

Quaid is a computer program that can analyze questions with respect to several aspects of questions namely: unfamiliar technical terms, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax and working memory overload. These judgments are based on long term research with respect to readability of texts (Graesser et al. 1994, Graesser et al. 2000a, Graesser et al. 2000b). Most of these checks are directed at problems of the form of the question, especially, at the complexity of the question and answer formulation with exception of the checks on vague or ambiguous noun phrases and vague or imprecise relative terms which are directed at the precision of the formulation. The attraction of the program is that one has to introduce the text of the questions and after a limited time one gets the results of the analysis. A disadvantage is that the program can only analyze questions in English and that the number of checks are limited. Suggestions for extension of the program are made for example by Faaß et al. (2008).

4.9. Latent Variable Models Like Test-retest, Factor Analysis and LCA

All latent variable models evaluate the quality of different questions for measurement of a latent variable. The quality of the question is based on the strength of the relationship between this latent variable and the observed variable. The difference between the models arises from the type of data, continuous or discrete, and the assumptions made about the latent variables and the relationship between the observed and the latent variable. The latent variable is a variable which all observed variables have in common. Whether this variable is what the researcher was supposed to measure cannot be determined by this method. So the validity is difficult to determine. If the observed variables measure the same variable, the models can evaluate which form of the question provides more information about the concept measured by the latent variable. If the observed variables contain unique components, the latent variable is a concept by postulation defined by different observed concepts by intuition. In that case the strength of the relationship between the observed variables and the latent variable is a combination of the quality of the question and the strength of the relationship between the concept by intuition and the concept by postulation.

Given this description of these evaluation procedures it follows that these procedures mainly provide information about the effect of the form of the question, because these approaches cannot provide information about the validity of the measure nor about the social desirability or lack of knowledge about the topic.

A limitation of these procedures is that for each set of questions a separate study has to be done. This means that the results cannot be generalized across topics.

Another limitation is that these methods are difficult to apply as well on background variables. This design requires variations of the question for the same concept. These variations are rather difficult for background variables and simple behavioral questions. Therefore these questions should be evaluated in a different way as mentioned below (quasi simplex models).

An extra limitation is that these procedures are normally applied in such a way that method variance cannot be estimated. To detect method effects, one needs a special design: the MTMM design.

4.10. Quasi-simplex Design and Model

A procedure that can be used for evaluation of background variables and simple behavioral questions is the quasi-simplex design and model. In this design, the same question is repeated at least three times in a panel study. If these data are available, the so-called quasi-simplex model, allowing for change through time and measurement error at each point in time, can be used to estimate the quality of the question. This model has been used intensively by Alwin (2007) to evaluate many different questions. The quality of the question is in this case the explained variance in the observed variable by the latent variable. In Alwin (2007), valuable information about the quality of many questions tested in this way can be found.

Given the form of these experiments, we would say that this approach provides information about the quality of the form of the specific question. The procedure does not provide information about validity, the social desirability of some categories, or lack of knowledge.

The limitation of this approach is that its application to more subjective variables leads to problems for two reasons. The first is the assumption that the latent variable may change but only with a lag of one time point. This means that an opinion that plays a role in the first moment, not in the second moment but again in the third moment cannot be specified in this model. This leads to identification problems (Coenders et al. 1999). The second problem is that all random changes in the latent variables are included in the error term. That means, for example, that in a measure of happiness the mood of a person, which is part of the happiness, will be included in the error and not in the latent variable. This characteristic of the model leads to serious problems with respect to the estimation of the quality of the questions, as was discussed by van der Veld (2006).

Another limitation of this approach is that method effects are ignored, while in general the same method is used at all points in time. The model does not allow the estimation of this effect. For background variables that may not be a serious problem, but for opinion questions it may cause a problem.

4.11. MTMM Design and Model

The multitrait multimethod (MTMM) design for evaluation of measurement instruments requires that for at least three different latent variables, at least three different forms that are however the same across latent variables are presented to the respondents (Campbell and Fiske 1959). On the basis of this design, a correlation matrix of nine variables is obtained. Different MTMM models have been developed for this matrix, which are special cases of latent variable models. Corten et al. (2002) and Saris and Aalberts (2003) showed that the classical MTMM model (Andrews 1984) and the equivalent true score model (Saris and Andrews 1991) fit the best to these matrices. This approach allows the estimation of reliability (the complement of random error variance) and internal validity (the complement of method variance). For details of this approach and for experiments to evaluate single questions, we refer to Saris and Gallhofer (2007). For evaluation of measures of concepts by postulation, we refer to Cote and Buckley (1987) and Lance et al. (2010).

The major advantage compared with the latent variable models discussed above is that with this design, besides the quality of the questions, the common method variance can also be estimated due to the use of the same method across questions. This is relevant because in survey research, batteries with the same form of questions are frequently used.

This approach provides estimates of the quality related with the different form of questions for the same latent variables. This procedure cannot say whether the specific latent variable is a good indicator for the concept of interest. Neither can social desirability and lack of knowledge be evaluated in this manner.

A limitation of this approach is that only a limited set of alternative forms for a specific latent variable are evaluated. The obtained results cannot be generalized. If meta-analyses across the existing MTMM experiments are conducted, a more general picture will arise. This was the basis for the SQP approach.

Another limitation is that the models used presently are based on the assumption of continuous observed variables. Whether this is a serious problem has yet to be studied in more detail. Some results suggest that it is not so serious an issue (Coenders et al. 1997). Only a start has been made with MTMM models for categorical variables (Oberski 2011).

This design has also problems with background variables and simple behavioral questions, because variations of these questions are difficult to formulate and to study.

4.12. Survey Quality Prediction: SQP

The computer program SQP 2.0 has been developed to generate predictions of the quality of questions, based at the moment on a data set of 4000 questions which have been involved in MTMM experiments. The quality is defined as the product of the reliability and validity of a question. The reliability and validity of a question are estimated in MTMM experiments. The program SQP 2.0 provides these estimates for all questions which have been involved in an MTMM experiment. But the program does more. Based on coding of the question characteristics of these 4,000 questions, a prediction procedure has been developed for the quality of the questions. The prediction of the quality of these 4,000 questions is rather good (close to .9), therefore, the program also offers the possibility to use this prediction procedure for predicting the quality of new questions. In order to do so, the user has to code the characteristics of the question, including some research characteristics, and the program then generates the prediction. It also provides suggestions for the improvement of the question, if necessary. For details of the procedure we refer to Saris and Gallhofer (2007) and a more recent publication by Saris et al. 2012.

Given that the predictions are based on coding of around 50 question characteristics and some research design characteristics, quality evaluation is mainly directed at the effects of the form of the questions, although the domain and the concept of the question and the social desirability and knowledge of the respondents of the issue are also taken into account in the prediction. An attractive feature is that form characteristics can be coded in all languages, and so the program can make predictions of the quality of questions in all languages that have been involved in the MTMM experiments, which are more than 20.

A limitation of the program is that it is concentrated on the form of single questions, keeping the concept by intuition the same. Whether this concept by intuition is a good indicator for the concept the respondent wants to measure is outside the scope of this program. So the validity coefficient predicted is the validity for a concept by intuition. The quality can be defined as the explained variance in the observed responses by the concept by intuition studied.

A second limitation of the program SQP is that it is based on MTMM experiments. These experiments are rather difficult for background variables and simple behavioral questions, as was mentioned above. So SQP cannot predict the quality of these questions.

4.13. Scaling Procedures

Most scaling procedures analyze the data of several questions simultaneously to test an expected structure between them. Typical examples are the Thurstone scale, Likert scale, etc. (Torgerson 1958), Rasch scale and item response theory (Hambleton et al. 1991), Gutmann scale, Mokken scale and the unfolding scale (van Schuur 1997), to mention some of them. These scales are based on different models, but all aim at ultimately deriving a score for a respondent on one or perhaps more scales. So these procedures claim to determine a score for the respondents on a scale for the variable of interest. However, the scaling procedure itself cannot guarantee that the score obtained really represents the variable of interest. In fact, like all model based methods mentioned, the procedure can only provide an estimate of the quality of the obtained score for whatever the latent variable may be.

The limitation of these approaches is therefore that they provide only an estimate of the quality of the observed scores, but not of the validity, the social desirability or the lack of knowledge of the respondents with respect to the issue.

Besides this, no attention is paid in these procedures to the problem of common method variance.

5. Conclusions

Looking at the given criteria, some obvious results can be observed:

1. All procedures based on personal judgments provide information about the validity, social desirability, and knowledge of the respondents about the issue of the question and much less about the effects of the form of the questions.
2. The model-based procedures provide rather precise information about the effect of the form of the question on the quality, and the quality can even be expressed in a number between 0 and 1. However, these procedures cannot provide information about the validity of the question for the concept of interest.
3. It is quite obvious that it makes no sense to start with the evaluation of the form of a question before the validity of the measure for a concept has been determined. This means that the personal judgment procedures, at the left side of Table 1, should play an important role in the first phase of questionnaire design.

   Based on our experience with questionnaire design, we have decided to spend extra time on the development of a procedure that can guarantee with more certainty that researchers measure what they are supposed to measure. This has become the three-step procedure of Saris and Gallhofer (2007). We are still convinced that this procedure requires more attention because it can prevent a lot of problems with respect to validity.
4. Evaluating the form of the questions, the model-based procedures, at the right side of Table 1, will be very helpful. In this context, a distinction should be made between evaluation procedures that can only evaluate single questions like SQP, the standard MTMM approach in survey research, and the quasi-simplex approach on the one side, and on the other side procedures that can evaluate measures for concepts by postulation like latent variable models and scaling procedures. In this respect the latter procedures have an advantage. However, they have also the disadvantage that they ignore method effects. In Saris and Gallhofer (2007, ch. 14) we have shown that this may lead to very different conclusions. In psychology, the MTMM approach has also been used for the evaluation of measures for concepts by postulation (Cote et al. 1987 and Lance et al. 2010).
5. There is a fundamental difference between the quality predictions of SQP, which are based on a multivariate prediction approach, and predictions of the quality of the empirical studies, such as latent variable models and also MTMM studies. In SQP, both results are available for all MTMM questions of the ESS. Most of the time the estimates are rather similar, but sometimes they are different. This can occur because the specific question is quite different from the other questions in the database, or in the study of this specific question something was different from normal. This is something one has to decide when looking at these results.

6. The procedures that do not need new data are obviously more attractive than procedures which require new data. On the personal judgment side, it would mean that asking experts for comments is a very attractive procedure before one starts to collect data. On the model-based side, Quaid and SQP seem to be attractive approaches to use before data collection.

6. References

Alwin, D.F. (2007). Margins of Error: A Study of Reliability in Survey Measurement. Hoboken: Wiley.

Andrews, F.M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Equation Approach. Public Opinion Quarterly, 48, 409–442.

Blalock, H.M. Jr, (1968). The Measurement Problem: A Gap Between Languages of Theory and Research. In Methodology in the Social Sciences, H.M. Blalock and A.B. Blalock (eds). London: Sage, 5–27.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrices. Psychological Bulletin, 56, 81–105.

Coenders, G., Saris, W.E., Batista-Foguet, J.M., and Andreenkova, A. (1999). Stability of Three-Wave Simplex Estimates of Reliability. Structural Equation Modeling, 6, 135–157.

Coenders, A.S. and Saris, W.E. (1997). Alternative Approaches to Structural Modeling of Ordinal Data: A Monte Carlo Study. Structural Equation Modeling, 4, 261–282.

Converse, P. (1964). The Nature of Belief Systems in Mass Publics. In Ideology and Discontent, D.A. Apter (ed.). New York: Free Press, 206–261.

Corten, I., Saris, W.E., Coenders, G., van der Veld, W., Albers, C., and Cornelis, C. (2002). The Fit of Different Models for Multitrait-Multimethod Experiments. Structural Equation Modeling, 9, 213–232.

Cote, J.A. and Buckley, M.R. (1987). Estimating Trait, Method and Error Variance; Generalizing Across 70 Construct Validity Studies. Journal of Marketing Research, 11, 535–559.

DeMaio, T. and Landreth, A. (2004). Cognitive Interviews: Do Different Methods Produce Different Results? In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (eds). Hoboken, NJ: John Wiley and Sons.

Faaß, T., Kaczmirek, L., and Lenzner, A. (2008). Psycholinguistic Determinants of Question Difficulty: A Web Experiment. Proceedings of the Seventh International Conference on Social Science Methodology (RC33) [cd-rom], University of Naples "Federico II", Italy.

Forsyth, B., Rothgeb, J., and Willis, G. (2004). Does Question Pretesting Make a Difference? An Experimental Test. In Methods for Testing and Evaluating Survey Questionnaires, Presser et al. (eds). Hoboken, NJ: Wiley.

Fowler, F.J. and Roman, A.M. (1992). A Study of Approaches to Survey Question Evaluation, Final Report for U.S. Bureau of the Census. Boston: Center for Survey Research.

Ginzburg, J. (1996). Interrogatives: Questions, Facts and Dialogue. In The Handbook of Contemporary Semantic Theory, S. Lappin (ed.). Cambridge, MA: Blackwell, 385–421.

Givon, T. (1984). Syntax. A Functional-Typological Introduction Vol. I–II. Amsterdam: J. Benjamin.

Graesser, A.C., McMahen, C.L., and Johnson, B.K. (1994). Question Asking and Answering. In Handbook of Psycholinguistics, M. Gernsbacher (ed.). San Diego, CA: Academic Press, 517–538.

Graesser, A.C., Wiemer-Hastings, P.K., Kreuz, R., and Wiemer-Hastings, P. (2000a). QUAID: A Questionnaire Evaluation Aid for Survey Methodologists. Behavior Research Methods, Instruments, and Computers, 32, 254–262.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., and Kreuz, R. (2000b). The Gold Standard of Question Quality on Surveys: Experts, Computer Tools, Versus Statistical Indices. Proceedings of the Section on Survey Research Methods of the American Statistical Association. Washington, DC: American Statistical Association, 459–464.

Groenendijk, J. and Stokhof, M. (1997). Questions. In Handbook of Logic and Language, J. van Benthem and A. ter Meulen (eds). Amsterdam: Elsevier, 1055–1124.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). Fundamentals of Item Response Theory. London: Sage.

Harris, Z. (1978). The Interrogative in a Syntactic Framework. In Questions, H. Hiz (ed.). Dordrecht: Reidel, 37–89.

Huddleston, R. (1994). The Contrast Between Interrogatives and Questions. Journal of Linguistics, 30, 411–439.

Jansen, H. and Hak, T. (2005). The Productivity of the Three-Step Test-Interview (TSTI) Compared to an Expert Review of a Self-Administered Questionnaire on Alcohol Consumption. Journal of Official Statistics, 21, 103–120.

Koning, P.L. and van der Voort, P.J. (1997). Sentence Analysis. Groningen: Wolters-Noordhoff.

Lance, C.E., Dawson, B., Birkelbach, D., and Hoffman, B.J. (2010). Method Effects, Measurement Error and Substantive Conclusions. Organizational Research Methods, 13, 435–455.

Oberski, D. (2011). Latent Class Multitrait- Multimethod models. In Measurement error in comparative research, D. Oberski (ed.). Unpublished PhD dissertation of the University of Tilburg.

Presser, S. and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? Sociological Methodology, 24, 73–104.

Rothgeb, J., Willis, G., and Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results. Proceedings of the Section on Survey Methods. Alexandria, VA: American Statistical Association.

Saris, W.E. and Andrews, F.M. (1991). Evaluation of Measurement Instruments Using a Structural Modeling Approach. In Measurement Errors in Surveys, P.P. Biemer, R.M. Groves, L.E. Lyberg, N. Mathiowetz and S. Sudman (eds). New York: Wiley, 575–599.

Saris, W.E. and Aalberts, C. (2003). Different Explanations for Correlated Errors in MTMM Studies. Structural Equation Modeling, 10, 193–214.

Saris, W.E., Oberski, D., Revilla, M., Zavalla, D., Lilleoja, L., Gallhofer, I., and Grüner, T. (2012). Final Report About the Project JRA3 as Part of ESS Infrastructure. RECSM Working paper, 24.

Saris, W.E. and Gallhofer, I.N. (2007). Design, Evaluation and Analysis of Questionnaires for Survey Research. Hoboken, NJ: Wiley.

Scherpenzeel, A.C. (1995). A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies. Leidschendam: KPN Research.

Schuman, H. and Presser, S. (1981). Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context. New York: Academic Press.

Torgerson, W.S. (1958). Theory and Methods of Scaling. London: Wiley.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). The Psychology of Survey Response. Cambridge, MA: Cambridge University Press.

Weber E.G. (1993). Varieties of Questions in English Conversations. Amsterdam: J. Benjamins Publ. Co.

Willis, G.B., Schechter, S., and Whitaker, K. (2000). A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What do They Tell Us? Proceedings of the Section on Survey Research Methods. American Statistical Association.

van der Veld, W. (2006). Judging Different Models to Estimate Survey Question Quality. In The Survey Response Dissected: A New Theory About the Survey Response Process, W. van der Veld (ed.). Unpublished PhD dissertation of the University of Amsterdam, Chapter 5.

van Schuur, W.H. (1997). Nonparametric IRT Models for Dominance and Proximity Data. In Objective Measurement: Theory into Practice, M. Wilson, G. Engelhard, Jr, and K. Draney (eds). Volume 4. Greenwich (Cn)/London: Ablex Publishing Corporation, 313–331.

Zaller, J.R. (1992). The Nature and Origins of Mass Opinion. Cambridge: Cambridge University Press.

Received October 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 553–554

Rejoinder

Ting Yan, Frauke Kreuter, and Roger Tourangeau

We thank the Editors-in-Chief of the Journal of Official Statistics, the reviewers, and the discussants for their comments on and discussion of our article. We especially appreciate it that they all point out, in various ways, the difficulties and challenges of conducting comparative studies like ours. We took a first attempt (perhaps an imperfect one) at it because we believe that difficulties and challenges are not an excuse for not trying, and that an imperfect attempt is better than no attempt at all.

One challenge with a comparison of question testing methods is the large differences between the different question evaluation methods. We decided on a basic metric – whether an item was classified as problematic – that could be easily implemented and compared across different question evaluation methods. This metric may not fully capture the products of a particular evaluation method. But we think many questionnaire designers sort draft items in a similar way, deeming some items as needing more work and other items as ready for administration. In addition, the use of this metric allows readers to easily connect our findings back to the existing literature on methods for testing questionnaire items. In the spirit of advancing research on question pretesting and evaluation, we encourage researchers to build on this simple metric and to propose other criteria that better capture the unique contribution of each question evaluation method. We are happy to make our data available to researchers who are interested in seeing whether alternative schemes for classifying our items would have produced different conclusions.

We do not necessarily disagree with the thinking that convergence should not be expected from these very different question evaluation methods. However, simply dismissing the convergence as a criterion for evaluating different question testing methods does not, it seems to us, push the science further. As we mentioned in our article (and we reiterate here), "the answers to the questions of whether converging conclusions should be expected and how to cope with diverging conclusions about specific items depends in part on how researchers conceive of the purpose of the different evaluation methods." In this regard, we agree with the discussants that the next steps for continuing this research is to outline circumstances under which convergence (or divergence) should be expected, and to identify circumstances under which each of the different methods is likely to be useful. Still, we continue to think it was quite reasonable for us to start with the assumption that the problems detected in cognitive interviews and those pointed out by expert reviewers *should* be related to the item's validity and reliability. If the "problems" detected by a given method are unrelated to whether the item produces reliable and valid answers, it is not clear to us what the value of the method is for evaluating questionnaire items.

We did not intend to criticize any qualitative question evaluation methods and we do not endorse any quantitative evaluation method either. However, we do think it is important for future research that those who advocate the use of a particular qualitative method make it clear what unique insights or contributions this method is supposed to provide so that these claims can be evaluated. For instance, one discussant pointed out that cognitive interviewing is practiced in various forms. A critical question then becomes what insights cognitive interviewing offers when the goal is to understand survey questions better, and what insights cognitive interviewing provides when the goal is to detect problems with a particular survey question and to fix those problems. We think it is equally important that advocates of each quantitative method make it clear what assumptions are required to apply the method and to specify the circumstances under which the method may fail because the assumptions are not met. In our examination of latent class analysis, we have demonstrated empirically that when the local dependence assumptions are violated or when the model-identifying assumptions are not met, the latent class method can yield inaccurate estimates of error rates and very implausible results about the differences across different modes of administration (Kreuter, Yan, and Tourangeau 2008; Yan, Kreuter, and Tourangeau 2012).

To advance research on question pretesting and evaluation and to enrich survey literature, we believe that the field needs more studies that include solid measures of validity and reliability on the one hand, and that employ multiple question evaluation methods on the other. In this way, question evaluation methods can be compared on questions with known psychometric properties. This is probably too ambitious a goal for one study. However, as studies and evidence cumulate over time, it will strengthen research on question testing and evaluation in particular and on survey research in general. Good examples of accumulating evidence from question evaluation studies include QBANK started by the National Center for Health Statistics (NCHS) in the United States (http://wwwn.cdc.gov/qbank/Home.aspx), QDDS in Germany (http://www.qdds.org/. See also Schnell and Kreuter 2001), and SQP (http://www.sqp.nl/. See also Saris et al. 2011). We advocate similar efforts to start accumulating experiments and other studies comparing different evaluation methods. Our main point is that we cannot simply continue to take it on faith that the methods we use for evaluating survey questions actually yield helpful insights.

References

Kreuter, F., Yan, T., and Tourangeau, R. (2008). Good Item or Bad – Can Latent Class Analysis Tell? The Utility of Latent Class Analysis for the Evaluation of Survey Questions. Journal of the Royal Statistical Society, Series A, 171, 723–738.

Saris, W.E., Oberski, D., Revilla, M., Zavala, D., Lilleoja, L., Gallhofer, I., and Gruner, T. (2011). Final report about the project JRA3 as part of ESS Infrastructure. Available from: http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf.

Schnell, R. and Kreuter, F. (2001). Neue Software-Werkzeuge zur Dokumentation der Fragebogenentwicklung. ZA-Informationen, 48, 56–70.

Yan, T., Kreuter, F., and Tourangeau, R. (2012). Latent Class Analysis of Response Inconsistencies across Modes of Data Collection. Social Science Research, 41, 1017–1027.

Received October 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 555–582

Disfluencies and Gaze Aversion in Unreliable Responses to Survey Questions

Michael F. Schober[1]*, Frederick G. Conrad*[2]*, Wil Dijkstra*[3]*, and Yfke P. Ongena*[4]

When survey respondents answer survey questions, they can also produce "paradata" (Couper 2000, 2008): behavioral evidence about their response process. The study reported here demonstrates that two kinds of respondent paradata – fluency of speech and gaze direction during answers – identify answers that are likely to be problematic, as measured by changes in answers during the interview or afterward on a post-interview questionnaire. Answers with disfluencies were less reliable both face to face and on the telephone than fluent answers, and particularly diagnostic of unreliability face to face. Interviewers' *responsivity* can affect both the prevalence and potential diagnosticity of paradata: both disfluent speech and gaze aversion were more frequent and diagnostic in conversational interviews, where interviewers could provide clarification if respondents requested it or the interviewer judged it was needed, than in strictly standardized interviews where clarification was not provided even if the respondent asked for it.

Key words: Respondent paradata; respondent cues of processing difficulty; interviewing mode; face-to-face interviewing; telephone interviewing; conversational interviewing; standardized interviewing.

1. Introduction

When survey respondents answer survey questions, they can provide information beyond the content of their answers. As Couper (2000, 2008) termed it, respondents provide *paradata* along with their answers (the survey data): extra evidence about their response process, and thus perhaps about the quality of their answers. Depending on the mode of the survey, different kinds of cues potentially constitute useful paradata (Conrad et al. 2008). For example, in a textual web survey a respondent's delay before answering can give evidence about how much trouble she is having answering the question (e.g., Conrad et al. 2007; Yan and Tourangeau 2008); in a telephone interview a respondent's *um*s and *uh*s can be informative about the extent to which she needs clarification (e.g., Schober and Bloom 2004), and her delays can signal various problems with answers (Bassili and Scott 1996; Draisma and Dijkstra 2004; Ehlen et al. 2007; Schaeffer and Maynard 2002). Paradata are almost certainly exploited by interviewers who adjust the tone or style of an interview to

[1] Department of Psychology, New School for Social Research, 80 Fifth Avenue, New York, NY 10003, U.S.A. Email: schober@newschool.edu. (Corresponding Author)
[2] Institute for Social Research, University of Michigan, Ann Arbor, MI, U.S.A.
[3] Social Research Methods, Free University of Amsterdam, Amsterdam, Netherlands.
[4] University of Groningen, Groningen, Netherlands

Acknowledgments: The authors thank Daniel Nielsen, Jan Smit, Gerhard Van de Bunt and Brady West for their help. This project was supported by the U.S. National Science Foundation Grant Nos. IIS-0081550, SES-0551294, and SES-1025645; the New School for Social Research; the Survey Research Center at the University of Michigan; and the Free University.

match the respondent's needs; they are also potentially exploitable by automated interviewing systems to provide tailored clarification or otherwise adapt to respondents (see papers in Conrad and Schober 2008).

Despite the general recognition that respondent paradata can be informative, survey researchers do not yet have a comprehensive body of knowledge about which kinds of paradata are dependable indicators of respondents' cognitive or emotional states, and under which circumstances. We propose that a careful analysis of the paradata available in different survey modes and the paradata that are produced in different interviewing techniques is needed to build on what is known thus far about details of interviewer-respondent interaction (see, e.g., Cannell et al. 1981; Dykema et al. 1997; Houtkoop-Steenstra 2000; Maynard et al. 2002; Oksenberg et al. 1991; Schaeffer 1991) and inform interviewer hiring, training and practice.

Such an analysis can potentially build on the larger body of evidence about paradata from studies of discourse in noninterview situations (although the term "paradata" is not used in these studies). For example, laboratory studies of answering trivia questions (Brennan and Williams 1995; Smith and Clark 1993; Swerts and Krahmer 2005) have demonstrated that paralinguistic displays (*um*s and repairs) and visual displays (eyebrow movement, smiles, gaze aversion, and "funny face" – diversion from a neutral expression) not only correspond with speakers' lack of confidence ("feeling of knowing") in their answers but can be used by observers to judge that confidence ("feeling of another's knowing"). Studies of other kinds of discourse demonstrate that disfluencies and speech errors can be evidence of speakers' planning and production difficulties (e.g., Fromkin 1973, 1980; Goldman-Eisler 1958; Levelt 1989) and of the complexity, conceptual difficulty or novelty of what they are trying to say (e.g., Barr 2003; Bortfeld et al. 2001; Fox Tree and Clark 1997).

But there is no guarantee that results from studies in other domains will generalize to survey interviewing situations. Survey respondents answer about their own behaviors and opinions rather than retrieving nonautobiographical facts from memory (answers to trivia questions) or referring to objects in scenes they are viewing (as in various psychology experiments). In addition, the particular nature of probing and questioning in subsequent dialogue takes a very particular form in interviews quite unlike other dialogue situations (see, e.g., Houtkoop-Steenstra 2000, papers in Maynard et al. 2002, and Schober and Conrad 2002), and quite unlike laboratory experiments that involve no dialogue. In survey interviews, respondents can have trouble answering for any number of reasons: they can have trouble recalling relevant information or deciding what they think, they can have comprehension problems (trouble knowing what the questioner intends by a term, trouble mapping the question concepts onto their personal circumstances), and they can have trouble formulating or articulating an answer. Any of these kinds of trouble could plausibly result in a problematic (unreliable or inaccurate) answer, and the associated processing difficulties might be evidenced in audio or visual paradata that are produced along with the answer – whether those are intentional communicative *signals* or unintended *symptoms* of processing difficulty (Clark 1996). Of course, problematic answers in surveys can be uttered without any potential indicators of trouble, and answers with potential indicators can be accurate; this is why additional research is needed to establish the relationship between how a survey answer is produced and the quality of that answer.

In the study reported here, our main research question is to what extent two kinds of respondent paradata – fluency of speech and direction of gaze – can diagnose or predict data quality of answers in a corpus of face to face (FTF) and telephone interviews asking nonsensitive factual and opinion questions. In particular, we ask whether the diagnosticity of these paradata is affected (a) by the mode of interviewing (telephone vs. FTF) and (b) by interviewers' responsivity to these paradata, that is, by whether interviewers clarify questions after respondents produce potential indicators of trouble.

Why Focus on Disfluencies and Gaze Aversion?

Disfluencies. Audio paradata in surveys include both linguistic and paralinguistic paradata. Linguistic paradata include words that respondents utter to explicitly inform the interviewer about their processing difficulty, the state of their comprehension (e.g., Mathiowetz 1998, 1999; Oksenberg et al. 1991) or their emotional state. For example, respondents can say that they didn't hear the question ("Could you repeat that?"), that they need clarification ("What do you mean by 'work for pay'?"), or that they feel uncomfortable ("I don't think I want to answer that question"). They can explicitly indicate various other kinds of reactions to the interview ("I never thought about that before"; "I have no idea"; "That's an interesting question"; etc.). They can also "report" rather than selecting a response option from those provided (see Drew 1984; Schaeffer and Maynard 2002, 2008; Schober and Bloom 2004), indicating a mismatch between the question and their circumstances; for example they might answer "I bought tires for a truck" rather than "yes" or "no" in response to the question "Last year, did you have any purchases or expenses for car tires?".

Paralinguistic paradata are the parts of respondents' answers that are not words. These can include speech disfluencies: *um*s and *uh*s (*em*s and *er*s in British transcriptions), pauses and hesitations either before or during an answer, repairs ("three- I mean two") and restarts ("thr- three"). They also include intonational contours: rising intonation in an answer may signal a respondent's uncertainty or need for clarification ("Three?"). Word stress can act as an implicit signal for the interviewer to correct what the respondent recognizes is a potential misinterpretation ("I bought *truck* tires"). Other acoustic cues can indicate emotional distress or irritation (see, e.g., Scherer 2003), and laughter can sometimes indicate discomfort with an answer (e.g., during an answer to a question about sexual behaviors), although it can also sometimes reflect and promote bonding and rapport with the interviewer (see Lavin and Maynard 2002).

We focus on disfluencies in particular for several reasons. First, as paralinguistic phenomena they are relatively frequent, unlike explicit linguistic paradata, which can be rarer; see, for example, Conrad and Schober (2000), in which respondents rarely explicitly requested clarification even when they needed it. Disfluencies are likely to be prevalent enough to allow statistical comparisons, and thus to be potentially practical on a large scale for interviewers or automated interviewing systems to exploit. Disfluencies have the advantage that there is relatively little ambiguity about their occurrence; rising intonation, in contrast, requires more complex measurement tools for researchers, and there may not be consistent agreement within linguistic subcultures about its meaning, as McLemore (1991) and Cameron (2001, pp. 112–114) have documented for speech styles with frequent rising intonation (“uptalk” or “talking in questions”). Finally, speech disfluencies have been argued to occur in potentially problematic answers in telephone interviews (Draisma and Dijkstra 2004; Schaeffer et al. 2008; Schaeffer and Maynard 2002; Schober and Bloom 2004).

Gaze aversion. Less is known about visual than audio paradata in surveys. It is likely that global information about the respondent’s appearance and demeanor can suggest whether the respondent is ready for and attending to the interview. The respondent’s posture, for instance, leaning forward or leaning back, may give evidence of their attentiveness, nervousness, or engagement (Person, D’Mello and Olney 2008). As communication researchers have demonstrated in non-survey situations, respondents’ facial expressions and head movements (furrowed brows, smiles, nods, head turns) potentially reflect (or explicitly signal) engagement, boredom, amusement, or confusion (see, e.g., Swerts and Krahmer (2005) on “audiovisual prosody” that reflects non-confidence in an answer to a trivia question). Eye movements have been shown to be particularly informative; direction of gaze can demonstrate what speakers are referring to (e.g., Hanna and Brennan 2007), when they are holding the floor or ready to pass the floor to another speaker (e.g., Goodwin 1991), or when they are searching for a word (Goodwin and Goodwin 1986). Gaze aversion – looking away from one’s conversational partner – is another cue of potential utility in face to face interviews; several studies have demonstrated that people tend to avert their gaze while answering difficult questions (e.g., Doherty-Sneddon et al. 2002; Glenberg et al. 1998) or when they are not confident in their answers (Swerts and Krahmer 2005). The argument is that people avert the gaze of the questioner in order to temporarily eliminate visual (facial) information which might be distracting and hard to ignore.

In the current study we focus on gaze aversion in particular for two reasons. First, the empirical literature on gaze aversion in non-survey situations points in the same direction: listeners look away from the speaker when engaged in difficult cognitive tasks about whose outcome they lack confidence. It is plausible that this extends to survey settings and reflects respondents’ processing difficulty. Second, unlike other visual paradata like facial expressions, gaze aversion is easy to observe without special training or aptitude, both for researchers and for interviewers; systematic coding of facial expressions, in contrast, can require extremely specialized knowledge, and interviewers may vary in face-reading skills. That is, interviewers might disagree on the meaning of a facial expression, but they are likely to agree, if they are paying attention, on whether a respondent has looked away during an answer.

Why Might Diagnosticity of Paradata Vary by Mode?

Audio paradata are transmitted in both FTF and telephone interviews, but the extent to which they are diagnostic of data quality may vary between modes. In telephone interviews, respondents only have the audio channel available for communication. They cannot assume that interviewers could possibly see their facial expressions or gaze direction; in fact, the notion of gaze aversion cannot even be defined when there is no interviewer from whom the respondent can avert their gaze. Thus it is only audio paradata that could be potentially diagnostic – at least for the interviewer – in telephone interviews. In FTF interviews, respondents can display (intentionally or not) evidence of processing difficulties not only through the auditory channel but also visually, and so they *can* assume that attentive interviewers have access to additional (potentially diagnostic) visual paradata.

The availability of perceptible visual displays in FTF interviews could change the diagnosticity of audio paradata, because visual displays might replace some of the audio paradata in expressing or communicating processing difficulty. If so, then some moments of processing difficulty might be expressed only visually and not audibly, and so the audio displays would diagnose a smaller proportion of episodes of difficult processing in FTF than telephone interviews. Thus the audio paradata in the aggregate would end up being less informative because there are fewer observations. Alternatively, on those fewer occasions in FTF interviews when only audio displays are produced they might be particularly diagnostic because the respondent has not exploited alternative visual means of expressing or communicating processing difficulty, placing the communicative burden entirely on what is audible.

There is reason to hypothesize that audio and visual paradata complement each other. We know from other domains that visual signals – e.g., physically placing an object – can take the place of words (Brennan 1990, 2004; Clark and Krych 2004). Perhaps when interviewers and respondents cannot see each other, as on the telephone, respondents compensate for the lack of visual information by displaying more verbal cues of comprehension difficulty (cf. Whittaker 2003). Swerts and Krahmer (2005) found that visual and audio paradata together allow observers to make better judgments of question-answerers' confidence in their answers to trivia questions than either alone, but whether this generalizes to interviewing situations is unclear. As far as we know, there are no studies on whether visual paradata are always redundant with audio paradata in FTF survey interviews, or whether visual paradata replace (or further emphasize) audio paradata in FTF interviews.

Why Might Diagnosticity Vary by Interviewers' Responsivity?

Interlocutors in general – not just in interviews – can respond to each other's communicative displays quite subtly, picking up on and changing what they say based on their partner's gaze cues, fleeting facial expressions, vocal signs of uncertainty or approval, and so on (see, e.g., Clark 1994, 1996; Goodwin 1991; Schegloff 1984, 1998; Schober and Brennan 2003 among many others). This raises the possibility that how an interviewer reacts to a respondent's audio and visual display could affect the kinds of display that a respondent produces, and thus the extent to which the corresponding paradata are diagnostic of the respondent's processing difficulty. In fact, Schober and Bloom (2004) have demonstrated exactly this in analyses of audio paradata in a corpus of telephone interviews in which respondents answered about fictional scenarios. In the current study we therefore compare the diagnosticity of respondent paradata under two different interviewing techniques: (1) one which encourages attention to and substantive reaction to respondent behaviors that could suggest need for clarification (e.g., "It sounds like you're having some trouble. What can I help you with?"), and (2) one which allows only nonsubstantive reactions (e.g., "let me repeat the question") to respondents' explicit or implicit requests for clarification.

The interviewing technique that encourages substantive reaction to any evidence that a respondent may need clarification is what we have called "conversational" interviewing (Conrad and Schober 2000; Schober and Conrad 1997; Schober et al. 2004). As further detailed below, interviewers using this approach are trained to say what they believe is required to ensure the respondent understands what the survey designers mean by the terms in their questions; interviewers should provide definitions when explicitly asked for them, and they should offer definitions whenever they get the sense that clarification might be helpful. Although the training does not discuss respondent paradata, interviewers are instructed to attend to anything in what a respondent says or does that might suggest that clarification is needed, whether it has been requested or not.

We contrast this with an interviewing technique that requires nonsubstantive reactions: strictly standardized interviewing, following Fowler and Mangione's (1990) prescriptions. In this technique, interviewers are required to administer nondirective probes like "let me repeat the question" when respondents explicitly ask for clarification, and they are expressly forbidden from providing substantive definitions. (The logic is that clarifying words in a question for some respondents would mean that not all respondents would receive the same stimulus). Although Fowler and Mangione (1990) do not explicitly mention respondents' audio or visual displays, their technique would prohibit interviewers from providing a definition in response to spoken or visual potential indicators of trouble.

Respondents' paradata may be differently diagnostic of the respondent's processing difficulty in conversational than in standardized interviews. The potential for a conversational interviewer to help when respondents provide evidence of their processing difficulty may increase respondents' likelihood of displaying such evidence (intentionally or not). This could accurately inform conversational interviewers about respondents' needs more often than in standardized interviews. It is, of course, entirely possible that audio or visual displays are produced by respondents in the same ways no matter how interviewers react; it is also possible that the effects of interviewer reaction may differ FTF and on the telephone. The current study allows us to find out.

Measuring Quality of Answers

To assess the diagnostic value of paradata in the current study, we need to measure which answers are problematic. In this study, respondents answer questions about their own lives rather than fictional scenarios (as they did in Schober and Bloom 2004), and so we cannot measure response accuracy (validity) directly as we have in prior laboratory studies (Conrad et al. 2007; Ehlen et al. 2007; Schober and Conrad 1997; Schober et al. 2004). Instead, we measure two different kinds of (un)reliability: (1) response change (or its complement, consistency) during a question-answer (Q-A) sequence, that is, the respondent first answers the question and then changes the answer before the interviewer asks the next question, and (2) change or consistency between responses in the interview and responses to the same questions in a self-administered post-interview questionnaire where definitions accompany the questions.

The logic for (1) is that answers that change during a Q-A sequence (with or without interviewer-provided clarification) are clearly problematic, even if we don't know whether the original or changed answer (or neither) is correct. At the very least changed answers of this sort reflect a lack of commitment to the original answer and potential uncertainty about which answer to provide. The logic for (2) follows that used in Conrad and Schober (2000): if answers change when definitions are provided (beyond the rate of answer change when no definitions are provided), this is evidence that initial (mis)interpretations have been corrected by the definitions. So response change (unreliability) when the respondent is presented with a definition is evidence that the earlier answer had been problematic. A consistent (reliable) response when the respondent is presented with a definition post-interview is evidence that the earlier answer was nonproblematic.

Note that the technique we use for assessing problematic answers intentionally supplements the wording of the re-asked questions in the post-interview questionnaire by adding definitions. This means that respondents who encountered a definition during a conversational interview will experience the same question and definition in the questionnaire; respondents who did not encounter a definition during the interview (either in standardized interviews or in conversational interviews in which they were not presented with a definition) will be encountering these post-interview definitions for the first time. It is these differences that allow us to assess whether respondents' interpretations of the questions in the original interview were consistent with the definitions presented in the post-interview questionnaire, and thus allow us to measure data quality of their original answers. In Conrad and Schober (2000) this interpretation of response change was supported by evidence that answers for which respondents elaborated their thinking (providing lists of purchased items) were more likely to fit what the survey definitions required when clarification had been provided.

2. Study

This study was carried out in a laboratory, as opposed to field, setting to guarantee suitable video views and audio quality for subsequent analysis, and to make sure that the physical setting was fully comparable in all conditions.

Interviewers were randomly assigned to conduct either strictly standardized or conversational interviews, either on the telephone or FTF; this led to four experimental groups. A total of eight experienced professional Dutch interviewers (all female) participated, with two interviewers assigned to each of the four experimental groups; interviewers had prior experience in both FTF and telephone interviewing. Each interviewer conducted five or six interviews for a total of 42. Respondents were Dutch university students (15 males, 27 females, mean age 22.3 ranging from 19 to 28 years) who were paid roughly the equivalent of US $25 to participate. There were eleven respondents in each of the two standardized groups (FTF and telephone) and ten respondents in each of the two conversational groups (FTF and telephone). The 42 respondents were randomly assigned to one of the four experimental treatments. All interviews were conducted in Dutch and carried out at the Free University of Amsterdam in June of 2000.

Interviewer Training

Interviewers were recruited to participate in a methodological study. They were told that they would be video-recorded for scientific purposes, to improve the quality of survey data collection.

Concepts. All interviewers were trained on the survey concepts being measured in each question (see Appendix A). This primarily involved a supervisor, who was blind to which interviewing technique the interviewer would be implementing, assessing interviewers' competence with the definition for each concept through mock interviews.

Interviewing Technique. Interviewers were then trained in one interviewing technique or the other. Standardized interviewers were trained to strictly follow the prescriptions of Fowler and Mangione (1990). Interviewers were required to read the question as worded; if the respondent did not provide an adequate answer, that is, did not select one of the response options presented with the question, the interviewer was instructed to administer a nondirective probe such as "Let me repeat the question" or "Is that a 'Yes' or a 'No?'" If the respondent requested clarification, the interviewers could only respond with nondirective probes such as "Whatever it means to you."

The instruction for conversational interviewers followed the approach of Schober and Conrad (1997) and Conrad and Schober (2000). In this technique interviewers also were to read the question as worded, but they could subsequently provide clarification if respondents explicitly asked for it or if in the interviewer's judgment the respondent seemed to need it. Interviewers were instructed to say whatever seemed necessary for the respondent to understand as intended, all or part of the definition, verbatim or in the interviewer's own words. Interviewers were not given any special instructions about attending or responding to visual or verbal evidence of difficulty answering.

Experimental Setting

In all of the interviews, the questionnaire was displayed on a laptop computer in front of the interviewer. She read aloud the questions from the computer and entered answers into the computer. The definitions of the survey concepts were printed on a sheet of paper available to the interviewer during the session. For the telephone interviews, the interviewer and respondent were situated in separate buildings. For the FTF interviews, the interviewer and respondent were seated at a table in the same room. All interview sessions, whether conducted on the telephone or in person, were video recorded with separate images of the interviewer's and the respondent's faces. In the FTF sessions, an additional video image was recorded of both parties together, so that we could determine where they were looking and when they were looking at each other.

Survey Questions

The questionnaire consisted of 18 questions, seven of which concerned nonsensitive facts or behaviors (e.g., student status, employment status, and membership in clubs) and eleven of which explored respondents' opinions (six questions about asylum seekers and five about illegal aliens). (See Appendix A for the full list of questions in English translation). In order to assess the impact of definitions on response change, we administered a paper questionnaire after the interview in which respondents were asked to answer the same questions they had just answered except the first five (for these five questions we believed we would have access to official records for students that would have allowed us to assess response validity by comparing access to those records, even if official records can themselves have errors in them; unfortunately, this access ultimately was denied for reasons beyond the authors' control). The questions in the paper questionnaire were accompanied by the definition for the relevant concept (see Appendix A). Respondents completed the questionnaire alone in the same room in which they had been interviewed; an experimenter entered the room to provide the questionnaire, and returned to the room when the respondent had finished.

3. Results

Interviewing Techniques

Before getting to analyses of the paradata, we first verified that the two interviewing techniques had indeed been implemented as interviewers had been trained and that the corpus of interviews had the characteristics of conversational and standardized interviews seen in prior studies (Conrad and Schober 2000; Schober and Conrad 1997; Schober et al. 2004) that would make it suitable for answering our research questions. Interviews were first transcribed and checked by a second transcriber to make sure that all disfluencies were accurately represented in the transcript, including all *um*s and *uh*s (*em*s and *eh*s in Dutch), perceptible pauses (judged by the transcribers as perceptible), repairs (immediate replacements of sounds, words or phrases) and (immediate) restarts. Pauses were notated with periods enclosed within parentheses, and repairs and restarts were notated with double dashes (- -). Interviews were then segmented into Q-A sequences: all behavior from the point at which the interviewer began to ask a question until the interviewer began to ask the next question. Each transcript was coded by one of 3 coders for functional events in the interview (e.g., asking the question, requesting clarification, providing an answer, repeating the question, providing clarification) using a coding scheme (see Appendix B); coders recorded their decisions in Sequence Viewer 4 (Dijkstra 2006; http://www.sequenceviewer.nl/) to allow the interaction and paradata analyses reported below.

To additionally verify reliability of transcription of disfluencies, 150 Q-A sequences (20%) were randomly selected from the total 756 sequences, equally distributed across telephone versus FTF and conversational versus standardized interviews, for independent transcription by a different researcher. Comparisons of these verification transcripts with the original transcripts revealed high reliability of the count of number of functional events with speech disfluencies (Pearson's $r = .946$) and of the number of speech disfluencies per Q-A sequence (which takes into account that in some events there may be more than one speech disfluency) (Pearson's $r = .933$). Reliability of the coding for functional events was measured through extra coding of the same 150 randomly selected Q-A sequences by an independent coder, and it proved to be *substantial* by Landis and Koch's (1977) criteria (Cohen's kappa = 0.743).

As a first piece of evidence on the suitability of the corpus for testing our research questions, interviewers provided clarification more often in conversational interviews (for an average of 33.5% of the questions per interview) than in standardized interviews (for an average of 0.5% of the questions per interview), $F(1,38) = 224.27$, $P < .0001$ (see Table 1). Clarification was provided at the same rates in telephone (16.8%) and FTF (17.2%) interviews, $F(1,38) = 0.02$, ns, and the differences in clarification rates for conversational and standardized interviewing did not differ in the different modes, interaction $F(1,38) = 0.36$, ns. All four conversational interviewers provided clarification

Table 1. Percent of questions per interview for which interviewers provided clarification (SE in parentheses)

	Telephone	FTF	Overall
Standardized	1.0 (2.1)	0 (2.1)	0.5 (1.5)
Conversational	32.7 (2.3)	34.3 (2.3)	33.5 (1.6)
Overall	16.8 (1.6)	17.2 (1.6)	17.0 (1.1)

at least some of the time but not all of the time, ranging from 21% to 44% of the Q-A sequences in the interviews they administered; this is consistent with their training to provide clarification when, in their judgment, clarification was needed. Thus we could be confident that interviewers implemented the technique as they had been trained.

A second piece of evidence that the corpus was suitable is that clarification did indeed affect data quality as in our prior studies. As Table 2 shows, for the questions included in the post-interview questionnaire (Questions 6–18), 89.3% of final answers were the same in the interview and in the post-interview questionnaire when conversational interviewers had provided a definition during the interview (that is, an average of 10.7% of answers changed in the post-interview questionnaire which included definitions). As expected, these answers were significantly more reliable than final answers in those Q-A sequences in which conversational interviewers hadn't provided clarification (67.2%), within-subjects $F(1,38) = 17.80$, $P < .001$, and in standardized interviews (78.1%), in which interviewers almost never provided clarification, between-subjects $F(1,39) = 6.11$, $P < .02$. There were no differences in reliability between telephone and FTF interviews, nor was there any interaction with interviewing technique.

A third piece of evidence on the suitability of the corpus is that conversational interviews took longer than standardized interviews, as one would expect when clarification (which takes time) is given versus when it is not. As Table 3 shows, Q-A sequences lasted 28.2 seconds on average in conversational interviews, but 16.4 seconds in standardized interviews, $F(1,38) = 61.0$, $P < .001$. Interview duration was no different in FTF and telephone modes (unlike in Groves and Kahn 1979), nor did interviewing technique interact with mode, Fs < 1.

Thus we are confident that the interviewers had administered the two interviewing techniques as intended and that our analyses of audio and visual paradata in the two techniques would be based on interviews with the qualities we expected.

Table 2. Reliability of final answers, Qs 6–18 (SE in parentheses)

	Telephone	FTF	Overall
Standardized	77.6 (4.5)	78.6 (4.5)	78.1 (3.2)
Conversational, Q-A sequences without clarification*	62.0 (6.0)	72.4 (5.6)	67.2 (4.1)
Conversational, Q-A sequences with clarification*	89.4 (4.9)	89.2 (4.7)	89.3 (3.4)

* This is a within-subjects comparison

Table 3. Q-A sequences' duration in secs (SE in parentheses)

	Telephone	FTF	Overall
Standardized	17.2 (1.5)	15.7 (1.5)	16.4 (1.1)
Conversational	28.1 (1.5)	28.3 (1.5)	28.2 (1.1)
	22.6 (1.1)	22.0 (1.1)	22.3 (0.8)

Paradata

We first verify that the potential indicators of trouble we are measuring are indeed frequent enough in the sample to ask our research questions. Note that this also gives practical evidence on whether those indicators are frequent enough that interviewers or automated interviewing systems could in principle benefit from exploiting them.

We then examine diagnosticity of the paradata. Our presentation of the findings on diagnosticity reflects the diagnostic problem that interviewers face: given an answer that includes potential indicators of trouble, how likely is it to be a good answer? An alternative analytic approach is to ask whether problematic answers are more likely to include diagnostic cues of response difficulty than nonproblematic answers, as in Schober and Bloom (2004). We have analyzed this data set in both directions (with paradata as independent and dependent variables) and both sets of analyses show essentially the same pattern of results.

For ease of exposition, we first report results about disfluencies, and then about gaze aversion.

Disfluencies

Prevalence of speech disfluencies. We coded every *um* and *uh*, perceptible pause within and between conversational turns, and every repair and restart in each Q-A sequence, through a Sequence Viewer utility that automatically assigned a code based on the notations in the transcript. We treated *um* and *uh* as instances of the same thing, although, as Clark and Fox Tree (2002) note, they may indicate different kinds of trouble.

Our first question was whether respondents produced disfluencies at different rates in telephone and FTF interviews. As Table 4 shows, counting all disfluencies – *um*s and *uh*s, pauses, and repairs and restarts – respondents produced at least one disfluency in their answer (wherever it appeared in the Q-A sequence) in a greater percentage of the Q-A sequences in telephone interviews (57.0%) than they did FTF (42.1%), F(1,38) = 10.56, $P = .002$. (Throughout, the patterns of results are the same whether one counts *um*s and

Table 4. Percent of Q-A sequences that included at least one respondent disfluency, that is, ums and uhs, pauses, and repairs and restarts (SE in parentheses)

	Telephone	FTF	Overall
Standardized	53.7 (4.5)	33.8 (4.5)	43.8 (3.1)
Conversational	60.3 (4.7)	50.5 (4.7)	55.4 (3.3)
Overall	57.0 (3.2)	42.1 (3.2)	

*uh*s or all disfluencies; we will report on all disfluencies, but note that the great majority of disfluencies – 78.9% – were *um*s and *uh*s). The overall pattern of a higher rate of disfluencies over the telephone than FTF is consistent with what has been found in studies of telephone conversation of other kinds (Williams 1977).

Unexpectedly, the rate of disfluencies on the telephone was higher than has been observed in studies of other kinds of discourse in which speakers could only hear each other (e.g., Bortfeld et al. 2001, who observed a rate of about 6 disfluencies per 100 words in a laboratory card-matching task in which participants could not see each other). Here respondents' rate of *um*s and *uh*s during their answer on the telephone was 12.2 per 100 words, reliably higher than the FTF rate of 6.4 per 100 words, $F(1,38) = 12.22$, $P = .001$ (see Table 5); no other effects of interviewing technique or interactions were significant. Disfluency rates varied substantially between different questions; for example, respondents were particularly disfluent (19.1 *um*s and *uh*s per 100 words at some point during the Q-A sequence) while answering the question about how many methods courses they had taken (Q7), compared to a rate of 5.7 per 100 words for Q1-Q3. To the extent that disfluencies reflect processing difficulty, this makes sense; answering Q7 involves demanding mental operations: recalling many courses, determining whether each qualifies, and incrementing a running tally, while Q1-Q3 simply require choosing one of two response options (e.g., whether one is a "full time" or "part time" student).

Interviewing technique also affected the prevalence of disfluencies. Respondents produced at least one disfluency during their answer in a significantly greater percentage of Q-A sequences in conversational interviews (55.4%) than in standardized interviews (43.8%), $F(1,38) = 6.46$, $P = .015$ (see Table 4). This is consistent with the pattern for disfluencies in the (telephone) interviews in Schober and Bloom (2004) and supports the proposal that the interviewer's responsivity can actually change the prevalence of disfluencies. There was no reliable interaction between interviewing mode and interviewing technique.

Can these findings be explained by the influence of individual interviewers? It is, in principle, possible that different interviewers elicited different rates of respondent disfluency, although it is difficult to imagine what interviewer behavior might be involved in such an effect. Nonetheless, if interviewers differ in the respondent disfluency rates with which they are associated and if those with higher rates happened to have been assigned to the telephone or conversational interviewing conditions, this could explain the disfluency results which we are attributing to mode and interviewing technique. To examine this possibility, we computed ρ_{int} for respondent *um* and *uh* rate. This statistic (also labeled "rho-int") was developed by Kish (1962) to measure the degree to which variance (usually

Table 5. Rate of respondent ums and uhs per 100 words (SE in parentheses)

	Telephone	FTF	Overall
Standardized	14.3 (1.6)	5.8 (1.6)	10.0 (1.1)
Conversational	10.2 (1.7)	7.1 (1.7)	8.7 (1.2)
Overall	12.2 (1.2)	6.4 (1.2)	9.3 (0.8)

of responses but in our case disfluency rates) is correlated with individual interviewers (see Biemer and Lyberg (2003) for an introduction).

We calculated ρ_{int} from a mixed model ANOVA consisting of four independent variables: respondents, interviewers, mode and interviewing technique, in which respondents were nested within interviewers and interviewers were nested within mode and interviewing technique. At first blush, interviewer variance for this measure was large (.069), but this is almost entirely attributable to the experimental treatments (mode and interviewing technique) rather than individual interviewers. That is, when we re-run these analyses removing mode and interviewing technique from the model, that is, carrying out a more pure test of different effects of individual interviewers, interviewer-related variance becomes so small that ρ_{int} is effectively zero, despite the fact that small numbers of interviewers can inflate ρ_{int} values. The bottom line is that it seems to be the treatments and not individual interviewers that are driving disfluency rates.

Diagnosticity of disfluencies: Reliability during Q-A sequence. As the first row of Table 6 shows, respondents overall were more likely to change their first answer during the Q-A sequence when it included a disfluency (changing on average 9.8% of their answers) than when it did not (2.1%), F(1,38) = 11.68, $P = .002$. These findings are based on a threeway ANOVA with one within-subjects factor, disfluency (present or absent), and two between-subjects factors, mode (telephone or FTF) and interviewing technique (standardized or conversational); as all respondents produced at least one answer with a disfluency, all 42 respondents are included in this analysis.

The diagnosticity of disfluencies during the first answer in the Q-A sequence varied by mode of interviewing. In particular, disfluencies during this first answer were significantly more diagnostic in FTF interviews (14.5% rate of change for disfluent answers vs. 1.6% for fluent answers) than in telephone interviews (5.1% rate of change for disfluent answers vs. 2.6% for fluent answers), F(1,38) = 5.30, $P = .027$ for the interaction of disfluency and mode (see Table 6 for the full set of means and SEs from this analysis). The diagnosticity of these disfluencies also varied (marginally) by interviewing technique. If we compare diagnosticity of disfluencies between conversational and standardized interviews, collapsing across telephone and FTF interviews, disfluencies were marginally more diagnostic in conversational interviews (15.4% rate of change for disfluent answers vs. 3.6% for fluent answers) than in standardized interviews (4.1% rate of change for disfluent

Table 6. Unreliability of responses: percent of initial answers changed during Q-A sequence (SE in parentheses)

	Fluent	Disfluent
Overall	2.1 (0.7)	9.8 (2.2)
Telephone	2.6 (1.0)	5.1 (3.1)
Standardized	1.1 (1.4)	1.6 (4.2)
Conversational	4.0 (1.4)	8.6 (4.4)
FTF	1.6 (1.0)	14.5 (3.1)
Standardized	0.0 (1.4)	6.7 (4.2
Conversational	3.3 (1.4)	22.3 (4.4)

These analyses exclude the three listing questions (Q4, Q5 and Q6) for which response change during an answer cannot be unambiguously coded because it is unclear when an initial response is unreliable or simply partial.

answers vs. 0.6% for fluent answers), F(1,38) = 3.39, $P = .073$ for the interaction of disfluency and interviewing technique. No other interactions were statistically significant.

Diagnosticity of disfluencies: Reliability of answers as measured post-interview. Recall that in this corpus conversational interviewing led to more reliable answers (as measured post-interview) than standardized interviewing particularly in those cases where the conversational interviewers provided clarification; when they did not, answers were no more reliable. Thus if disfluencies are diagnostic of unreliable answers (as measured post-interview), they should predict response change in those cases where respondents' interpretations were *not* corrected during the interview, that is, in conversational interviews when clarification was not given and in standardized interviews. When clarification had been given, disfluencies in the original answer should not predict response change, because the problems diagnosed by the disfluency should have been resolved by the clarification.

This was exactly the pattern observed. In order to compare reliability for disfluent and fluent answers in conversational interviews where no clarification had been given and in standardized interviews, we carried out a threeway ANOVA with one within-subjects factor, disfluency (present or absent), and two between-subjects factors, mode (telephone or FTF) and interviewing technique (standardized or conversational without clarification). If we collapse the data for all respondents included in the analysis, the overall pattern is that in both cases (standardized interviews and conversational interviews where no clarification had been given) respondents' disfluent answers were more likely to be unreliable (32.1%) than their fluent answers (21.9%), F(1,36) = 4.55, $P < .05$. The means and SEs for all experimental conditions are presented in Table 7A. As expected, this did not vary by interviewing technique (conversational interviews without clarification are essentially standardized) or by mode, nor were there any interactions.

In contrast, disfluencies were no longer predictive of post-experiment response change when interviewers *had* provided clarification in conversational interviews. This can be seen when we compare reliability for disfluent and fluent answers in conversational interviews where clarification had been given and in standardized interviews, in a threeway ANOVA with one within-subjects factor, disfluency (present or absent), and two between-subjects factors, interviewing technique (standardized or conversational with clarification) and mode (telephone or FTF). As Table 7B shows, disfluent answers in the conversational interviews with clarification were 100% reliable (0% response change on

*Table 7A. Unreliability of final answers, Qs 6–18, compared to answers on post-interview questionnaire: percent of changed answers (SE in parentheses)**

	Fluent	Disfluent
Standardized (n = 21)	15.5 (5.3)	30.3 (4.8)
Telephone	16.3 (7.7)	26.3 (7.0)
FTF	14.8 (7.4)	34.4 (6.7)
Conversational interviews with Q-A sequences without clarification (n = 19)	28.2 (5.6)	34.0 (5.1)
Telephone	32.4 (8.1)	36.4 (7.4)
FTF	24.0 (7.7)	31.5 (7.0)

* These analyses include all respondents but two, who either were not disfluent or did not receive clarification.

*Table 7B. Unreliability of final answers, Qs 6–18, compared to answers on post-interview questionnaire: percent of changed answers (SE in parentheses)**

	Fluent	Disfluent
Standardized (n = 21)	15.5 (4.7)	30.3 (4.8)
Telephone	16.3 (6.9)	26.3 (6.9)
FTF	14.8 (6.5)	34.4 (6.6)
Conversational interviews with Q-A sequences with clarification (n = 8)	18.7 (7.7)	0.0 (7.7)
Telephone	37.5 (10.8)	0.0 (10.9)
FTF	0.0 (10.9)	0.0 (10.9)

* These analyses include those respondents in conversational interviews who had at least one Q-A sequence with a disfluent answer followed by clarification.

the post-experiment questionnaire, versus 18.7% response change for fluent answers), while disfluent answers in standardized interviews were unreliable 30.3% of the time (compared to 15.5% response change for fluent answers), interaction of disfluency and interviewing technique F(1,25) = 6.95, $P = .014$.

Altogether, these results show that speech disfluencies are indeed diagnostic of unreliable answers. They are frequent enough to be useful, and they are produced in predictably different ways in different modes (respondents were more likely to be disfluent during an answer on the telephone than FTF) and with differential interviewer responsivity (respondents were more likely to be disfluent in conversational than standardized interviews). And by two different measures of unreliability, answers with disfluencies were more likely to be unreliable. First, they were more likely to change within the Q-A sequence. Second, they were more likely to be corrected post-survey when respondents were provided with clarification – unless respondents had already been provided with clarification during the interview itself.

Gaze Aversion

Prevalence of gaze aversion. The video recordings of the FTF interviews allowed clear views of when respondents looked away from interviewers, turning their heads and averting their gaze (see Figure 1). (Of course we could not examine respondents' direction of gaze in the telephone interviews because the respondent was alone in the room without an interviewer so there was no stable fixation point from which to measure deviation). The start of gaze aversion was defined by eye movement away from the interviewer; the precise moment in time (to within one video frame) at which gaze aversion started could be unambiguously measured by moving the video one frame backwards or forwards. Based on double-coding of a sample of 79 randomly selected Q-A sequences (20% of all Q-A sequences in FTF interviews, with roughly half in conversational and half in standardized interviews), measurement was indeed unambiguous; the two coders' identification of the number of instances of gaze aversion correlated r(79) = .990, $P < .0001$, and measures of the duration of gaze aversion correlated r(79) = .996, $P < .0001$.

Based on this measurement, there were 65 identifiable Q-A sequences in the 21 FTF interviews in which there was at least one instance of gaze aversion. Almost all respondents (19 of 21) averted their gaze at least once during an answering phase, and many did so on

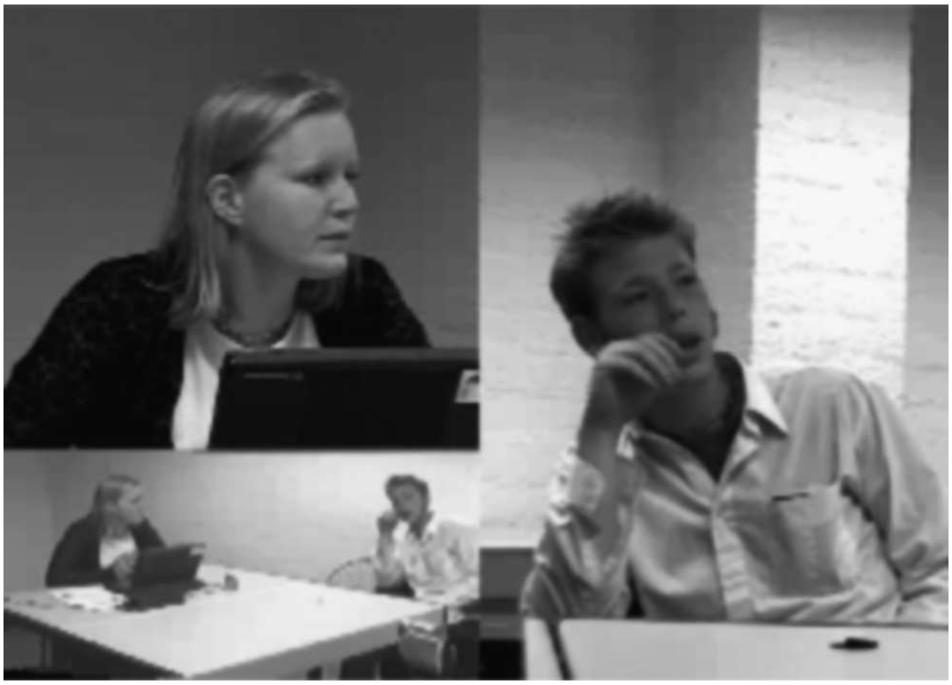


Fig. 1. Respondent (right) averting gaze from interviewer while answering question. (Fotographer: Wil Dijkstra, VU University, Amsterdam)

several questions, up to a maximum of eleven questions. Note that this creates a smaller sample than for the audio paradata, which were observable in both telephone and FTF interviews, but with enough statistical power to carry out a parallel set of analyses.

Respondents averted their gaze at least once during a greater percentage of their answers in conversational interviews (24.7%) than in standardized (11.4%) interviews, $F(1,19) = 5.16$, $P = .035$. Thus, as with the audio paradata, it seems that interviewing technique affects how often respondents produce this visual display. Certainly different interviewing techniques lead to different opportunities to produce visual indicators of trouble; conversational interviews are longer because they sometimes include the presentation of definitions, and so there is simply more time in which gaze aversion could occur. It is also possible that respondents in a FTF conversational interview use gaze aversion to display communication difficulty, much as in ordinary interaction – because interviewers, like ordinary conversational partners, can react substantively to evidence of need for clarification. As was the case with audio paradata, there is no evidence that different interviewers elicited different amounts of gaze aversion: ρ_{int} was effectively zero for FTF interviewers.

Diagnosticity of Gaze Aversion: Reliability During Q-A Sequence

The evidence is that gaze aversion did indeed predict unreliability of answers within a Q-A sequence. Among the 21 FTF interviews, there were 17 respondents (9 conversational and 8 standardized) who produced at least one answer with gaze aversion, which allowed us to compare reliability of answers with and without gaze aversion within-subjects. To do this,

Table 8. *Unreliability of responses: percent of initial answers changed during Q-A sequence, FTF interviews (SE in parentheses)*

	No gaze aversion	Gaze aversion
Overall	4.3 (1.8)	24.7 (8.5)
Standardized	1.0 (2.7)	18.8 (12.3)
Conversational	7.6 (2.5)	30.6 (11.6)

Analysis based on the 17 FTF respondents who produced at least one answer with gaze aversion.

we carried out a two-way ANOVA with one within-subjects factor, gaze aversion (present or absent), and one between-subjects factor, interviewing technique (standardized or conversational). As Table 8 shows, answers with gaze aversion were more likely to be unreliable within the Q-A sequence (24.7%) than answers without gaze aversion (4.3%), $F(1,15) = 4.94$, $P < .05$. The pattern was the same in both interviewing techniques, interaction $F(1,15) = 0.08$, *n.s.*, although perhaps we would see an interaction with a larger sample.

Diagnosticity of Gaze Aversion: Reliability of Answers As Measured Post-interview

Unlike disfluencies, gaze aversion did not predict unreliable answers between the interview and the post-experiment questionnaire. Answers with gaze aversion were no more likely to be unreliable (20.6%) than answers without gaze aversion (24.2%), $F(1,15) = 0.29$, n.s. Following our earlier logic, gaze aversion should predict response change only in the cases where interviewers had not provided clarification: in standardized interviews and in conversational interviews without clarification. Unfortunately we have too few cases for the full within-subjects comparisons we were able to do for disfluencies, but we can compare the cases where interviewers did not provide clarification in both kinds of interviewing. In this comparison, answers with gaze aversion were no more likely to be unreliable (27.3%) than answers without gaze aversion (27.3%), $F(1,15) = 0.0$, ns. And there was no evidence for an effect of interviewing technique on diagnosticity: in conversational interviews 26.3% of answers were unreliable with gaze aversion versus 32.3% without, and in standardized interviews 28.3% of answers were unreliable with gaze aversion versus 22.3% without, interaction $F(1,15) = 0.40$, n.s.

On the other hand, there were five respondents in conversational interviews for whom we could compare (within-subjects) the rate of unreliability for answers in which they averted their gaze and received clarification versus the rate for answers where they averted their gaze and did not receive clarification; the other respondents did not avert their gaze and both receive and not receive clarification. When these five respondents exhibited gaze aversion and received clarification, the rate of unreliable answers (0%) was significantly lower than the rate (32.8%) when they exhibited gaze aversion and did not receive clarification ($F(1,4) = 7.98$, $P < .05$). This is consistent with the notion that gaze aversion followed by clarification leads to more reliable answers than gaze aversion not followed by clarification. So at least part of the logic about unreliability of answers with gaze aversion as measured post-interview holds for a very small sample of respondents, but with only five respondents we see this result as more suggestive than conclusive.

Altogether, these results show that for one kind of visual paradata – gaze direction – respondents in FTF interviews were more likely to avert their gaze during an answer in an interview where the interviewer could provide clarification than in one where the interviewer couldn't. Answers with gaze aversion were more likely to be unreliable within the Q-A sequence than answers without gaze aversion. Answers with gaze aversion were not more likely to be unreliable as measured post-interview, in contrast to disfluencies for which there was such an effect.

4. Discussion

The findings in this study demonstrate that two kinds of respondent paradata – fluency of speech and the direction of gaze during answers to survey questions – can provide evidence about data quality in face to face interviews, and that speech disfluencies can provide evidence about data quality in both face to face and telephone interviews. For both interview modes, answers with these behaviors were more likely to be of poorer quality. The findings extend evidence from other domains of interaction that utterances with these behaviors are more likely to be problematic (unreliable, unconfident, wrong) than utterances without them. They also extend the related Schober and Bloom (2004) finding on speech disfluencies into interviews about autobiographical information and into FTF interviews.

Regarding our first research question, whether the diagnosticity of speech disfluencies is affected by the mode of interviewing (FTF vs. telephone), the evidence is clear. Although answers with disfluencies were less reliable in both modes, disfluencies were particularly diagnostic of unreliability FTF. Disfluencies were also less frequent in FTF interviews than on the phone, possibly because respondents have visual channels for displaying response difficulty beyond audio.

Regarding our second research question, the current findings demonstrate that in both FTF and telephone interviews the interviewer's ability to respond when the paradata indicate trouble affects the respondent's likelihood of indicating that trouble. That is, respondents produced more disfluencies and averted interviewers' gazes more often during answers in conversational interviews, a technique in which interviewers were trained to provide clarification if they got the sense that respondents needed it. And the evidence was that this was not an effect of individual interviewers' somehow eliciting more disfluencies, but rather the result of the experimental treatment – a more collaborative interviewing style that promotes clarification. To our knowledge this provides the only evidence thus far that an interlocutor's potential uptake increases a speaker's likelihood of producing a disfluency or averting gaze. (Oviatt (1995) found that speakers were more likely to be disfluent when speaking to another human than to a computer, but this could be the case for many reasons besides the interlocutor's potential uptake.)

How might these findings be usefully applied to reduce measurement error in survey interviews? We propose several different possibilities, each of which would require additional research in order to be effectively implemented. First, one could imagine implementing new selection criteria for interviewers to hire those who are intuitively able to recognize and make use of visual and auditory evidence of response difficulty. It is possible that current hiring practices already favor interviewers who are interpersonally sensitive on multiple fronts, including the ability to attend to a respondent's audio and visual displays; but it is an empirical question whether this is in fact the case. If so one could imagine making the practice more deliberate.

Second, one could imagine explicitly training already-hired interviewers to detect and make use of the presence of these behaviors, assuming that attentiveness to them can be trained (an open question). Interviewers could be trained to use whatever interviewing techniques are available to them when they encounter evidence of a problematic answer, from additional neutral probing to engaging in clarification dialogue to resolve the trouble. Training materials could be created from existing audio and video recordings of interviews, demonstrating which kinds of verbal and visual behaviors are informative about problematic answers and what the possible subsequent interviewer actions might be.

If such attentiveness turns out not to be easily trainable (interpersonal skill does seem to vary across interviewers), one could imagine designing automated real-time support for helping less sensitive interviewers to recognize potential need for clarification, either for training or production purposes. For example, one could design automated speech recognition systems to monitor and provide evidence to interviewers about delays in the respondent's speech or *um*s and *uh*s (see Ehlen et al. 2007, for a preliminary system of this sort); one could design automated vision tools that could inform an inattentive interviewer about a respondent's gaze aversion, for example processing the video feed in a videomediated interview, or even from an interviewer's laptop in a FTF interview. With such tools, one could even imagine fully automated detection of gaze direction or speech disfluencies in an automated interviewing system. This, of course, would require additional knowledge about whether respondents avert gaze or produce disfluencies in the same way with an automated partner as with a human interviewer.

The findings in this study open the door to additional research on the uses of paradata in interviews and interviewing systems. First, beyond speech fluency and gaze direction it is plausible that other paradata – for example, response latency, vocal stress and tone, facial expressions, gestures, and posture, among others – are systematically related to the quality of responses. Which of these occur frequently enough to be useful, and how universally they are diagnostic across different respondent cultures, dialects, and individual expressive styles, is unknown. We assume that the base rates of potentially diagnostic behaviors – either within an interview mode or technique, across a culture, in an individual, or across different topics (see, e.g., Schachter et al. 1991) – are likely to be important factors in judging the utility of any particular instance of paradata. That is, an *um* produced by a respondent who never *um*s, or averted gaze by a respondent who mostly stares right at the interviewer, should be far more informative about the respondent's cognitive or interactive processes than an *um* produced by a respondent who is chronically disfluent or averted gaze by a respondent who barely maintains eye contact with the interviewer.

Another important arena for additional research is the extent to which different paradata co-occur or supplement one another, and the extent to which they replace each other. In our data set there is a hint that the co-occurrence of audio and visual paradata in FTF interviews is particularly diagnostic: Among the 36 sequences (of 315 FTF sequences) that involved both gaze aversion and disfluency, 9 (25%) resulted in answers that were unreliable between the interview and post-experiment questionnaire. The percentage of unreliable answers was notably lower among the 75 sequences that involved disfluencies alone, where seven answers (9%) were unreliable; among the 13 sequences that involved gaze aversion alone, where one answer was unreliable (a rate of 8%); and among the 191 sequences involving neither disfluency nor gaze aversion, where only two of the answers (1%) were unreliable. Although these are so few cases that we would not want to conclude too much from them, they nonetheless are consistent with the possibility that answers that include displays in more than one channel may be particularly problematic.

Further research is also needed on whether the diagnosticity of different paradata varies for different kinds of questions than those examined here: open-ended questions that require more speech planning, sensitive or personal distress questions for which respondents may feel a greater need to present themselves in a positive light, or particularly complex and difficult questions that require deeper thought. We hypothesize that, in general, the prevalence and diagnosticity of behaviors that provide evidence of trouble answering will be greater for questions for which respondents must construct answers on the fly. And based on our findings, we assume that the diagnosticity of particular paradata is likely to vary in different modes. Given the proliferation of new modes and platforms of interviewing beyond FTF and telephone, it will be important to understand the availability and diagnosticity of different paradata in modes that implement survey dialogue differently, from videomediated interviews to web surveys to speech-IVR interviews, on desktop or mobile multimodal devices, and more.

Presumably not every piece of paradata is revealing about the accuracy or reliability of the speaker's utterance, nor about the speaker's affect or motivation or confidence. The practical challenge for survey researchers will be to understand when interventions that make use of respondent paradata – either by interviewers or automated interviewing systems during the interview itself, or in subsequent data analysis – lead to improved data quality. The theoretical challenge will be to map out, in different domains and styles of discourse, when which paradata are informative of which cognitive and affective states.

Appendix A: Questions and Definitions

Question 1
The first questions in this interview are about your education.
Are you a full-time or part-time student?

1. full-time
2. part-time

definition:
Whether a student is called a part-time of full-time student depends on the official registration form. This seems logical, but many part-time students (officially) participate for whatever reason in the full-time program, and consequently consider themselves (incorrectly) full-time students.

Question 2
Is this a full or shortened course of study?

1: normal
2: reduced

definition:
no definition

Question 3
What is your year of study?
definition:
The registration date determines which year of study a student is in. For instance, if a student was registered as a student by September 1999, he/she is a first year student. This also holds for students who participate in the shortened program (2 instead of 4 years), because the exemptions are based upon prior education (completed outside the Faculty of Social-Cultural Sciences). If a regular student takes up a second course of study, exemptions count. For instance, if a student decides to take up a second course of study and he/she is exempted from the first year, he/she is called a second year student.

Question 4
(not posed to freshmen: 17 respondents)
What is your field of study? [more than one answer is possible]
definition:
no definition

Question 5
Which methodological/statistical courses have you completed during your course of study?
definition:
English:
A course is considered an M&T (methodological/statistical) course when an employee of the Research Methodology Department teaches it and this department is responsible for the course.
In order to complete a course a student must sit for and pass an exam. The course is also considered as completed when a student is exempted from the course due to previous education at another institution.

Question 6
Now I will ask some questions about your membership in clubs.
Can you name all the clubs in which you are a member?
definition:

- An ‘association’ is a legal entity (local authorities and natural persons are legal entities as well).
- An association has members and aims for a certain goal which need not be idealistic.
- A person cannot be the owner of the association; there is no owner.
- An association has a non-profit seeking goal.
- Any profit may not be divided among its members but should be spent on the goal of the association.
- An association is normally established by a notarial deed, containing the articles of vereniging.

- Members of the board as well as of the association according to certain provisions bear personal responsibility for debts and the like.
- Membership is personal (unless it is stated otherwise in the articles of association).
- The members of the board are normally nominated from the members by the general meeting. Each member has a right to vote.
- Within six months (11 at most) the board should publish an annual report, including a financial report.

To mention:

- personal membership
- non-profit seeking goal, no division of profit among members, profit should be spend on the goal
- general meeting of members, board, annual report
- no owner

Question 7
How many paid jobs on the side have you had since July 1, 1999?
definition:
A respondent can have a job on the side only if the job is not his/her main activity. The number of jobs on the side depends on the number of employment contracts. If multiple duties are mentioned in one contract only one job is counted. In the case of multiple employers but the same kind of job multiple jobs are counted. In the case of moonlighting there is no legal contract and therefore no job. An employment contract is simply nothing more than a written or oral agreement between employer and employee.

Question 8
I would like to present some statements about asylum seekers and illegal aliens in the Netherlands. First I will present some questions about asylum seekers. We would like to know to what extent you agree or disagree with these statements. You have the following alternatives to choose from: fully agree; agree; neither agree nor disagree; disagree; fully disagree.
Asylum seekers come to Europe because they are in danger in their own country.

1: fully agree
2: agree
3: neither agree nor disagree
4: disagree
5: fully disagree

definition:
An asylum seeker is a person who irrespective of the reason, which can vary a great deal, seeks asylum in The Netherlands. Reasons may be:

- political and religious reasons,
- social-economical reasons,
- ethnic reasons and/or
- social reasons.

An illegal alien is

- a person who is refused asylum, has no status as a recognized fugitive and doesn't have permission to stay in The Netherlands
- a person who never applied as an asylum seeker and stays in the Netherlands without permission (except for holidays), or
- a "white illegal" person

A white illegal person

- is an undocumented alien who has worked for six continuous years in the Netherlands (and is able to show and prove this), and who has a social security number and valid passport.
- Until 1 January 1998 they were qualified for a residence permit.
- Each case is treated separately.
- A "white illegal" person is an illegal alien until he/she obtains the status of recognized fugitive.

For all remaining questions interviewers presented the same response alternatives (1–5) as those for Question 8, and the same definitions were used.

Question 9
Asylum seekers come to Europe to profit from welfare.

Question 10
The Netherlands should close its borders to all asylum seekers.

Question 11
Asylum seekers should make more efforts to adjust to Dutch norms.

Question 12
The areas surrounding asylum seekers' centers are unsafe.

Question 13
The Netherlands should receive asylum seekers with political grounds with open arms.

Question 14
The following statements are about illegal aliens and not about asylum seekers any more. We would like to know to what extent you agree or disagree with these statements. You have the following alternatives to choose from: fully agree; agree; neither agree nor disagree; disagree; fully disagree.
There is enough room in our country for everyone.

Question 15
Illegal aliens should not receive food stamps.

Question 16
Illegal aliens have rights, too.

Question 17
Illegal aliens should not be discriminated against.

Question 18
All illegal aliens deserve the same rights as Dutch citizens.

Appendix B: Coding Scheme for Functional Events

Respondent:

(1) Answers a question from the questionnaire (e.g.,: “I’m a member of a tennis club,” “I’m not a member of any club”
(2) Answers question (or gives information) relevant to the definition (e.g., “I make a contribution,” “There is an annual meeting”)
(3) Any don’t know answer (e.g., “Don’t know if there is an annual meeting”)
(4) Request clarification
(5) Standalone filler (*em* or *eh*)
(6) Answer other question, e.g., about other characteristics (e.g., “It’s in Amsterdam”, “The name is X”)
(7) Report (describe circumstances) (e.g., “I play tennis”)
(8) Request repeat of survey question/present survey question for confirmation
(9) Repeat previous answer at request of interviewer
(10) No more information (e.g., “That’s all”)
(11) Other, including confirmation of other’s utterances and own repetitions

Interviewer:

(1) Read question exactly as worded (include corrected disfluencies)
(2) Read question with change in wording
(3) Repeat question or part of question
(4) Paraphrase question (re-present question or parts of question, deviating from original wording)
(5) State response alternatives
(6) Neutral probe (e.g., “whatever it means to you,” “we need your interpretation,” “let me repeat the question,” “anything else?”, “take your time to think”
(7) Read definition verbatim
(8) Paraphrase parts of definition (includes answering respondent’s question about definition) (e.g., “A club has an annual meeting,” “A sports club is also a club”
(9) request information from respondent pertaining to definition (“Do you make a contribution?” “Is there an annual meeting?”, “is that a real club?”)
(10) request description of other characteristics (“What is the name of the club?”)
(11) Repeat/restate/elaborate respondent’s answer
(12) Request repetition of answer from questionnaire
(13) Back channel (e.g., “uh-huh,” “okay”)
(14) Other

9. References

Barr, D.J. (2003). Paralinguistic Correlates of Conceptual Structure. Psychonomic Bulletin & Review, 10, 462–467.

Bassili, J.N. and Scott, B.S. (1996). Response Latency as a Signal to Question Problems in Survey Research. Public Opinion Quarterly, 60, 390–399.

Biemer, P. and Lyberg, L. (2003). An Introduction to Survey Quality. Hoboken, NJ: Wiley.

Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.R., and Brennan, S.E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. Language and Speech, 44, 123–149.

Brennan, S.E. (1990). Seeking and Providing Evidence for Mutual Understanding. Unpublished doctoral dissertation, Stanford University.

Brennan, S.E. (2004). How Conversation is Shaped by Visual and Spoken Evidence. In Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions, J.C. Trueswell and M.K. Tanenhaus (eds). Cambridge, MA: MIT Press, 95–130.

Brennan, S.E. and Williams, M. (1995). The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners About the Metacognitive States of Speakers. Journal of Memory and Language, 34, 383–398.

Cameron, D. (2001). Working with Spoken Discourse. Thousand Oaks, CA: SAGE Publications, Inc.

Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981). Research on Interviewing Techniques. In Sociological Methodology, S. Leinhardt (ed.). San Francisco: Jossey-Bass, 389–437.

Clark, H.H. (1994). Managing Problems in Speaking. Speech Communication, 15, 243–250.

Clark, H.H. (1996). Using Language. Cambridge: Cambridge University Press.

Clark, H.H. and Fox Tree, J.E. (2002). Using Uh and Um in Spontaneous Speaking. Cognition, 84, 73–111.

Clark, H.H. and Krych, M. (2004). Speaking While Monitoring Addressees for Understanding. Journal of Memory and Language, 50, 62–81.

Conrad, F.G. and Schober, M.F. (2000). Clarifying Question Meaning in a Household Telephone Survey. Public Opinion Quarterly, 64, 1–28.

Conrad, F.G. and Schober, M.F. (2008). Envisioning the Survey Interview of the Future. Hoboken, NJ: Wiley.

Conrad, F.G., Schober, M.F., and Coiner, T. (2007). Bringing Features of Dialogue to Web Surveys. Applied Cognitive Psychology, 21, 165–187.

Conrad, F.G., Schober, M.F., and Dijkstra, W. (2008). Cues of Communication Difficulty in Telephone Interviews. In Advances in Telephone Survey Methodology, J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster (eds). New York: Wiley, 212–230.

Couper, M.P. (2000). Usability Evaluation of Computer Assisted Survey Instruments. Social Science Computer Review, 18, 384–396.

Couper, M.P. (2008). Designing Effective Web Surveys. New York: Cambridge University Press.

Dijkstra, W. (2006). Sequence Viewer, version 4. Available at: http://www.sequenceviewer.nl.

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., and Doyle, C. (2002). Development of Gaze Aversion as Disengagement from Visual Information. Developmental Psychology, 38, 438–445.

Draisma, S. and Dijkstra, W. (2004). Response Latency and (para)Linguistic Expressions as Indicators of Response Error. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: Wiley.

Drew, Paul (1984). Speakers' Reportings in Invitation Sequences. In Structures of Social Action: Studies in Conversation Analysis, J.M. Atkinson and J. Heritage (eds). New York: Cambridge University Press, 129–151.

Dykema, J., Lepkowski, J.M., and Blixt, S. (1997). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 287–310.

Ehlen, P., Schober, M.F., and Conrad, F.G. (2007). Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces. Discourse Processes, 44, 245–265.

Fox Tree, J.E. and Clark, H.H. (1997). Pronouncing "The" as "Thee" to Signal Problems in Speaking. Cognition, 62, 151–167.

Fowler, F.J. and Mangione, T.W. (1990). Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Newbury Park, CA: SAGE Publications, Inc.

Fromkin, V.A. (1973). Speech Errors as Linguistic Evidence. The Hague, Netherlands: Mouton.

Fromkin, V.A. (1980). Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand. New York: Academic Press.

Glenberg, A.M., Schroeder, J.L., and Robinson, D.A. (1998). Averting the Gaze Disengages the Environment and Facilitates Remembering. Memory & Cognition, 26, 651–658.

Goldman-Eisler, R. (1958). Speech Production and the Predictability of Words in Context. Quarterly Journal of Experimental Psychology, 10, 96–106.

Goodwin, C. (1991). Conversational Organization: Interaction Between Speakers and Hearers. New York: Academic Press.

Goodwin, M.H. and Goodwin, C. (1986). Gesture and Coparticipation in the Activity of Searching for a Word. Semiotica, 62(1/2), 51–75.

Groves, R.M. and Kahn, R.L. (1979). Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press.

Hanna, J.E. and Brennan, S.E. (2007). Speakers' Eye Gaze Disambiguates Referring Expressions Early During Face-to-Face Conversation. Journal of Memory and Language, 57, 596–615.

Houtkoop-Steenstra, H. (2000). Interaction and the Standardized Survey Interview: The Living Questionnaire. Cambridge: Cambridge University Press.

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

Landis, J.R. and Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33, 159–174.

Lavin, D. and Maynard, D.W. (2002). Standardization vs. Rapport: How Interviewers Handle the Laughter of Respondents During Telephone Surveys. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 335–364.

Levelt, W.J.M. (1989). Speaking: From Intention to Articulation. Cambridge, MA: MIT Press.

Mathiowetz, N.A. (1998). Respondent Expressions of Uncertainty: Data Source for Imputation. Public Opinion Quarterly, 62, 47–56.

Mathiowetz, N.A. (1999). Respondent Uncertainty as Indicator of Response Quality. International Journal of Public Opinion Research, 11, 289–296.

Maynard, D.W., Houtkoop-Steenstra, H., Schaeffer, N.C., and van der Zouwen, J. (2002). Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview. New York: Wiley.

McLemore, C.A. (1991). The Pragmatic Interpretation of English Intonation: Sorority Speech. Unpublished doctoral dissertation, University of Texas, Austin.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New Strategies for Pretesting Survey Questions. Journal of Official Statistics, 7, 349–365.

Oviatt, S. (1995). Predicting Spoken Disfluencies During Human-Computer Interaction. Computer Speech and Language, 9, 19–35.

Person, N.K., D'Mello, S., and Olney, A. (2008). Toward Socially Intelligent Interviewing Systems. In Envisioning the Survey Interview of the Future, F.G. Conrad and M.F. Schober (eds). New York: Wiley, 195–214.

Schachter, S., Christenfeld, N., Ravina, B., and Bilous, F. (1991). Speech Disfluency and the Structure of Knowledge. Journal of Personality and Social Psychology, 60, 362–367.

Schaeffer, N.C. (1991). Conversation with a Purpose – or Conversation? Interaction in the Standardized Interview. In Survey Measurement and Process Quality, P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: John Wiley, 367–391.

Schaeffer, N.C. and Maynard, D.W. (2002). Occasions for Intervention: Interactional Resources for Comprehension in Standardized Survey Interviews. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 261–280.

Schaeffer, N.C. and Maynard, D.W. (2008). The Contemporary Standardized Survey Interview for Social Research. In Envisioning the Survey Interview of the Future, F.G. Conrad and M.F. Schober (eds). New York: Wiley, 31–57.

Schaeffer, N.C., Dykema, J., Garbarski, D., and Maynard, D.W. (2008). Verbal and Paralinguistic Behaviors in Cognitive Assessments in a Survey Interview. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Schegloff, E.A. (1984). On Some Gestures' Relation to Talk. In Structures of Social Action: Studies in Conversation Analysis, J.M. Atkinson and J. Heritage (eds). Cambridge: Cambridge University Press, 266–298.

Schegloff, E.A. (1998). Body Torque. Social Research, 65, 535–596.

Scherer, K.R. (2003). Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication, 40, 227–256.

Schober, M.F. and Bloom, J.E. (2004). Discourse Cues that Respondents have Misunderstood Survey Questions. Discourse Processes, 38, 287–308.

Schober, M.F. and Brennan, S.E. (2003). Processes of Interactive Spoken Discourse: The Role of the Partner. Handbook of Discourse Processes, A.C. Graesser, M.A. Gernsbacher, and S.R. Goldman (eds). Mahwah, NJ: Lawrence Erlbaum Associates, 123–164.

Schober, M.F. and Conrad, F.G. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? Public Opinion Quarterly, 61, 576–602.

Schober, M.F. and Conrad, F.G. (2002). A Collaborative View of Standardized Survey Interviews. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 67–94.

Schober, M.F., Conrad, F.G., and Fricker, S.S. (2004). Misunderstanding Standardized Language in Research Interviews. Applied Cognitive Psychology, 18, 169–188.

Smith, V.L. and Clark, H.H. (1993). On the Course of Answering Questions. Journal of Memory and Language, 32, 25–38.

Swerts, M. and Krahmer, E. (2005). Audiovisual Prosody and Feeling of Knowing. Journal of Memory and Language, 53, 81–94.

Whittaker, S. (2003). Mediated Communication. In Handbook of Discourse Processes, A.C. Graesser, M.A. Gernsbacher, and S.R. Goldman (eds). Mahwah, NJ: Erlbaum, 243–286.

Williams, E. (1977). Experimental Comparisons of Face-to-Face and Mediated Communication: A Review. Psychological Bulletin, 84, 963–976.

Yan, T. and Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age. Experience and Question Complexity on Web Survey Response Times. Applied Cognitive Psychology, 22, 51–68.

Received October 2011
Revised July 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 583–590

Inferentially Valid, Partially Synthetic Data: Generating from Posterior Predictive Distributions not Necessary

Jerome P. Reiter[1] *and Satkartar K. Kinney*[2]

To avoid disclosures in public use microdata, one approach is to release partially synthetic data sets. These comprise the units originally surveyed with some collected values, for example sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. In practice, partially synthetic data typically are generated from Bayesian posterior predictive distributions; that is, one draws repeated values of parameters in the synthesis models before generating data from them. We show, however, that inferentially valid, partially synthetic data can be generated by fixing the parameters of the synthesis models at their modes. We do so with both a theoretical example and illustrative simulation studies. We also discuss implications of these results for agencies generating synthetic data.

Key words: Confidentiality; disclosure; imputation; microdata; privacy; survey.

1. Introduction

To limit the risks of disclosures when releasing public use data on individual records, statistical agencies and other data disseminators can release multiply imputed, partially synthetic data (Little 1993; Reiter 2003). These comprise the units originally surveyed with some collected values, for instance, sensitive values at high risk of disclosure or values of quasi-identifiers, replaced with multiple imputations. Partially synthetic data can protect confidentiality, since identification of units and their sensitive data can be difficult when select values in the released data are not actual, collected values. And, with appropriate estimation methods based on the concepts of multiple imputation (Rubin 1987), they enable data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Because of these appealing features, partially synthetic data products have been developed for several major data sources in the U.S., including the Longitudinal Business Database (Kinney et al. 2011), the Survey of Income and Program Participation (Abowd et al. 2006), the American Community Survey group quarters data (Hawala 2008), and the OnTheMap database of where people live and work (Machanavajjhala et al. 2008). Other examples of partially synthetic data are described in Abowd and Woodcock (2004), Little et al. (2004), Drechsler et al. (2008), and Drechsler and Reiter (2010).

[1] Duke University, Box 90251, Durham, NC 27708, U.S.A. Email: jerry@stat.duke.edu
[2] National Institute of Statistical Sciences, Research Triangle Park, NC 27709, U.S.A. Email: saki@niss.org

Acknowledgment: This research was supported by a grant from the National Science Foundation (SES-11−31897).

In the statistical theory underlying the generation of partially synthetic data, as well as typical implementations in practice, replacement values are sampled from posterior predictive distributions. That is, the agency repeatedly draws values of the model parameters from their posterior distributions, and generates a set of replacement values based on each parameter draw. The motivation for sampling from posterior predictive distributions derives from multiple imputation of missing data, in which drawing the parameters is necessary to enable approximately unbiased variance estimation (Rubin 1987, Chapter 4).

In this article, we argue that it is not necessary to draw parameters to enable valid inferences with partially synthetic data. Instead, data disseminators can estimate posterior modes or maximum likelihood estimates of parameters in synthesis models, and simulate replacement values after plugging those modes into the models. Using a simple but informative case, we show mathematically that point and variance estimates based on the plug-in method can be approximately unbiased. We also illustrate this fact via simulation studies and include a comparison to generating partially synthetic data from posterior predictive distributions.

The remainder of the article is organized as follows. Section 2 reviews existing methods of generating and making inferences from partially synthetic data. Section 3 offers the mathematical example, and Section 4 presents results of the simulation studies. Section 5 concludes with implications of these results for agencies seeking to generate partially synthetic data.

2. Review of Partially Synthetic Data

To review partially synthetic data, we closely follow the description and notation of Reiter (2003). Let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise. Let $I = (I_1, \ldots, I_N)$. Let Y_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; let Y_{nobs} be the $(N - n) \times p$ matrix of unobserved survey data for the units with $I_j = 0$; and let $Y = (Y_{obs}, Y_{nobs})$. For simplicity, we assume that all sampled units fully respond to the survey; see Reiter (2004) for simultaneous imputation of missing and synthetic data. Let X be the $N \times d$ matrix of design variables for all N units in the population, for instance, stratum or cluster indicators or size measures. We assume that such design information is known approximately for all population units. It may come, for example, from census records or the sampling frame(s).

The agency releasing synthetic data constructs synthetic data sets based on the observed data, $D = (X, Y_{obs}, I)$, in a two-part process. First, the agency selects the values from the observed data that will be replaced with imputations. Second, the agency imputes new values to replace those selected values. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \ldots, Z_n)$. Let $Y_{rep,i}$ be all the imputed (replaced) values in the ith synthetic data set, and let $Y_{nrep,i}$ be all unchanged (unreplaced) values of Y_{obs}. In Reiter (2003), $Y_{rep,i}$ is assumed to be generated from the Bayesian posterior predictive distribution of $(Y_{rep,i}|D, Z)$. The values in Y_{nrep} are the same in all synthetic data sets. Each synthetic data set, d_i, then comprises $(X, Y_{rep,i}, Y_{nrep}, I, Z)$. Imputations are made independently for $i = 1, \ldots, m$ to yield m different synthetic data sets. These synthetic data sets are released to the public.

Reiter (2003) also describes methods for analyzing the m public use, synthetic data sets. Let Q be the analyst's scalar estimand of interest, for example the population mean of Y or some coefficient in a regression of Y on X. In each d_i, the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator u. The analyst determines the q and u as if the synthetic data were in fact collected data from a random sample of (X, Y) based on the actual survey design used to generate I.

For $i = 1, \ldots, m$, let q_i and u_i be respectively the values of q and u computed with d_i. The following quantities are needed for inferences:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i/m. \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_p = b_m/m + \bar{u}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$. When n is large, inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_p = (m-1)\left(1 + r_m^{-1}\right)^2$, where $r_m = (m^{-1}b_m/\bar{u}_m)$. Extensions for multivariate Q are presented in Reiter (2005a) and Kinney and Reiter (2010).

3. Example Showing That Sampling Parameters is Unnecessary

In this section, we provide for one scenario a mathematical proof that the estimators $\bar{q}_m$ and T_p are approximately unbiased for Q and the variance of $\bar{q}_m$, respectively, when generating partially synthetic data without drawing model parameters. For the scenario, we seek to estimate the population mean of a single variable, which we denote $\bar{Y}$, in a simple random sample of size n. We do not utilize additional variables for this example; Section 4 displays simulation results involving regressions.

We suppose that the agency replaces all values of Y_{obs} with draws from some distribution, that is all values of Y_{obs} are confidential. Setting $Z_j = 1$ for all j is common in practice; for example, the synthesis for the Longitudinal Business Database, the Survey of Income and Program Participation, and OnTheMap do so. We assume that a reasonable model for the data is $Y|\mu, \sigma^2 \sim N(\mu, \sigma^2)$. Of course, since we have only n observations in Y_{obs}, we do not know μ and σ^2. Let $\bar{y}$ be the sample mean and s^2 be the sample variance, both computed with Y_{obs}. We propose to generate m partially synthetic data sets with two steps.

D1. Sample n values independently from $N(\bar{y}, s^2)$, resulting in $Y_{rep,i}$.

D2. Repeat step D1 independently for $i = 1, \ldots, m$ to create m partially synthetic data sets that are released to the public.

We note that this process is not sampling from a Bayesian posterior predictive distribution, since we do not draw (μ, σ^2) from their posterior distribution before sampling any $Y_{rep,i}$.

Using data generated via D1 and D2, in each d_i we let $q_i = \bar{y}_i$, that is, the sample mean in d_i, and let $u_i = (1 - n/N)s_i^2/n$, where s_i^2 is the usual sample variance of the values in d_i. Hence, we have $\bar{q}_m = \sum_{i=1}^{m} \bar{y}_i/m;$ $\bar{u}_m = \sum_{i=1}^{m}(1 - n/N)s_i^2/(nm)$; and $b_m = \sum_{i=1}^{m}(\bar{y}_i - \sum_{i=1}^{m}\bar{y}_i/m)^2/(m-1)$. We now derive the expected values of $\bar{q}_m$ and T_p over repeated samples of Y_{obs} from the population, that is, over repeated realizations of (I, Z). Since Z is a vector of ones for all I, we drop it from further notation.

We first show that simulating via D1 and D2 results in an unbiased estimate of $\bar{Y}$ when averaging over repeated samples I. By D1, the $E(\bar{y}_i|Y, I) = E(\bar{y}|Y)$. Hence,

$$E(\bar{q}_m|Y) = E(E(\bar{q}_m|Y, I|Y) = E(\bar{y}|Y) = \bar{Y}. \tag{5}$$

We next show that T_p is unbiased for the actual variance of $\bar{q}_m$ when averaging over repeated samples I. To begin, we write $Var(\bar{q}_m|Y) = E(Var(\bar{q}_m|Y, I)|Y) + Var(E(\bar{q}_m|Y, I)|Y)$. From D1, we have

$$Var(E(\bar{q}_m|Y, I)|Y) = Var(\bar{y}|Y) = (1 - n/N)S^2/n, \tag{6}$$

where $S^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/(N-1)$ is the population variance. Also from D1 and D2, we have $Var(\bar{q}_m|Y, I) = (s^2/n)/m$, so that

$$E(Var(\bar{q}_m|Y, I|\mathrm{Y}) = E(s^2/(nm)|Y) = S^2/(nm). \tag{7}$$

Hence, we have $Var(\bar{q}_m|Y) = S^2/(nm) + (1 - n/N)S^2/n$. Moving to $E(T_P|Y)$, from D1 we have that $E(u_i|Y, I) = (1 - n/N)s^2/n$, so that $E(\bar{u}_m|Y) = (1 - n/N)S^2/n$. Additionally, from D1 we have $E(b_m|Y, I) = s^2/n$. Hence, we have

$$E(T_p|Y) = E(\bar{u}_m + b_m/m|Y) = (1 - n/N)S^2/n + S^2/(nm) = Var(\bar{q}_m|Y). \tag{8}$$

We note that none of the derivations for the t-reference distribution in Reiter (2003) require sampling from posterior distributions. Hence, with approximately unbiased point and variance estimates, we can obtain valid variance inferences with those methods.

4. Simulation Studies

In this section, we illustrate that partial synthesis without posterior predictive simulation can result in well-calibrated inferences. To do so, we generate 10,000 observed data sets D, each comprising $n = 1{,}000$ observations and nine variables. For each D, we sample seven of the variables, denoted as $(X_1, \ldots, X_7)$, from independent $N(0, 1)$. For each observation $j = 1, \ldots, 1{,}000$, let $x_j' = (1, x_{j1}, \ldots, x_{j7})$. For $j = 1, \ldots, 1{,}000$, we draw a continuous variable, Y_1, from the regression $y_{1j} = x_j'\beta + \epsilon_j$, where $\beta = (0, -1, 2, -.5, .1, .1, .1, 3)$, $\epsilon_j \sim N(0, \tau^2)$, and $\tau^2 = 1$. We also draw a binary variable, Y_2, using independent Bernoulli distributions such that $\mathrm{logit}(P(y_{2j} = 1)) = x_j'\alpha + y_{1j}\gamma$. Here, $\alpha = \beta/3$ and $\gamma = -1/3$. This results in values of $P(y_{2j} = 1)$ that are between .2 and .8 with high probability. We treat (Y_1, Y_2) as sensitive variables and synthesize all of both. We do not change values of $X = (X_1, \ldots, X_7)$.

To generate partially synthetic data, we consider two possible strategies. The first is to sample from posterior predictive distributions as recommended in Reiter (2003). We estimate the posterior distributions of β and τ^2 based on the default improper prior distribution, $p(\beta, \tau^2) \propto 1/\tau^2$. Let $\hat{\beta}$ be the maximum likelihood estimate (MLE) of β, and let $s^2_{y_1|x} = \sum_{j=1}^{n}(y_{1j} - x_j'\hat{\beta})^2/(n-p)$ be the usual unbiased estimate of τ^2. Let $(\hat{\alpha}, \hat{\gamma})$ be the MLE of (α, γ), and let $\hat{\Lambda}$ be the estimated covariance matrix of $(\hat{\alpha}, \hat{\gamma})$. These quantities are obtainable from standard logistic regression output. The synthesis process following Reiter (2003) proceeds as follows.

P1. Sample a value of τ^2, say $\tau^{2}*$, from its inverse χ^2 distribution.
P2. Sample a value of β, say $\beta*$, from a normal distribution with mean $\hat{\beta}$ and variance $(X'X)^{-1}\tau^{2*}$.
P3. Sample $n = 1{,}000$ values of Y_1 from $N(X\beta*, \tau^{2}*)$, resulting in $Y_{1rep,i}$.
P4. Sample a value of (α, γ), say $(\alpha*, \gamma*)$, from a multivariate normal with mean $(\hat{\alpha}, \hat{\gamma})$ and covariance matrix $\hat{\Lambda}$.
P5. Sample $n = 1{,}000$ values of Y_2 from independent Bernoulli distributions such that $\text{logit}(P(y_{2j} = 1)) = x_j'\alpha^* + y_{1rep,i,j}\gamma^*$, resulting in one partially synthetic data set $(X, Y_{1rep,i}, Y_{2rep,i})$.
P6. Repeat steps P1 to P5 independently $m = 5$ times.

We note that P4 approximates the posterior distribution of (α, γ) as a multivariate normal with known covariance. For large n, this approximation is reasonable and is typically used in practice.

The second strategy is to sample without drawing parameters. It involves only three steps.

R1. Sample $n = 1{,}000$ values of Y_1 from $N\left(X\hat{\beta}, s^2_{y_1|x}\right)$, resulting in $Y_{1rep,i}$.
R2. Sample $n = 1{,}000$ values of Y_2 from independent Bernoulli distributions such that $\text{logit}(P(y_{2j} = 1)) = x_j'\hat{\alpha} + y_{1rep,i,j}\hat{\gamma}$, resulting in one partially synthetic data set $(X, Y_{1rep,i}, Y_{2rep,i})$.
R3. Repeat step R1 to R2 independently $m = 5$ times.

Table 1 displays the simulated coverage rates of 95% confidence intervals, as well as the simulated variances of $\bar{q}_m$, for the mean of Y_1, five coefficients in the regression of Y_1 on X, the percentage of observations with $Y_1 > 1$, the mean of Y_2, and six coefficients in the regression of Y_2 on (Y_1, X). The simulated coverage rates in each case are close to the 95% nominal rate, indicating that steps R1–R3 are sufficient for inferential validity in this simulation. The variances of $\bar{q}_m$ across the 10,000 replications when data are generated from R1–R3 are always smaller than those when data are generated from P1–P6.

Table 1. Comparison of simulated coverage rates for 95% confidence intervals and simulated variances of $\bar{q}_m$ when partially synthetic data are created with (Draws) and without (No draws) sampling from the posterior distributions of the parameters. Results based on 10,000 replications. Variances are reported in parentheses after multiplying by 10^3.

	$E(Y_1)$	β_1	β_2	β_3	β_4	β_5	$P(Y_1 > 1)$
Draws	94.8 (15.6)	94.9 (1.4)	95.4 (1.4)	94.8 (1.4)	94.7 (1.4)	95.2 (1.4)	97.0 (.21)
No draws	94.8 (15.4)	94.9 (1.2)	94.8 (1.2)	94.9 (1.2)	94.6 (1.2)	95.2 (1.2)	97.1 (.21)
	$E(Y_2)$	α_1	α_2	α_3	α_4	α_5	γ
Draws	94.9 (.35)	94.6 (12.6)	94.8 (32.0)	95.1 (7.7)	95.2 (6.0)	95.1 (6.0)	94.9 (6.5)
No draws	95.1 (.30)	94.6 (10.9)	94.5 (27.5)	94.6 (6.6)	94.7 (5.3)	94.9 (5.3)	94.8 (5.5)

The magnitude of the variance reduction is minor for the mean of Y_1 and the $P(Y_1 > 1)$, but it is generally between 15% and 20% for the other parameters.

We also ran a simulation with $n = 10{,}000$ and otherwise the same design. The 95% confidence interval coverage rates were well-calibrated. The variances of $\bar{q}_m$ across the 10,000 replications when data were generated from R1–R3 continued to be always smaller those when data were generated from P1–P6.

5. Concluding Remarks

Based on the mathematical example and simulations, it appears that agencies do not need to sample from the posterior distributions of parameters to facilitate valid inference from partially synthetic data. This has considerable implications for the generation of partially synthetic data in practice. First, sampling from posterior distributions can be time consuming, as it may require running MCMC algorithms to get posterior distributions. Simply plugging in modes, which often can be computed with off-the-shelf software routines, can reduce this cost. Second, it lends support to the use of synthesizers based on algorithmic methods from machine learning, such as regression trees (Reiter 2005b), random forests (Caiola and Reiter 2010), and support vector machines (Drechsler 2010). These are difficult to justify from the perspective of posterior predictive distributions, since they do not have readily identified model parameters. However, in practice they have been shown to perform reasonably well as data synthesizers (Drechsler and Reiter 2011). Third, it offers agencies a way to reduce variances of secondary analyses of the released synthetic data.

While synthesizing based on plug-in modes has analytical advantages, it could have disadvantages from the perspective of confidentiality protection. In the setting of Section 3, for example, suppose that an ill-intentioned data snooper knows all values of the variable Y except for one, say y_j. If the data snooper can get a sharp estimate of $\bar{y}$ from the synthetic data, he effectively learns the unknown y_j. When synthetic data are generated from $N(\bar{y}, s^2)$, the data snooper may be able to use $\bar{q}_m$ and $\bar{u}_m$ to get close estimates of $(\bar{y}, s^2)$, and therefore closely estimate the unknown y_j. On the other hand, when synthetic data are generated by drawing (μ, σ^2) first, the data snooper's estimate of $(\bar{y}, s^2)$ has greater uncertainty, and hence his estimate of the unknown y_j is likely to have higher error. Of course, the "intruder knows all values but one" scenario is an unlikely one in many surveys, and the two approaches may have similar disclosure risk profiles in practice. Nonetheless, the example suggests that evaluating trade offs in risk and utility from the two partial synthesis strategies is an area for future research.

Many data sets also contain missing values. Reiter (2004) presents an approach to multiple imputation of missing data and synthetic data simultaneously, in which the agency (i) fills in the missing data by sampling from posterior predictive distributions to create m completed data sets, and (ii) replaces confidential values in each completed dataset with r partially synthetic imputations. Hence, a total of mr nested data sets is released. With this approach, it is necessary to sample from posterior predictive distributions in the first stage of completing the missing values. However, the results in Section 3 and 4 here imply that it is not necessary to use posterior predictive simulation at the second stage.

We also note that it remains necessary to draw from posterior predictive distributions for fully synthetic data (Rubin 1993; Raghunathan et al. 2003; Si and Reiter 2011). In fully synthetic data, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these data sets to the public. Fully synthetic data essentially involve filling in missing values for records that were not in the original sample. Since one needs to predict values that are not observed, one needs to account for parameter uncertainty in the synthesis models.

6. References

Abowd, J., Stinson, M., and Benedetto, G. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at http://www.census.gov/sipp/synth_data.html.

Abowd, J.M. and Woodcock, S.D. (2004). Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. In Privacy in Statistical Databases, J. Domingo-Ferrer and V. Torra (eds). New York: Springer, 290–297.

Caiola, G. and Reiter, J.P. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. Transactions on Data Privacy, 3, 27–42.

Drechsler, J. (2010). Using Support Vector Machines for Generating Synthetic Datasets. In Privacy in Statistical Databases, J. Domingo-Ferrer and E. Magkos (eds). New York: Springer, 148–161.

Drechsler, J., Bender, S., and Rässler, S. (2008). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. Transactions on Data Privacy, 1, 105–130.

Drechsler, J. and Reiter, J.P. (2010). Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata. Journal of the American Statistical Association, 105, 1347–1357.

Drechsler, J. and Reiter, J.P. (2011). An Empirical Evaluation of Easily Implemented, Non-parametric Methods for Generating Synthetic Datasets. Computational Statistics and Data Analysis, 55, 3232–3243.

Hawala, S. (2008). Producing Partially Synthetic Data to Avoid Disclosure. Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association.

Kinney, S.K. and Reiter, J.P. (2010). Tests of Multivariate Hypotheses when Using Multiple Imputation for Missing Data and Partial Synthesis. Journal of Official Statistics, 26, 301–315.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., and Abowd, J.M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review, 79, 363–384.

Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Little, R.J.A., Liu, F., and Raghunathan, T.E. (2004). Statistical Disclosure Techniques Based on Multiple Imputation. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, A. Gelman and X.L. Meng (eds). New York: John Wiley and Sons, 141–152.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory Meets Practice on the Map. IEEE 24th International Conference on Data Engineering, 277–286.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1–16.

Reiter, J.P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 29, 181–189.

Reiter, J.P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology, 30, 235–242.

Reiter, J.P. (2005a). Significance Tests for Multi-Component Estimands from Multiply-Imputed, Synthetic Microdata. Journal of Statistical Planning and Inference, 131, 365–377.

Reiter, J.P. (2005b). Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, 21, 441–462.

Rubin, D.B. (1987b). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 462–468.

Si, Y. and Reiter, J.P. (2011). A Comparison of Posterior Simulation and Inference by Combining Rules for Multiple Imputation. Journal of Statistical Theory and Practice, 5, 335–347.

Received February 2012
Revised August 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 591–613

Confidentialising Exploratory Data Analysis Output in Remote Analysis

Christine M. O'Keefe[1]

This article is concerned with the problem of balancing the competing objectives of allowing statistical analysis of confidential data while maintaining privacy and confidentiality. Traditional approaches to reducing the risk of disclosure typically involve modifying or *confidentialising* data before releasing it to users. In contrast, *remote analysis* enables analysts to submit statistical queries and receive output without direct access to data.

In this article we discuss the implementation of remote analysis allowing exploratory data analysis on confidential data, where the system outputs are modified to protect confidentiality. To illustrate the effect of the modifications, we provide a comprehensive example comparing traditional and confidentialised output for a range of common exploratory data analyses on discrete and continuous data.

We believe that confidentialised exploratory data analysis output is still useful, provided the analyst understands the confidentialisation process and its potential impact. Where the potential impact is judged to be too great, the analyst will need to seek another mode of access to the data.

Key Words: Confidentiality; privacy; remote access; remote data access; output checking.

1. Introduction

This article addresses the challenge of balancing the competing objectives of allowing statistical analysis of confidential or private data and maintaining standards of privacy and confidentiality. Such standards can include those imposed by relevant privacy legislation and regulation, as well as assurances provided by data custodians to data contributors.

This balance is often characterised as a trade-off between disclosure risk and data utility (see Duncan et al. 2001). Disclosure risk attempts to capture the probability of a data release resulting in a disclosure, while data utility attempts to capture some measure of the usefulness of the released data.

A high-level discussion of the problem of achieving this balance typically covers two broad approaches, which are often used in combination. The first approach is *restricting access,* where access to data is granted under strong controls including researcher training and registration, supervised secure data laboratories or secure remote access environments, analysis output checking as well as legal and operational protections and agreements. Many national statistical agencies allow researcher access to confidential data in secure, on-site research data centres. Examples include the Australian Bureau of

[1] CSIRO Mathematics, Informatics and Statistics, GPO Box 664, Canberra ACT 2601 Australia.
Email: Christine.O'Keefe@csiro.au

Acknowledgment: Thanks to Scott Furman for assistance with Figure 7(b).

Statistics (ABS) On-site Data Laboratory (Australian Bureau of Statistics n.d.), the United Kingdom Office For National Statistics (ONS) Virtual Microdata Laboratory (Office for National Statistics n.d.) and the Census Bureau Research Data Centers (RDC) (United States Census Bureau n.d.). An example of the remote access approach is the US NORC Data Enclave, which provides a confidential, protected environment within which authorised social science researchers can access sensitive microdata remotely (University of Chicago n.d.). In the NORC Enclave, researchers do not have access to the internet and cannot move files into or out of the secure environment without review approval. Any export request from a researcher is scrutinised by a NORC statistician to ensure that it does not contain disclosive data. If there are any disclosure concerns, the researcher is notified and the output is not released. If no concerns exist, the output is cleared and uploaded to a transfer site from where the researchers can download the output. Similar systems include the UK Secure Data Service, which provides secure remote access to data operated by the Economic and Social Data Service (UK Data Archive n.d.) and the Australian Bureau of Statistics (ABS) Remote Access Data Laboratory (RADL) (Australian Bureau of Statistics n.d). While these systems are very successful, manual output checking is highly context dependent, requires specialised statistical skills and can be very time consuming. In particular, it is normally not possible to define common rules for deciding in advance whether an output can be released or not. In December 2009 the ABS noted that it was experiencing high user demand for access to more detailed unit record data in a more flexible way, across a wider array of datasets (such as business data and longitudinal linked datasets; see Australian Bureau of Statistics 2009). In order to manage the risk of inability to meet this demand, the ABS is pursuing a strategy of progressive replacement of RADL with a new system, primarily for table generation and basic statistical analysis. It is proposed that this new system will enable access to detailed de-identified microdata, and will make use of automated output confidentialisation routines to ensure that system outputs meet ABS legislative requirements. The system outputs will be able to be released as public use outputs, that is, they will be able to be published and shared with others without restrictions.

The second approach is *restricting or altering data,* where less than the full dataset is released or the data are altered in some way before release to analysts, in order to provide enhanced confidentiality protection. First, identifying attributes such as name and address are usually removed, as well as other sensitive attributes or observations. Often, this is followed by the application of *statistical disclosure control* methods such as aggregation of geographic classifications, rounding, swapping or deleting values, and adding random noise to data. The application of statistical disclosure control techniques also requires specialised statistical skills and is highly context dependent, and it can be extremely difficult to quantify the level of protection achieved. Unfortunately, statistical disclosure control methods can also result in information loss and/or biased estimation. For more information on statistical disclosure control methods, see, for example Adam and Wortmann 1989; Domingo-Ferrer and Magkos 2010; Domingo-Ferrer and Saygin 2008; Domingo-Ferrer and Torra 2004; Doyle et al. 2001; Office of Information and Regulatory Affairs 1994; Willenborg and de Waal 2001). Motivated by the drawbacks associated with statistical disclosure control, Rubin (1993) suggested the alternative of generating and releasing *synthetic data* (see also Little 1993; Reiter 2005). In this approach, the data custodian fits a model to the original data, then repeatedly draws from the model to generate multiple synthetic datasets which are released for analysis. The recently-developed *differential privacy* approach seeks to formalise the notion of confidentiality in the context of the output of algorithms performed on confidential databases, which includes statistical analysis (see Dwork et al. 2006; Dwork and Smith 2009). The most common method for achieving differential privacy is to add Laplace-distributed noise to the algorithm output, which unfortunately often results in inaccurate or misleading analysis results. The alternative approach of *remote analysis* has also been proposed (see for example Gomatam et al. 2005; Reiter 2003 and Sparks et al. 2008), and is the approach under active investigation by the ABS. A remote analysis system accepts a query from an analyst, runs it on data held in a secure environment, then returns confidentialised results to the analyst.

From the above discussion it should be clear that there are a number of different approaches to achieving a balance between allowing statistical analysis of confidential or private data and maintaining standards of privacy and confidentiality. Each approach has its own strengths and weaknesses, which means that there is no common approach suitable for every situation. It is important in any given situation to select the method which is most suitable for the given dataset, custodian, researcher, research project and regulatory environment. In this article we are interested in the remote analysis approach, which is being considered by at least one national statistical agency as a suitable replacement for remote access with manual output checking. It is our purpose to give an example of the sort of impact that confidentialisation of remote analysis outputs may have on exploratory data analysis, in order to better inform future research and choices about which confidentialisation approach to use in a given situation.

1.1. Remote Analysis

A remote analysis system accepts a query from an analyst, runs it on data held in a secure environment, then returns results to the analyst. In particular, the analyst does not have direct access to the data at all. In designing a remote analysis system to deliver useful results with acceptably low disclosure risk, restrictions can be imposed on the queries, the analysis itself can be modified and the results can be modified. In addition, the data can be restricted or altered, though this measure would seem to reduce the benefits of remote analysis over statistical disclosure control. A remote analysis system could be fully automated, or could involve some manual checking of queries or outputs. In the fully automated remote analysis system investigated in this article, we assume that the data are not restricted or altered, and we only suggest restricting the queries, modifying the analyses and modifying the results. We will call the modified results *confidentialised output.*

For reviews of remote analysis systems in use or in development in national statistical agencies, see (Brandt and Zwick 2010; Lucero and Zayatz 2010; Reuter and Museux 2010).

1.2. Scenarios in Which Remote Analysis May be Useful

It is unlikely in the foreseeable future that remote analysis systems will completely replace other data access modes such as the release of de-identified data or data which has undergone a statistical disclosure control process, or indeed remote access with manual output checking. This is largely because remote servers significantly reduce flexibility in analysis. However, there are some scenarios in which a remote analysis system may usefully augment these approaches, including:

- A remote analysis system could be used by an analyst as preparation before visiting a secure data laboratory. This would enable the analyst to learn about the data and formulate some initial analysis approaches with low disclosure risk. The analyst would then be able to make efficient, effective and informed use of a later session in a secure data laboratory. This is important because of the cost of secure data laboratory access to both the analyst and the administrative organisation.
- Access to confidential data through a remote analysis system may be viewed as "low risk" and so may require only a lightweight ethics approval process. This would enable an analyst to have an initial exploration of the data and perhaps find out whether a full ethics application for access to the data itself would be worthwhile.
- A remote analysis system could be used by an analyst to conduct preliminary investigations and obtain preliminary results, such as assessment of number of cases and statistical power through exploratory data analysis. Funding applications can be more favourably considered if these preliminary results have been obtained.

1.3. Related Work

Early proposals for remote analysis systems combined query restriction with statistical disclosure control on the source data (Duncan and Mukherjee 1991; Duncan and Pearson 1991; Keller-McNulty and Unger 1998; Schouten and Cigrang 2003). Later, the problem was considered in the special case of *table servers* designed to disseminate allowable marginal subtables of large, high-dimensional contingency tables (Dandekar 2004; Karr et al. 2003; Karr et al. 2002). An early discussion of remote analysis appeared in Reiter 2004.

A number of authors have addressed the problem of checking the output from an on-site data laboratory within a national statistical agency (see Corscadden et al. 2006; Honinger et al. 2010; Reznek 2003, 2006; Reznek and Riggs 2004, 2005; Ritchie 2006, 2007; and the summary guidelines in Brandt et al. 2010). In this approach, analysis outputs are classified as either *safe* or *unsafe*. Safe outputs are those which the researcher should expect to have cleared for release with no or minimal further changes, for example, the coefficients estimated from a survival analysis. Analytical outputs and estimated coefficients are usually classified as safe, except for a well-defined and limited number of exceptions. Unsafe outputs will not be cleared unless the researcher can demonstrate, to the output checker's satisfaction, that the particular context and content of the output makes it nondisclosive. For example, a table will not be released unless it can be demonstrated that there are enough observations, or the data have been transformed enough, so that the publication of that table would not lead to identification of outputs.

In this article we will compare our approach with the guidelines for the checking of output based on microdata research published in Brandt et al. (2010), as they represent the most recent and comprehensive treatment available. The paper also addresses the applicability of the guidelines to automatic disclosure control for remote data centres, and remote execution. The paper is an output of ESSnet SDC, a Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control (European Union n.d.).

The *differential privacy* approach seeks to formalise the notion of privacy in the context of algorithms performed on confidential information, which includes statistical analysis (see Dwork et al. 2006; Dwork and Smith 2009). An algorithm is differentially private essentially if its application to any two datasets that differ in a single element gives similar answers. Under the most common method for generating differentially private algorithms, Laplace-distributed noise is added to the algorithm output, which unfortunately often results in low data utility. Several improvements have been proposed in the literature (see for example Barak et al. 2007; Dwork and Lei 2009; Dwork et al. 2006 for results relevant to exploratory data analysis), however the problem of appropriately balancing disclosure risk and data utility in differentially private algorithms is not completely solved.

In the case of remote analysis for model fitting, most effort to date has been directed at linear regression. Gomatam et al. (2005) suggested ways to mitigate the effects of attacks for linear regression on a remote analysis system using transformations of variables (see Bleninger et al. 2010 for an empirical investigation).The authors also described disclosure risks associated with multiple, interacting queries to remote analysis systems, primarily in the context of remote regression analysis, and proposed quantifiable measures of risk and data utility. The challenge of confidentialising regression diagnostics has been addressed by Reiter (2003), Reiter and Kohnen (2005) and Sparks et al. (2008); see O’Keefe and Good (2009) for a detailed discussion and empirical investigation. Algorithms for obtaining differentially private regression coefficients are provided in Chaudhuri and Monteleoni 2008 and Smith 2009.

More generally, Sparks et al. (2008) proposed a range of measures for addressing disclosure risks in exploratory data analysis and model fitting for discrete or continuous response variables, and provided examples from biostatistics. O’Keefe et al. (2012) explored disclosure risks associated with survival analysis, and proposed measures to reduce the disclosure risk. The *Privacy-Preserving Analytics* (*PPA*) software demonstrator, described in Sparks et al. (2008), is an implementation of these measures for exploratory data analysis, statistical modelling including Generalised Linear Modelling, survival analysis, time series and clustering. Some of the measures involve the modification or restriction of standard statistical analyses submitted through a menu-driven interface, whereas others involve modifications to the output of fitted models. In particular, they do not involve applying any traditional statistical disclosure techniques to the underlying microdata (except in the case of using a random 95% sample of the microdata in some analyses).

The particular case of confidentialising exploratory data analysis output in remote analysis systems was discussed in Sparks et al. (2005) and later expanded in Sparks et al. (2008). The generality of the treatment in Sparks et al. (2008) makes it very difficult to see the range of disclosure risk reduction measures proposed for particular types of analysis. To address this gap, O’Keefe and Good (2008) and O’Keefe and Good (2009) provided a detailed discussion of the explicit confidentialisation measures in the case of linear regression, including a side-by-side comparison of the proposed confidentialised residual plots (using parallel boxplots) with plots of synthetic residuals. The current paper addresses the important case of exploratory data analysis in a similar way.

Apart from the problem of balancing disclosure risk with data utility, remote analysis systems present additional technical challenges in addressing, for example, missing data, outliers, selection bias testing, assumption checking and additional disclosure risks due to multiple, interacting queries.

1.4. Contents of This Article

As mentioned above, Sparks et al. (2008) have proposed methods by which the outputs from a range of individual statistical queries can be modified to reduce disclosure risk. Exploratory data analysis is an important special case, since it would be normal for an analyst approaching statistical analysis of any dataset to commence with exploratory data analysis. However, it is not easy to determine the applicable disclosure risk reduction methods proposed in Sparks et al. (2008), nor to understand their impact.

To address this gap, in this paper we provide a detailed and systematic study of the confidentialisation of exploratory data analysis output, such as could be implemented on a remote analysis system. We provide an analysis of relevant disclosure risks, and describe methods for addressing these risks. We also provide detailed examples which enable a side-by-side comparison of traditional with confidentialised exploratory data analysis output. We compare our approach with the guidelines for the checking of output based on microdata research developed by Brandt et al. (2010).

2. Exploratory Data Analysis in Remote Analysis

In this section we give a brief overview of exploratory data analysis, including some terminology, and describe the types of exploratory data analysis which will be the focus of this article.

We also discuss the main disclosure risks and associated confidentiality objectives for exploratory data analysis output from a remote analysis system.

2.1. Exploratory Data Analysis

Exploratory data analysis is concerned with developing an understanding of data, including exploring the nature of the distributions of the variables involved, and the relationships between the variables, (For more information on exploratory data analysis, see McNeil 1977; Mosteller and Tukey 1977; Tukey 1977; Velleman and Hoaglin 1981).

Velleman and Hoaglin (1981) outline four basic elements of exploratory data analysis, namely, data visualisation, residual analysis after model fitting, data transformation or re-expression and resistant procedures. For confidentialising residuals after model fitting and data transformation or re-expression, see the references in Section 1.3. In Sparks et al. (2008, Section 1.3) it is recommended that robust statistical methods be used when confidentialising output from a remote analysis system, and we will not directly address robust procedures further here.

The focus of this article will therefore be on exploratory data analysis through data visualisation. Methods for data visualisation commonly include:

Univariate Exploratory Data Analysis

1. Discrete variable
   (a) Frequency table
   (b) Bar chart or pie chart
2. Continuous variable
   (a) Summary statistics such as: number of observations, number of missing values, mean, median, sample minimum, sample maximum, quantiles such as quartiles, and standard deviation
   (b) Plot, dot chart, histogram or density estimate
   (c) Box plot
3. Discrete or continuous variable
   (a) Q-Q plots and P-P plots
   (b) Corresponding correlation coefficients

Bivariate and Multivariate Exploratory Data Analysis

4. Tabulation of frequencies for two or more discrete variables
5. Scatter plot of two continuous variables or scatter plot matrix for more than two continuous variables
6. Principal components analysis for two or more continuous variables
7. Parallel box plots or dot charts for a discrete and a continuous variable
8. Correlation coefficient for two variables or correlation matrix for more than two variables

In practice the analyst will choose which method(s) to use depending on their task at hand.

2.2. *Confidentialising Exploratory Data Analysis in Remote Analysis*

The key means by which identification of an individual might occur through an information release are direct identification, spontaneous recognition and matching to an external dataset. Direct identification occurs when an identifier such as name and address is read directly from a dataset. Spontaneous recognition occurs when an analyst recognises a data subject from an unusual combination of characteristics, such as being 105 years old and living in a certain suburb. Matching to an external dataset uses one or more variables common to both datasets as a matching key. If a match is found to an external dataset containing identifying information, then direct identification occurs. Otherwise, a match may be found to an external dataset with sufficient characteristics that spontaneous recognition occurs.

As in Sparks et al. (2008), the risk of direct identification can be minimised by ensuring that the results do not contain any directly identifying information. It is important to determine which variables are identifying, but examples include name, address and unique identifiers like government health care number. The risk of spontaneous recognition is minimised if the exact values of the variables are not disclosed for any individual. It may be important to know which variables carry the highest risk of spontaneous recognition to identify those which must be most strongly protected. The risk of matching is minimised if the exact values of the variables are not disclosed for any individual. Again, it may be important to know which variables are most useful as matching key variables to identify those that must be most strongly protected. For example, exact dates such as date of admission are extremely useful as matching keys. Thus, the results of a statistical analysis are unlikely to lead to identification of an individual if they contain no identifying information and if the exact values of variables corresponding to an individual (the unit record) are not disclosed. On the other hand, it is not always problematic to release a data value; for example if it is impossible to assign the data value to an individual data subject. In this article we have chosen to take the most conservative position of seeking to release no exact value of any variable corresponding to an individual, for two reasons. One is that we are envisaging an automated system which may have difficulty distinguishing risky from non-risky releases. The other is that we are interested in whether output could be useful even given this conservative position.

In the following, we consider only disclosure risk from a single exploratory data analysis request, though this might include a number of different analyses. In order to reduce risks associated with multiple, interacting queries, it would be necessary to implement a request tracking system which would identify and alert the system administrator to suspicious queries or query streams. While a full discussion of the identification of suspicious queries or query streams is beyond the scope of this article, examples might include a vast number of similar queries within a very short time frame, or queries for subsets that differ in only one individual data subject.

One of the main ways that disclosures of information about discrete variables can occur is through the existence of small numbers of data cases with a given combination of values (this is the problem of so-called *small cells* in tabular data). In addition, if a cell has a dominant observation (contributing more than 90% of the cell value, for example) or if it contains most (more than, say, 90%) of the observations in one of its variables, then disclosure risk can be unacceptable. Therefore many of the measures taken to confidentialise the output of exploratory data analysis simply ensure that each combination of variable values has sufficient data cases represented, through data winsorising or aggregation, and by rounding or smoothing of the results. (Under data winsorising, any observation which is more than 2.6 standard deviations above or below the mean is set to the mean plus or minus 2.6 standard deviations, respectively.)

The risk that the exact value of a variable is released in exploratory data analysis output is reduced by the following measures suggested in Sparks et al. (2008):

- Replace each table with a correspondence analysis plot
- Replace each scatter plot with confidentialised parallel box plots, where the procedure for constructing confidentialised parallel box plots is as follows:
  1. Determine which variable will be on the x-axis and which will be on the y-axis
  2. Determine the number of box plots to be constructed, by specifying intervals of the x-axis variable so that each interval has frequency at least at a minimum threshold value
  3. If the difference between the median and either the lower or upper quartile on a box plot is zero, amalgamate that interval with an adjacent interval and repeat until all box plots have distinct median, lower and upper quartiles
  4. For each interval, draw a confidentialised box plot as follows
     (a) Winsorise the data

   (b) Compute the new five summary statistics (minimum, lower quartile, median, upper quartile and maximum)
   (c) If the difference between the median and the lower or upper quartile is zero, then:
      (i) If the discrete variable is nominal (that is, categorical in which the categories have no natural order) then there is no natural way to amalgamate box plots, so provide no output
      (ii) If the discrete variable is ordinal (that is, categorical in which the categories have a natural order) then merge adjacent box plots until there is no remaining box plot with zero difference between the median and the lower or upper quartile
   (d) Round the resulting final values of the five summary statistics
   (e) Draw the parallel box plots using these final rounded values.
- Replace each plot of the estimate of an underlying probability density function (density estimate) with a confidentialised version, obtained by winsorising the data and rounding the sample minimum and maximum
- Replace each Q-Q plot or P-P plot with a confidentialised version, obtained as follows:
   1. Winsorise the data
   2. Fit a robust nonparametric regression line to the points *(x, y)* of the traditional Q-Q (respectively P-P) plot on the winsorised data.
- Replace each trend line with a confidentialised trend line, obtained as follows:
   1. Use Loess or Lowess (locally weighted scatter plot smoothing) to plot a smooth curve through the set of data points in the scatter plot (see Cleveland 1979; Cleveland and Devlin 1988)
   2. Winsorise or add noise to the end points of the curve to ensure that they do not reveal exact data values
- Round or otherwise perturb values of statistics such as medians, upper and lower quartiles, maxima and minima, as well as Pearson χ^2 statistics and Pearson product-moment correlation coefficients, since these are functions of the data values

In replacing a scatter plot with confidentialised parallel box plots, it is desirable that box plots of constant width be used to represent x variable intervals of the same length. For example, several different divisions into equal-width intervals could be tried until a division is found with each frequency at least at the minimum threshold value. However, it may occur that no such reasonable division can be found, and it is necessary to combine adjacent box plots to meet the frequency threshold. In this case, using a box plot of double width may be visually misleading as it tends to suggest double the mass of observations on that interval. An alternative is to delete one of the two combined box plots, as is in fact done in Figure 7(b).

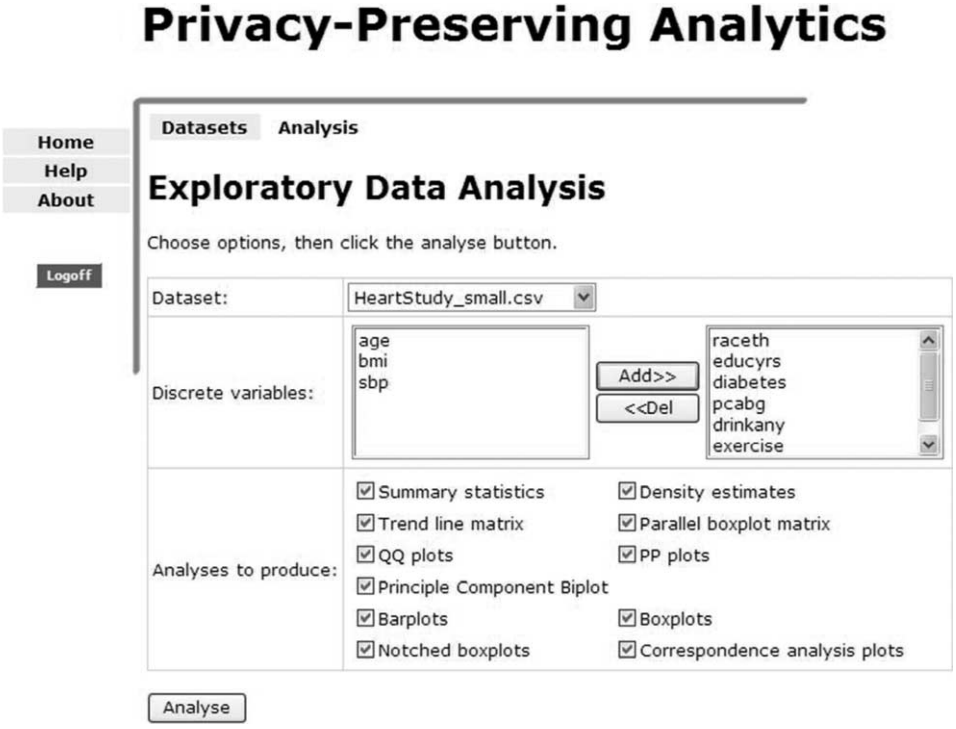
The suggested treatment of outliers with winsorisation has serious drawbacks. Analysts are not permitted to view outliers (since these present confidentiality risks) and so cannot make their own removal or treatment decisions. Instead, the remote analysis system removes outliers in the presented results, and alerts the analyst to the fact that removal has occurred. If these disadvantages are judged too serious in a given situation, the analyst may have to seek access to the unconfidentialised dataset through a different access mode.

3. Example of Remote Exploratory Data Analysis Output

In this section we provide a comprehensive example demonstrating the impact of implementing remote analysis system output confidentialisation measures, including those described in Section 2.2.

Figure 1 shows an example query input screen for all the exploratory data analyses conducted. After selecting the dataset from the drop down *Dataset:* menu, the analyst manually selects the *Discrete Variables.* (This should be unnecessary in a production system which would automate this step.) The analyst selects the desired exploratory data analyses and clicks the *Analyse* box. This menu-driven interface restricts the analyst to standard exploratory data analyses. Also, transformations or re-expressions of variables can reveal information about outliers, so these are not permitted. This restriction could potentially be relaxed in a production system after further disclosure risk evaluation.

In Sections 3.1 and 3.2 we provide comprehensive and representative examples of traditional and confldentialised exploratory data analysis outputs, on a publicly available dataset. In comparing the outputs, it is important to note differences in the scales because the removal of dataset outliers in the confidentialised output may cause a compression of the plot scale in comparison with the traditional output. We do not uncompress the scale, since the point of the example is to evaluate the information that can be deduced from the confidentialised output. The unconfidentialised output is provided to assist this evaluation. If we manipulate the confidentialised output, then it no longer represents the output



Privacy-Preserving Analytics

Exploratory Data Analysis

Choose options, then click the analyse button.

Dataset: HeartStudy_small.csv

Discrete variables: age, bmi, sbp

Analyses to produce:

☑ Summary statistics ☑ Density estimates
☑ Trend line matrix ☑ Parallel boxplot matrix
☑ QQ plots ☑ PP plots
☑ Principle Component Biplot
☑ Barplots ☑ Boxplots
☑ Notched boxplots ☑ Correspondence analysis plots

Fig. 1. Screen shot of query input interface for Exploratory Data Analysis

available to the analyst. We cross-reference the guidelines for the checking of output based on microdata research developed by Brandt et al. (2010).

While it would be ideal to use an example dataset from a national statistical agency, the confidentiality concerns which are the subject of this paper prevent it. Instead, we use a publicly available dataset with mostly categorical variables and some continuous variables, which is similar in this respect to many datasets housed in national statistical agencies. For the examples, we will use an extract of data from a study to test the safety and efficacy of estrogen plus progestin therapy to prevent recurrent coronary heart disease in postmenopausal women. The *Heart and Estrogen/Progestin Replacement Study (HERS)* data (Grady et al. 1998) contain information on the characteristics of 2763 participants in the HERS study. For our example, we will use the continuous variables: age in years (age), body mass index (bmi) and systolic blood pressure (sbp), and the discrete variables: ethnicity (raceth), years of education (educyrs), diabetes comorbidity (diabetes), insulin used (insulin), previous coronary artery bypass graft surgery (pcabg), at least one drink per day (drinkany) and attendance at exercise program or walking (exercise). The data are used for illustrative purposes only.

For the examples, the traditional output was generated within the R environment (R Development Core Team 2012), while most of the confidentialised output was generated with the PPA software demonstrator (see Sparks et al. 2008), however some of the confidentialised output was generated directly within the R environment.

3.1. Univariate Exploratory Data Analysis

3.1.1. Univariate Discrete Variable

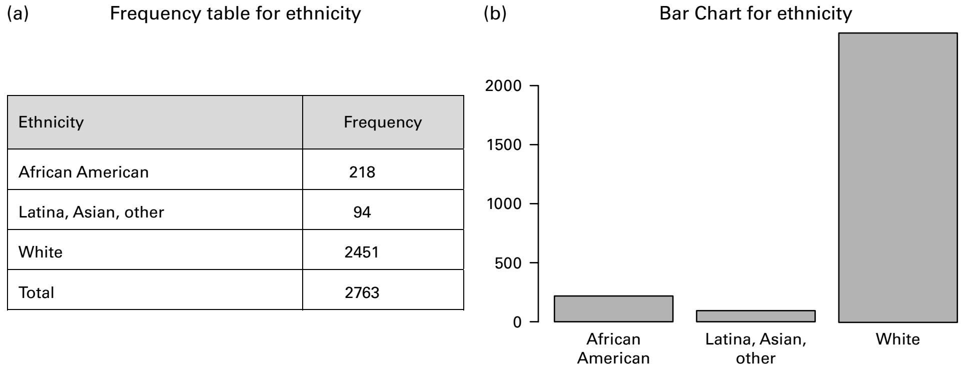
Exploratory data analysis output for a discrete variable would normally comprise a frequency table and bar chart. Confidentialising these outputs involves suppression or aggregation of categories to ensure that no category has less than a minimum threshold number of values (which could be set by the custodian) and no category contains more than, say, 90% of the observations. In this case, there are no small cells, so confidentialised output coincides with traditional output. An example of this type of output for the discrete variable ethnicity (raceth) in the HERS data is provided in Figure 2.

In this case there is no difference between traditional and confidentialised output. However, in general output may be suppressed or categories may be amalgamated in the confidentialised case.

For comparison, Brandt et al. (2010) also classify frequency tables as unsafe due to potential issues with small cells and cells which contain more than 90% of the total number of observations in one of its variables. If a frequency table is classified as unsafe, then it would either be suppressed or a tabular statistical disclosure limitation procedure would be applied; see Section 1 for references.

3.1.2. Univariate Continuous Variable

The mean and standard deviation would meet the disclosure risk objectives in Section 2.2 provided that there are sufficiently many observations contributing to their calculation. The minimum and maximum reveal data values and cannot be released. Similarly, the



Ethnicity	Frequency
African American	218
Latina, Asian, other	94
White	2451
Total	2763

Fig. 2. Traditional/confidentialised exploratory data analysis output for the discrete variable ethnicity (raceth) in the HERS data

median and lower and upper quartiles may reveal data values (depending on, for example, the parity of the number of observations), and are (conservatively) not released. Disclosure risk is reduced for these quantities through dataset winsorising and/or rounding of the values. A histogram would meet the disclosure risk objectives in Section 2.2 provided that there are no low interval frequencies. A density estimate may give information about outliers and minimum and maximum value in the dataset, and is confidentialised with the method described in Section 2.2. A plot, a dot chart and a box plot reveal observed data values, and so would not be permitted in confidentialised output of exploratory data analysis in a remote analysis system. Each of them can be replaced by a confidentialised box plot, constructed with the method described in Section 2.2.

In Figure 3 we show examples of traditional and confidentialised output of exploratory data analysis for the continuous variable age in years (age) in the HERS data. Traditional output in the form of a histogram and a box plot is shown in Figures 3(a) and 3(c) respectively, while confidentialised output in the form of a confidentialised density estimate and a confidentialised box plot is shown in Figures 3(b) and 3(d) respectively. The traditional histogram has one interval with a very small number of values which would be suppressed in confidentialised output. The text on Figure 3(b) and the '***' symbol in Figure 3(d) alert the analyst to the fact that the data in these cases have been winsorised.

The main difference between the traditional and confidentialised output is due to the data winsorising. Given only the confidentialised output in Figures 3(b) and 3(d), the analyst would only know that outliers had been removed. The analyst would not know the number of outliers removed, and would not know whether they were outliers with low or high age, or both. Despite this difference, the confidentialised density estimate in Figure 3(b) and the confidentialised box plot in Figure 3(d) both provide good general information about the shape of the variable distribution.

For comparison, Brandt et al. (2010) also classify mean, maximum, minimum and percentiles as unsafe due to concerns regarding small cells, dominant observations and cells which contain more than 90% of the total number of observations in one of its variables. Mode and standard deviation are classified as safe if there is no cell which contains more than 90% of the total number of observations in one of its variables. Graphs are generally classified as unsafe unless the underlying modified information used to

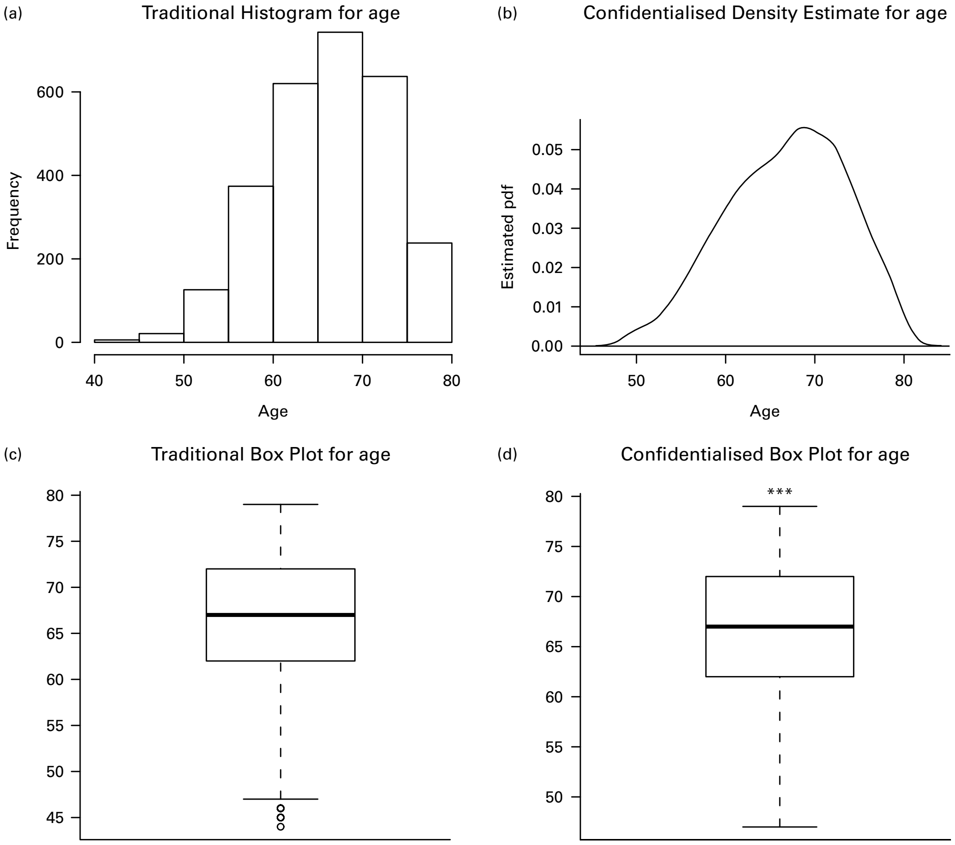


Fig. 3. Traditional and confidentialised exploratory data analysis output for the continuous variable age in years (age) in the HERS data

construct the graph has been classified as safe. For example, a safe graph would have no significant outliers and would not reveal any individual observation value.

3.1.3. Univariate Discrete or Continuous Variable

The confidentialisation of Q-Q plots and P-P plots is discussed in Section 2.2. The Pearson χ^2 statistic corresponding to a P-P plot meets the disclosure risk objectives in Section 2.2.

Figure 4 provides examples of traditional and confidentialised Q-Q plots (in Figures 4(a) and 4(b) respectively) and traditional and confidentialised P-P plots (in Figures 4(c) and 4(d) respectively). The Q-Q plots provide a comparison of the continuous variable age in years (age) sample data with the normal distribution. The P-P plots provide a comparison of the discrete variable years of education (educyrs) sample data with the Poisson distribution.

The rounded value of the Pearson χ^2 statistic for comparing the discrete variable years of education (educyrs) sample data with the Poisson distribution is 0.988, rounded from the true value of 0.9877984.

The confidentialised Q-Q plot in Figure 4(b) clearly shows the issues at the tails of the distribution apparent in the traditional Q-Q plot in Figure 4(a). The confidentialised and traditional P-P plots in Figures 4(c) and 4(d) are also of very similar shape to one another.

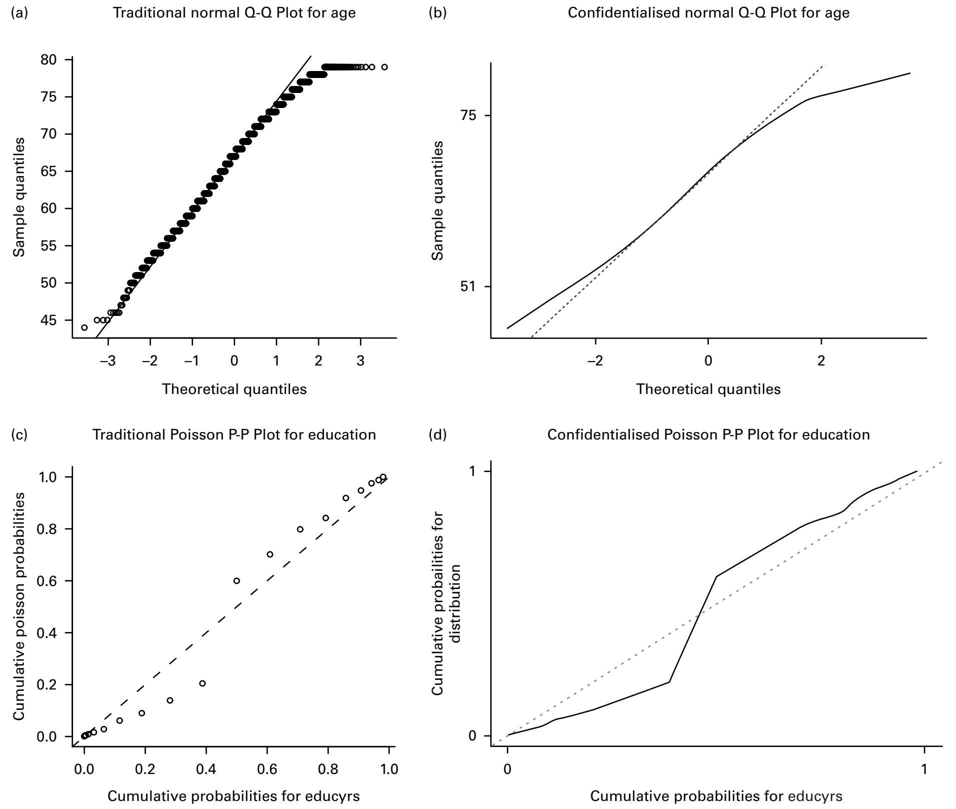


Fig. 4. Traditional and confidentialised exploratory data analysis output for the continuous variable age and discrete variable years of education (educyrs) in the HERS data

Although the confidentialised plots indicate that outliers have been deleted, in this case the confidentialisation procedure has not adversely affected the information provided in the plots. However, an analyst would need to be aware that in general the deletion of outliers in the plots may degrade the information presented at the tails of the plots.

For comparison, Brandt et al. (2010) classify plots as unsafe unless the underlying modified information used to construct the graph has been classified as safe. Test statistics such as χ^2 are classified as safe provided the model has at least ten degrees of freedom and at least ten units to produce the model.

3.2. Bivariate and Multivariate Exploratory Data Analysis

3.2.1. Two or More Discrete Variables

It has been long recognised that contingency tables in which there are cells with small counts or dominant observations represent a disclosure risk, since the existence of such cells increases the risk that individuals can be identified. There are many techniques proposed in the literature for confidentialising such tables, including rounding and cell suppression (see for example Domingo-Ferrer and Magkos 2010; Domingo-Ferrer and Torra 2004; Doyle et al. 2001).

Sparks et al. (2008, Section 2.2) propose that a correspondence analysis plot be provided instead of a confidentialised table. The plot would display the variable names, but individual data points would not appear on the plot. (The authors also suggest fitting a log-linear model, but we do not discuss this option here.) *Correspondence Analysis* (Benzecri 1973; Greenacre 2007) is a multivariate method for transforming a number of possibly correlated discrete variables into a number of uncorrelated variables *(principal components).* A *Correspondence Analysis plot of counts* is a graphical representation of the associations between the variables found during the correspondence analysis, see, for example Figure 5(c). As discussed in Sparks et al. (2008, Appendix A), the marginal totals of the matrix of counts together with the basic values can reveal information about the actual counts if the correspondence analysis explains nearly all of the variation. For this reason, the information is suppressed.

In Figure 5, we show examples of contingency tables and correspondence analysis plots for subsets of discrete variables in the HERS data. For the purpose of the tables and plots, the following further variable abbreviations are used: raceth = R, educyrs = Ed, diabetes = Di, insulin = I, pcabg = P, drinkany = Dr and exercise = Ex. The (partial)

(a) Traditional partial contingency table of ethnicity (R) and years of education (Ed)

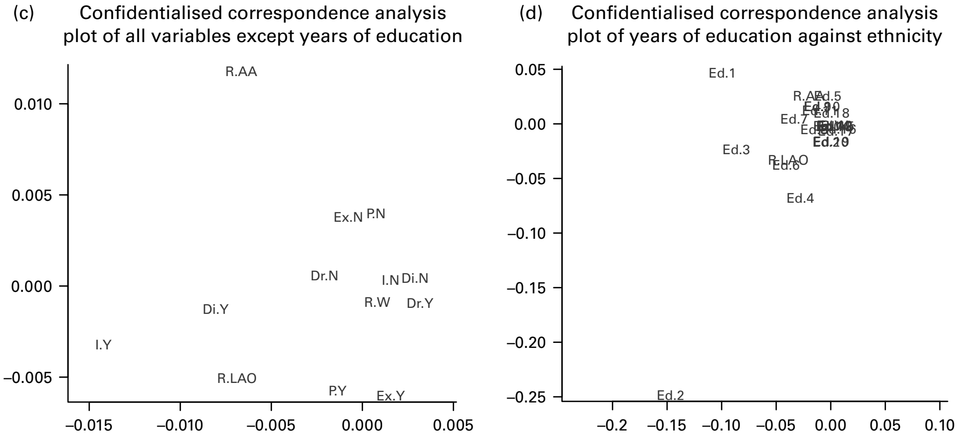
```
, , Di = N, P = N, Dr = N, Ex = N, I = N

     Ed
R      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  AA   0   0   0   0   0   0   2   2   2   5   6  13   3   3   0   0   1   1   0   0
  LAO  0   1   0   1   0   1   0   2   0   0   1   5   0   1   0   1   1   0   0   0
  W    0   0   0   1   3   3   6  14   9  18  38 158  34  32   7  21   9   7   2   4
```

(b) Traditional partial contingency table of ethnicity (R) and years of education (Ed)

```
, , Di = Y, P = N, Dr = N, Ex = Y, I = Y

     Ed
R      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  AA   0   0   0   0   0   0   0   0   2   1   1   2   0   0   0   1   0   1   0   0
  LAO  0   0   0   0   0   0   1   0   1   0   0   0   0   0   0   0   0   1   0   0
  W    0   0   0   0   0   0   0   0   0   0   1   6   3   1   0   1   1   0   0   0
```


(c) Confidentialised correspondence analysis plot of all variables except years of education

(d) Confidentialised correspondence analysis plot of years of education against ethnicity

Fig. 5. Traditional and confidentialised exploratory data analysis output for two or more discrete variables in the HERS data

contingency table in Figure 5(a) tabulates the values of the educyrs variable (Ed) against raceth (R), for the values Diabetes = No, pcabg = No, Drinkany = No, Exercise = No and Insulin = No, while the (partial) contingency table in Figure 5(b) tabulates the values of the educyrs variable (Ed) against raceth (R), for the values Diabetes Di = Yes, pcabg = No, Drinkany = No, Exercise = Yes and Insulin = Yes. The confidentialised correspondence analysis plot in Figure 5(c) shows the relationship between all variables except educyrs, while the correspondence analysis plot in Figure 5(d) shows the relationship between the variables educyrs (Ed) and raceth (R), where the number of years of education is indicated by a number and ethnicity codes are African American = AA, Latin, Asian or Other = LAO and White = W.

In a correspondence analysis plot, the distance between points indicates association, with the strength of the relationship indicated by the distance from the origin (0,0). For example, Figure 5(c) shows a strong relationship between insulin = N and diabetes = N, as would be expected. Figure 5(d) shows strong relationships between most pairs of values of educyrs and ethnicity, except, somewhat inexplicably, educyrs = 2.

There is information lost in replacing the contingency tables as in Figures 5(a) and 5(b) with correspondence analysis plots as in Figures 5(c) and 5(d). However, at least in this case, the sheer number of contingency tables makes it quite difficult to gain an overall view of the data. An analyst would be likely to try another exploratory data analysis approach or even some simple modelling. On the other hand, the correspondence analysis plots give overall trend information without underlying detailed information.

Brandt et al. (2010) classify frequency tables as as unsafe due to potential issues with small cells and cells which contain more than 90% of the total number of observations in one of their variables. If a frequency table is classified as unsafe, then it would either be suppressed or a tabular statistical disclosure limitation procedure would be applied; see Section 1 for references.

3.2.2. Two or More Continuous Variables

For several continuous variables, confidentialising a matrix of scatter plots would involve replacing it with a matrix of confidentialised parallel box plots and a matrix of confidentialised trend lines, as in Section 2.2. A principal components biplot can also be provided (Gabriel 1971; Greenacre 2010).

Figure 6 shows traditional output comprising two-dimensional scatter plots in Figure 6(a) and confidentialised output comprising parallel box plots in Figure 6(b) and confidentialised trend lines in Figure 6(c). Recall that the procedures for drawing these confidentialised plots are provided in Section 2.2.

The confidentialised output in Figure 6(b) shows similar information about the spread of variable values as the traditional output in Figure 6(a), although the analyst does need to take account of the fact that outliers have been removed.

It is perhaps surprising to note that the confidentialised output in Figures 6(b) and 6(c) arguably provide more information about variable value trends and the trends of relationships between variables, in comparison with the traditional output in Figure 6(a), The applicability of this observation is not restricted to remote analysis, and in fact it may be that analysts should construct un-confidentialised displays of parallel boxplots and un-confidentialised trend line matrices as part of routine exploratory data analysis.

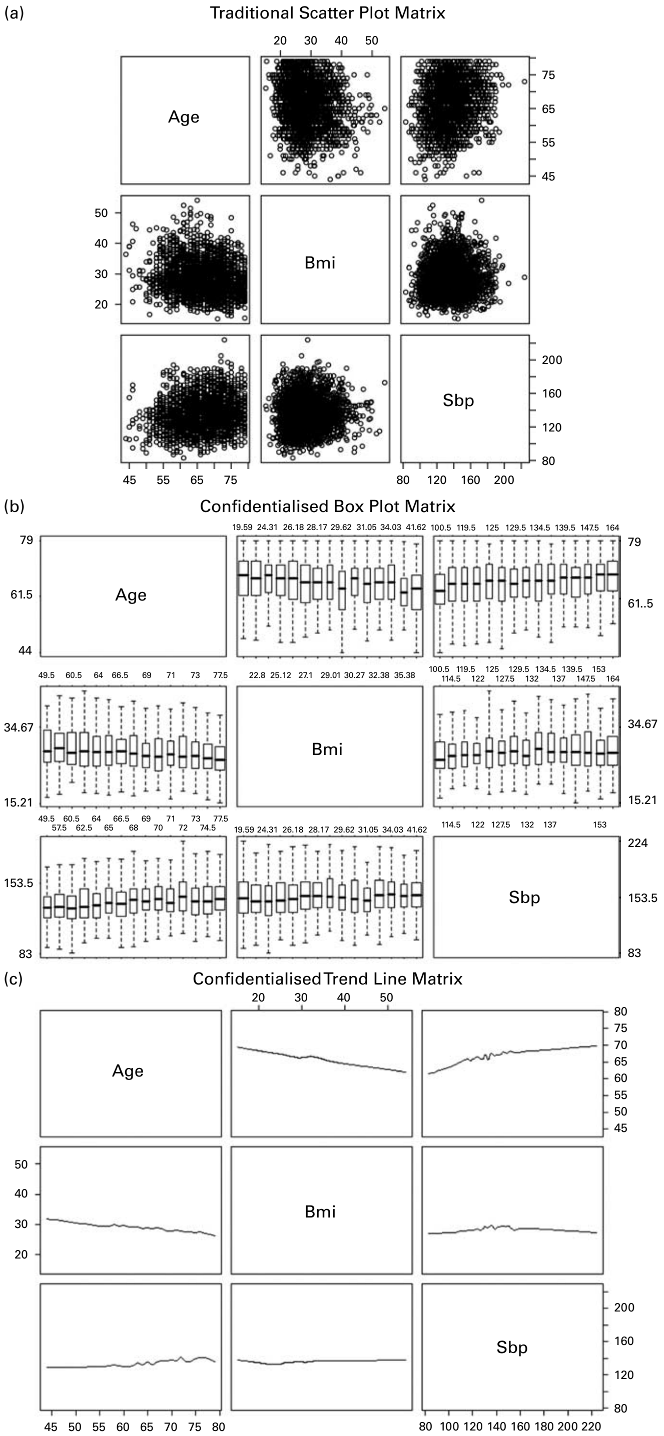


Fig. 6. Traditional and confidentialised exploratory data analysis output for pairwise continuous variables age in years (age), body mass index (bmi) and systolic blood pressure (sbp).

As noted in earlier sections, Brandt et al. (2010) classify graphs as unsafe unless the underlying modified information used to construct the graph has been classified as safe.

3.2.3. A Discrete and a Continuous Variable

Confidentialised parallel box plots can be provided as confidentialised output of exploratory data analysis for a discrete and a continuous variable in a remote analysis system. A dot chart reveals observed data values, and so would not be permitted in confidentialised output of exploratory data analysis in a remote analysis system.

Figure 7(a) shows traditional parallel box plots and Figure 7(b) shows confidentialised parallel box plots for the variables years of education and age in the HERS data.

The confidentialised plot alerts the analyst to the fact that outliers have been removed, and the analyst would be aware that the bin for the value 2 of years of education is missing, so must have had a small count and therefore have been amalgamated with an adjacent bin. However, the information provided to the analyst by the confidentialised parallel box plots output is very similar to the information provided in the unconfidentialised output.

Again, Brandt et al. (2010) classify graphs as unsafe unless the underlying modified information used to construct the graph has been classified as safe.

3.2.4. Correlation Coefficients

The rounded or perturbed Pearson product-moment correlation coefficient (Pearson 1896; Rodgers and Nicewander 1988) can be provided as confidentialised output of bivariate exploratory data analysis variables in a remote analysis system.

For comparison, Brandt et al. (2010) classify correlation coefficients as safe provided there are at least ten units contributing. However, they note that the publication of a correlation matrix which contains 0 or 1 and is connected to summary statistics may need further confidentialisation measures.

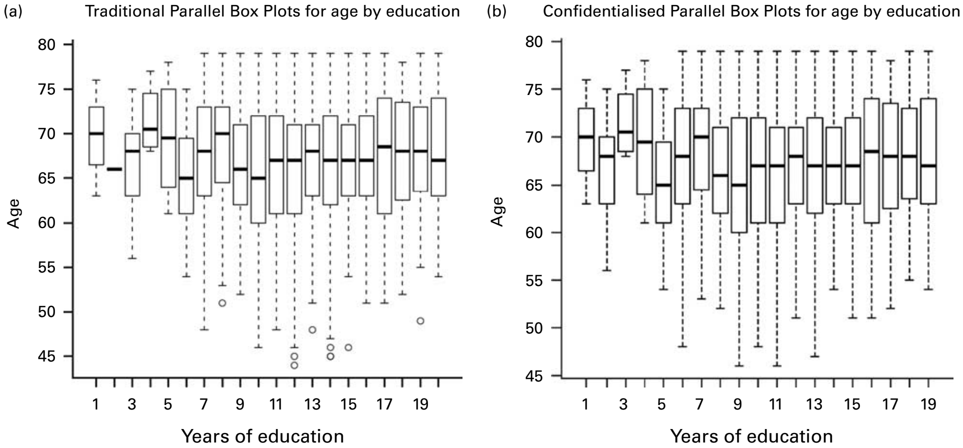


Fig. 7. Traditional and confidentialised exploratory data analysis output for the continuous variable age in years (age) by the discrete variable years of education (educyrs) in the HERS data

4. Discussion and Conclusions

In this article we have described a remote analysis system allowing exploratory data analysis on confidential data, including describing a number of scenarios in which this sort of functionality may be useful.

We provided an overview of disclosure risks and technical challenges in a remote analysis system. We then gave a detailed description of measures to confidentialise exploratory data analysis output, designed to achieve the disclosure risk objectives. The work clarifies and builds on the confidentiality objectives and some of the measures as discussed in Gomatam et al. (2005), Sparks et al. (2005), and Sparks et al. (2008). The measures are broadly in agreement with the guidelines for the checking of output based on microdata research developed by Brandt et al. (2010).

To illustrate the effect of the proposed confidentialisation methods, we provided a comprehensive example enabling a side-by-side comparison of traditional output and confidentialised output for a range of common exploratory data analyses.

The main differences between the traditional and confidentialised outputs were:

- Some plots showed differences in the scales because the removal of outliers in confidentialised plots caused compression of the plot scale.
- Data for some discrete variable categories could be suppressed or aggregated in the confidentialised output.
- Data winsorisation may mask information about outliers and or behaviour at the extremes of the dataset. The analyst would be aware that outliers had been removed, but would have no information about their number or values.
- The remote analysis system would not provide contingency tables, but rather would provide correspondence analysis plots. The analyst would have to obtain contingency tables using a different data access method.
- Continuous data are aggregated before presentation, for example as parallel box plots and trend lines instead of a scatter plot.
- Values of statistics and correlation coefficients would be rounded.

In the example presented, the confidentialised output generally provided good information about the data, except that outliers were removed and there was a general reduction in the amount of detail available.

In summary, we believe that the confidentialised output is still useful for exploratory data analysis, provided the analyst understands the confidentialisation process and its potential impact. Where the potential impact is judged to be too great, the analyst would need to seek another mode of access to the data.

It seems to be generally agreed that remote analysis servers will play an important role in the future of data dissemination (see for example Bleninger et al. 2010; Reiter 2004).

5. References

Adam, N. and Wortmann, J. (1989). Security-control Methods for Statistical Databases: A Comparative Study. ACM Computing Surveys, 21, 515–556.

Australian Bureau of Statistics (2009). Methodological News.

Australian Bureau of Statistics (n.d.). Available at: http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+(ABSDL) (accessed 20 December 2012).

Australian Bureau of Statistics (n.d.). Remote Access Data Laboratory (RADL). Available at: http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL) (accessed 20 December 2012).

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, Accuracy, and Consistency Too: a Holistic Solution to Contingency Table Release. In Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), 273–282.

Bénzecri, J.-P. (1973). L'Analyse des Données. Paris: Dunod.

Bleninger, P., Drechsler, J., and Ronning, G. (2010). Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study. In Privacy in Statistical Databases, Lecture Notes in Computer Science, J. Domingo-Ferrer and E. Magkos (eds). Vol. 6344. New York: Springer, 220–233.

Brandt, M., Franconi, L., Gurke, C, Hundepol, A., Lucarelli, M., Mol, J., Ritchie, F., Seri, G. and Welpton, R. (2010). Guidelines for the Checking of Outputs Based on Microdata Research. ESSnet SDC, A Network of Excellence in the European Statistical System in the Field of Statistical Disclosure Control. Available at neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf.

Brandt, M. and Zwick, M. (2010). Improvement of Data Access. The Long Way to Remote Data Access in Germany. Privacy in Statistical Databases Conference PSD. Short paper in CD proceedings.

Chaudhuri, K. and Monteleoni, C. (2008). Privacy-Preserving Logistic Regression. In Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS), 289–296.

Cleveland, W. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74, 829–836.

Cleveland, W. and Devlin, S. (1988). Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association, 83, 596–610.

Corscadden, L., Enright, J., Khoo, J., Krsinich, F., McDonald, S. and Zeng, I. (2006). Disclosure Assessment of Analytical Output. Statistics New Zealand Preprint.

Dandekar, R. (2004). Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data. In Privacy in Statistical Databases, Lecture Notes in Computer Science J. Domingo-Ferrer and V. Torra (eds), Vol. 3050. New York: Springer, 121–135.

Domingo-Ferrer, J. and Magkos, E. (2010). Privacy in Statistical Databases, Lecture Notes in Computer Science, Vol. 6344. New York: Springer.

Domingo-Ferrer, J. and Saygin, Y. (2008). Privacy in Statistical Databases, Lecture Notes in Computer Science, Vol. 5262. New York: Springer.

Domingo-Ferrer, J. and Torra, V. (2004). Privacy in Statistical Databases, Lecture Notes in Computer Science, Vol. 3050. New York: Springer.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam: North-Holland.

Duncan, G. and Mukherjee, S. (1991). Microdata Disclosure Limitation in Statistical Databases: Query Size and Random Sample Query Control. In Proceedings of the 1991 IEEE Symposium on Security and Privacy, 278–287.

Duncan, G. and Pearson, R. (1991). In Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future. Statistical Science, 6, 219–239.

Duncan, G.T., Keller-McNulty, S.A. and Stokes, S.L. (2001). Disclosure Risk vs Data Utility: The r-u confidentiality Map, Technical Report LA-UR-01-6428, Los Alamos National Laboratory.

Dwork, C. and Lei, J. (2009). Differential Privacy and Robust Statistics. In Proceedings of the 41st ACM Symposium on Theory of Computing (STOC), 371–380.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In 3rd IACR Theory of Cryptography Conference, 265–284.

Dwork, C. and Smith, A. (2009). In Differential Privacy for Statistics: What We Know and What We Want to Learn. Journal of Privacy and Confidentiality, 1, 135–154.

European Union (n.d.) Essnet project. http://neon.vb.cbs.nl/casc/index.htm (accessed 20 December 2012).

Gabriel, K. (1971). The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. Biometrika, 58, 453–467.

Gomatam, S., Karr, A., Reiter, J., and Sanil, A. (2005). Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Systems. Statisical Science, 20, 163–177.

Grady, D., Applegate, W., Bush, T., Furberg, C., Riggs, B., and Hulley, S. (1998). Heart and Estrogen/Progestin Replacement Study (hers): Design, Methods, and Baseline Characteristics. Controlled Clinical Trials, 19, 314–335.

Greenacre, M. (2007). Correspondence Analysis in Practice. London: Academic Press.

Greenacre, M. (2010). Biplots in Practice. Madrid: BBVA Foundation.

Honinger, J., Pattloch, D., and Voshage, R. (2010). On-site Access to Micro Data: Preserving the Treasure, Preventing Disclosure. Preprint.

Karr, A., Dobra, A., and Sanil, A. (2003). Table Servers Protect Confidentiality in Tabular Data Releases. Communications of the ACM, 46, 57–58.

Karr, A., Lee, J., Sanil, A., Hernandez, J., Karimi, S., and Litwin, K. (2002). Web-Based Systems that Disseminate Information but Protect Confidentiality. Advances in Digital Government: Technology, Human Factors and Public Policy, W. Mclver and A. Elmagarmid (eds). Kluwer: Amsterdam, 181–196.

Keller-McNulty, S. and Unger, E. (1998). A Database System Prototype for Remote Access to information Based on Confidential Data. Journal of Official Statistics, 14, 347–360.

Little, R. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Lucero, J. and Zayatz, L. (2010). The Microdata Analysis System at the U.S. Census Bureau. Privacy in Statistical Database, Lecture Notes in Computer Science J. Domingo-Ferrer and E. Magkos (eds), Vol. 6344. New York: Springer, 234–248.

McNeil, D. (1977). Interactive Data Analysis. Hoboken, NJ: Wiley.

Mosteller, F. and Tukey, J. (1977). Data Analysis and Regression. Boston: Addison-Wesley.

Office for National Statistics (n.d.). Available from: http://www.ons.gov.uk/ons/about-ons/who-we-are/services/vml/index.html (accessed 20 December 2012).

Office of Information and Regulatory Affairs (1994). Statistical policy working paper 22 – report on statistical disclosure limitation methodology, Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

O'Keefe, C.M. and Good, N. (2008). A Remote Analysis System – What Does Regression Output Look Like? In Privacy in Statistical Databases, number 5262 Lecture Notes in Computer Science J. Domingo-Ferrer and Y. Saygin (eds). New York: Springer, 270–283.

O'Keefe, C.M. and Good, N. (2009). Regression Output From a Remote Analysis System. Data & Knowledge Engeneering, 68, 1175–1186.

O'Keefe, C.M., Sparks, R., McAullay, D. and Loong, B. (2012). Confidentialising Survival Analysis output in a Remote Data Access System. Journal of Privacy and Confidentiality, 4, 127–154.

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution, iii. Regression, Heredity and Panmixia. Philisophical Transactions of the Royal Society A, 187, 253–318.

R Development Core Team (2012). R:A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Available from: www.R-project.org/.

Reiter, J. (2003). Model Diagnostics for Remote-Access Regression System. Statistical Computing, 13, 371–380.

Reiter, J. (2004). New Approaches to Data Dissemination: A Glimpse into the Future (?). Chance, 17, 12–16.

Reiter, J. (2005). Using Cart to Generate Partially Synthetic Public Use Microdata. Journal of Official Statistics, 21, 441–462.

Reiter, J. and Kohnen, C. (2005). Categorical Data Regression Diagnostics for Remote Systems. Journal of Statistical Computation and Simulation, 75, 889–903.

Reuter, W. and Museux, J.-M. (2010). Establishing an Infrastructure for Remote Access to Microdata at Eurostat. Privacy in Statistical Databases, Lecture Notes in Computer Science J. Domingo-Ferrer and E. Magkos (eds), Vol. 6344. New York: Springer, 249–257.

Reznek, A. (2003). Disclosure Risks in Cross-Section Regression Models. American Statistical Association 2003, Proceedings of the Section on Government Statistics and Section on Social Statistics, CD, 3444–3451.

Reznek, A. (2006). Recent Confidentiality Research Related to Access to Enterprise Microdata. Prepared for the Comparative Analysis of Enterprise Microdata (CAED) Conference, Chicago IL, USA.

Reznek, A. and Riggs, T. (2005). Disclosure Risks in Releasing Output Based on Regression Residuals. American Statistical Association 2005 Proceedings of the Section on Government Statistics and Section on Social Statistics (available on CD), 1397–1404.

Reznek, A. and Riggs, T.L. (2004). Disclosure Risks in Regression Models: Some Further Results. American Statistical Association 2004 Proceedings of the Section on Government Statistics and Section on Social Statistics, 1701–1708.

Ritchie, F. (2006). Disclosure Controls for Regression Outputs. Mimeo, Office of National Statistics, London.

Ritchie, F. (2007). Disclosure Detection in Research Environments in Practice. Working paper 37 in the Joint UNECE/Eurostat work session on statistical data confidentiality. Topic (iii): Applications; United Nations Statistical Commission and Economic Commission for Europe Conference of Europe Statisticians, European Commission Statistical Office of the European Communities (Eurostat), Manchester. Available at: www.unece.org/stats/documents/2007/12/confidentiality/wp. 37.e.pdf

Rodgers, J. and Nicewander, W. (1988). Thirteen Ways to Look at the Correlation Coefficient. The American Statistician, 42, 59–66.

Rubin, D. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 462–468.

Schouten, B. and Cigrang, M. (2003). Remote Access Systems for Statistical Analysis of Microdata. Statistical Computing, 13, 371–380.

Smith, A. (2009). Asymptotically Optimal and Private Statistical Estimation. Proceeding of CANS 2009. LNCS 5888, J. Garay, A. Miyaji, and A. Otsuka (eds), Berlin: Springer.

Sparks, R., Carter, C, Donnelly, J., Duncan, J., O'Keefe, C. and Ryan, L. (2005). A Framework for Performing Statistical Analyses of Unit Record Health Data Without Violating Either Privacy or Confidentiality of Individuals. In Proceedings of the 55th Session of the International Statistical Institute, Sydney.

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., Keighley, T., and McAullay, D. (2008). Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-preserving Analytics™, Computer Methods and Programs in Biomedicine, 91, 208–222.

Tukey, J. (1977). Exploratory Data Analysis: Addison-Wesley.

UK Data Archive (n.d.). Secure data service, Available from: http://securedata.data-archive.ac.uk/ (accessed 20 December 2012).

United States Census Bureau (n.d.). Available from: http://www.census.gov/ces/rdcresearch/ (accessed 20 December 2012).

University of Chicago (n.d.). Available at: www.norc.org (accessed 20 December 2012).

Velleman, P. and Hoaglin, D. (1981). The ABC's of EDA: Applications, Basics and Computing of Exploratory Data Analysis. Boston, MA: Duxbury Press.

Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control, Lecture Notes in Statistics, Vol. 155: Springer.

Received 2012
Revised November 2012

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 615–619

Editorial Collaborators

The editors wish to thank the following referees who have generously given their time and skills to the Journal of Official Statistics during the period October 1, 2011–September 30, 2012. An asterisk indicates that the referee served more than once during the period.

Abts, Koen, KU Leuven, Leuven, Netherlands
Adiguzel, Feray, Erasmus University, Rotterdam, The Netherlands
Alam, Moudud, Dalarna University, Borlänge, Sweden
Alldritt, Richard, UK Statistics Authority, London, UK
Antal, Erika, University of Neuchâtel, Neuchâtel, Switzerland
Ash, Stephen, U.S. Census Bureau, Arlington, VA, U.S.A.
Ballano, Carlos, Instituto Nacional de Estadística de España, Madrid, Spain
Barboza, Wendy, USDA/NASS, Fairfax, VA, U.S.A.
Beaumont*, Jean-Francois, Statistics Canada, Ottawa, Canada
Beck, Jennifer, U.S. Census Bureau, Washington DC, U.S.A.
Bediako, Grace, Ghana Statistical Service, Accra, Ghana
Berent Matthew K, Sharon, PA, U.S.A.
Bethlehem, Jelke, Statistics Netherlands, The Hague, The Netherlands
Bhattacharya, Debopam, University of Oxford, Oxford, UK
Billiet, Jaak, Katholieke Universiteit Leuven, Leuven, Belgium
Blom, Annelies, University of Mannheim, Mannheim, Germany
Booleman, Max, Statistics Netherlands, Voorburg, The Netherlands
Brion, Philippe, INSEE, Paris, France
Broome, Jessica, University of Michigan, Ann Arbor, MI, U.S.A.
Brunton-Smith, Ian, University of Surrey, Surrey, UK
Burton, Levine, RTI International, NC, U.S.A.
Callegaro, Mario, Google, Mountain View, CA, U.S.A.
Carey*, James, Centers for Disease Control & Prevention, Atlanta, GA, U.S.A.
Chauvet, Guillaume, ENSAI, Bruz Cedex, France
Chen, Qixuan, Columbia University, New York, U.S.A.
Chipperfield*, James, University of Wollongong, Wollongong, Australia.
Cho, Eungchun, Kentucky State University, Kentucky, U.S.A.
Cho, Moon, U.S. Bureau of Labor Statistics, Washington, U.S.A.
Christian, Leah, Pew Research Center, Washington DC, U.S.A.
Chun Young, Asaph, NORC University of Chicago, Washington DC, U.S.A.
Cohen*, Stephen, National Science Foundation, Arlington, Virginia, U.S.A.
Coleman, Shirley, Newcastle University, Newcastle, UK
Creel, Darryl V., RTI International, Research Triangle Park, NC, U.S.A.
Cyr, André, Statistics Canada, Ontario, Ottawa, Canada
Dahlhamer, James, National Center for Health, Hyattsville, MD, U.S.A.
Das, Marcel, Tilburg University, Tilburg, The Netherlands
D'Aurizio, Leandro, Bank of Italy, Rome, Italy
Davidson, Russell, McGill University, Montreal Quebec, Canada

De Beer, Joop, NIDI, The Hague, The Netherlands
Del Barrio Castro, Tomas, University of the Balearic Islands, Palma de Mallorca, Spain
De Luna, Xavier, Umeå University, Umeå, Sweden
Dixon, John, U.S. Bureau of Labor Statistics, Washington DC, U.S.A.
Dorfman, Alan, U.S. Bureau of Labor Statistics, Washington DC, U.S.A.
Durand, Claire, University of Montreal, Montreal, Québec, Canada
Eckman, Stephanie, University of Maryland, College Park, Maryland, U.S.A.
Escobar, Emilio, University of Southampton, Southampton, UK
Falorsi*, Stefano, ISTAT, Roma, Italy
Frank*, Ove, Stockholm University, Stockholm, Sweden
Franses, Hans, Erasmus University Rotterdam, Rotterdam, The Netherlands
Freiman, Michael*, U.S. Census Bureau, Colombia, Maryland, U.S.A.
Fuchs, Marek, Technische Universität Darmstadt, Darmstadt, Germany
Funke, Frederik, Staufenberg, Germany
Galesic Mirta, Max Planck Institute for Human Development, Berlin, Germany
Ganninger, Matthias, ZUMA, Mannheim, Germany
Gao, Jingjing, Emory University, Atlanta, Georgia, U.S.A.
Garcia, Maria M, U.S. Census Bureau, Washington DC, U.S.A.
Gareth, James, UK Office for National Statistics, Newport, UK
Giles, David, University of Victoria, Victoria, B.C., Canada
Giusti, Caterina, University of Pisa, Pisa, Italy
Golinelli, Roberto, University of Bologna, Bologna, Italy
Gonzalez, Jeffrey, U.S. Census Bureau of Labor Statistics, Washington, DC, U.S.A.
Grafström, Anton, University of Umeå, Umeå, Sweden
Graham*, Patrick, University of Otago, Christchurch, New Zealand
Graziadei, Connie, Statistics Canada, Ottawa, Canada
Groen, Jeffrey, U.S. Bureau of Labor of Statistics, Washington DC, U.S.A.
Göritz, Anja, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany
He, Chong, University of Missouri, Columbia, MO, U.S.A.
Herrmann, Douglas, Terre Haute, U.S.A.
Hidiroglou, Mike, Statistics Canada, Ottawa, Ontario, Canada
Holbrook, Allyson, Survey Research Laboratory, Chicago, U.S.A.
Hoogendoorn, Adriaan, GGZ inGeest, Amsterdam, Amsterdam, The Netherlands
Ilves, Maiki, Örebro University, Örebro, Sweden
Jansson, Thomas, Sveriges Riksbank, Stockholm, Sweden
Jäckle, Annette, University of Essex, Colchester, UK
Kadane, Joseph, B. Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.
Kaminska, Olena, University of Essex, Colchester, United Kingdom
Kennel*, Timothy, U.S Census Bureau, Washington, DC U.S.A.
Kim, Jae-Kwang, Iowa State University, Ames, IA, U.S.A.
Kinney, Satkartar, NISS, Research Triangle Park, NC, U.S.A.
Kirkendall, Nancy, The committee on National Statistics, Washington, DC, U.S.A.
Kloek, Wim, Eurostat, Luxembourg, Luxemburg
Koch, Achim, ZUMA, Mannheim, Germany
Kohler, Ulrich, Wissenschaftszentrum Berlin, Berlin, Germany
Kozak, Marcin, Warsaw Agricultural University, Warsaw, Poland
Kuusela, Vesa, Espoo, Finland
Laflamme, Francois, Statistics Canada, Ottawa, Canada

Laitila, Thomas, Statistics Sweden, Örebro, Sweden
Lane, Julia, National Science Foundation, Arlington, U.S.A.
Laurie, Heather*, University of Essex, Colchester, UK
Lavrakas*, Paul J., Flagstaff, Arizona, U.S.A.
Lee, Katherine, Murdoch Children Research Institute, Melbourne, Australia
Lee, Sunghee, University of Michigan, Ann Arbor, Michigan, U.S.A.
Lehtonen, Risto, University of Helsinki, Helsinki, Finland
Levine, Burton, RTI International, NC, U.S.A.
Linacre, Susan, Australian Bureau of Statistics, Canberra, Australia
Lipps, Oliver, FORS, University of Lausanne, Lausanne, Switzerland
Little J.A. Roderick*, University of Michigan, Ann Arbor, Michigan, U.S.A.
Ljones, Olav, Statistics Norway, Oslo, Norway
Loosveldt*, Geert, Katholieke Universiteit Leuven, Leuven, Belgium
Lopez-Escobar Emilio, University of Southampton, Southampton, UK
Lundqusit, Peter, Statistics Sweden, Stockholm, Sweden
Lussier, Robert, Quebec, Canada
Mahon-Haft, Taj, Radford University, Radford VA, U.S.A.
Malhotra, Neil, Stanford University, Stanford, CA, U.S.A.
Malmdin, Joakim, Statistics Sweden, Stockholm, Sweden
McConway, Kevin, The Open University, Milton Keynes, UK
McDonald, John, Institute of Education, Quantitative Social Statistics, London, UK
McGonagle, Katherine, University of Michigan, Ann Arbor, Michigan, U.S.A.
McIsaac, Michael, University of Waterloo, Ontario, Canada
Mecatti*, Fulvia, University of Milan-Bicocca, Milan, Italy
Meyermann, Alexia, University of Bielefeld, Bielefeld, Germany
Miller, Peter, U.S. Census Bureau, Washington, U.S.A.
Montaquila, Jill, Westat, Rockville, Maryland, U.S.A.
Mulry, Mary, U.S. Census Bureau, Washington, U.S.A.
Nandram, Balgobin, Worcester Polytechnic Institute, Worcester, MA, U.S.A.
Nerbonne*, John, Rijksuniversiteit Groningen, Groningen, The Netherlands
Nordbotten, Svein, University of Bergen, Bergen, Norway
Oganyan, Anna, Georgia Southern, University, Statesboro, GA, U.S.A.
O'Keefe, Christine, CSIRO, Melbourne, Australia
Olin, Jens, Statistics Sweden, Örebro, Sweden
Olson, Kristen, University of Nebraska-Lincoln, Lincoln, NE, U.S.A.
Opsomer, Jean, Colorado State University, Fort Collins, Colorado, U.S.A.
Pannekoek, Jeroen, Statistics Netherlands, The Hague, The Netherlands
Pascale, Joanne, U.S. Census Bureau, Washington D.C., U.S.A.
Peytchev, Andrey, RTI International, Research Triangle Park, NC, U.S.A.
Persson, Andreas, Statistics Sweden, Örebro, Sweden
Phelps, Andrew, NatCen Social Research, London, UK
Phipps, Polly, U.S. Bureau of Labor Statistics, Washington, U.S.A.
Piersimoni, Federica, ISTAT, Rome, Italy
Pink, Brian, Australian Bureau of Statistics, Canberra, Australia
Plewis, Ian, University of Manchester, Manchester, UK
Polasek, Wolfgang, Institute for Advanced Studies, Vienna, Austria
Pratesi, Monica, University of Pisa, Pisa, Italy
Pursiainen, Heikki, VATT, Helsinki, Finland

Qualité, Lionel, University of Neuchâtel, Neuchâtel, Switzerland
Raghunathan, Trivellore, University of Michigan, Ann Arbor, MI, U.S.A.
Ranalli, Giovanna, University of Perugia, Perugia, Italy
Reichert, Jennifer, U.S. Census Bureau, Washington, U.S.A.
Reips, Ulf-Dietrich, Univerisidad de Deusto, Bilbao, Spain
Reist, Benjamin, U.S. Census Bureau, Washington, U.S.A.
Ritchie, Felix, University of the West of England, Bristol, UK
Roberts, Caroline, University of Lausanne, Lausanne, Switzerland
Roose, Henk, Ghent University, Ghent, Belgium
Rosen, Jeffrey, RTI International, Chicago, Illinois, U.S.A.
Ryten, Jacob, London Ontario, Ontario, Canada
Sakshaug*, Joseph, University of Michigan, Ann Arbor, MI, U.S.A.
Salamin*, Paul-André, University of Applied Sciences, Sion, Switzerland
Scheuren, Fritz, NORC, Alexandria, VA, U.S.A.
Schouten, Barry, Statistics Netherlands, The Hague, The Netherlands,
Schräpler, Jörg-Peter, Ruhr-Universität Bochum, Bochum, Germany
Schulte Nordholt, Eric, Statistics Netherlands, Voorburg, Netherlands
Seastrom, Marilyn, National Center for Education Statistics, Washington, U.S.A.
Shikano, Susumu, University of Konstanz, Konstanz, Germany
Sikov, Anna, Hebrew University, Jerusalem, Israel
Sillajõe, Tuulikki, Statistics Estonia, Tallinn, Estonia
Slanta, John, Arlington, Virginia, U.S.A.
Slavec, Ana, University of Ljubljana, Ljubljana, Slovenia
Slud, Eric, University of Maryland, College Park, MD, U.S.A.
Sverchkov, Michail, U.S. Bureau of Labor Statistics, Washington DC, U.S.A.
Smith, Paul, Office of National Statistics, Newport, UK
Smyth, Jolene, University of Nebraska-Lincoln, Lincoln, NE U.S.A.
Sova, Markus, Office for National Statistics, Newport, United Kingdom
Spagat, Michael, Royal Holloway, University of London, Egham Surrey, UK
Stolzenberg, Stephanie, RTI International, Research Triangle Park, NC U.S.A.
Stoop, Ineke, I&A – Data Services and IT, The Hague, The Netherlands
Tang*, Yuqing, U.S. food and Drug Admin, Yorkville, IL, U.S.A.
Thompson, Katherine, U.S. Census Bureau, Washington, U.S.A.
Thompson, Mary, University of Waterloo, Canada
Thorburn, Daniel, Stockholm University, Stockholm, Sweden
Toepoel, Vera, Tilburg University, Tilburg, The Netherlands
Tucker, Clyde, Independent Consultant, Vienna, VA, U.S.A.
Valente, Paolo, United Nations Economic Commission for Europé, Geneva, Switzerland
Valliant, Richard, University of Maryland, College Park, MD, U.S.A.
Vandenplas, Caroline, UNIL-FORS, Lausanne, Switzerland
Van der Loo, Mark, Statistics Netherlands, The Hague, The Netherlands
Van Oest, Rutger, BI Norwegian Business School, Oslo, Norway
Vercruyssen, Anina, Ghent University, Gent, Belgium
Vorst, Harrie, University of Amsterdam, Amsterdam, The Netherlands
Wagner*, James, University of Michigan, Michigan, U.S.A.
Watson, Nichole, University of Melbourne, Melbourne, Australia
Weale, Martin, NIESR, London, UK
Weisman, Ethan, IMF Statistics, District of Columbia, U.S.A.

Wenemark, Marika, University of Linköping, Linköping, Sweden
Wieczorek, U.S. Census Bureau, Washington, U.S.A.
Wilke, Ralf, Nottingham School of Economics, Nottingham, UK
Wooden, Mark, University of Melbourne, Melbourne, Australia
Wretman, Jan, Stockholm University, Stockholm, Sweden
Wrighte, Duncan, Statistics Canada, Ottawa, Ontario, Canada
Wu, Changbao, University of Waterloo, Waterloo, Canada
Xianchao, Xie, Harvard University, Cambridge, MA, U.S.A.
Yang, Michael Y., NORC, University of Chicago, Bethesda, MD, U.S.A.
Young*, Linda, University of Florida, Gainesville, Florida, U.S.A.
Yung, Wesley, Statistics Canada, Ottawa, Canada
Zio Di, Marco, ISTAT, Rome, Italy
Zwick, Markus, Goethe-Universität Frankfurt am Main, Germany

Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 621–623

Index to Volume 28, 2012

Contents of Volume 28, Numbers 1–4

Author Index

Book Reviews

Printed in December 2012