Journal of Official Statistics vol. 32, i. 2 (2016)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 259–286, http://dx.doi.org/10.1515/JOS-2016-0013

On a Modular Approach to the Design of Integrated Social Surveys

Evangelos Ioannidis[1], *Takis Merkouris*[2], *Li-Chun Zhang*[3], *Martin Karlberg*[4], *Michalis Petrakos*[5], *Fernando Reis*[6], *and Photis Stavropoulos*[7]

This article considers a modular approach to the design of integrated social surveys. The approach consists of grouping variables into 'modules', each of which is then allocated to one or more 'instruments'. Each instrument is then administered to a random sample of population units, and each sample unit responds to all modules of the instrument. This approach offers a way of designing a system of integrated social surveys that balances the need to limit the cost and the need to obtain sufficient information. The allocation of the modules to instruments draws on the methodology of split questionnaire designs. The composition of the instruments, that is, how the modules are allocated to instruments, and the corresponding sample sizes are obtained as a solution to an optimisation problem. This optimisation involves minimisation of respondent burden and data collection cost, while respecting certain design constraints usually encountered in practice. These constraints may include, for example, the level of precision required and dependencies between the variables. We propose using a random search algorithm to find approximate optimal solutions to this problem. The algorithm is proved to fulfil conditions that ensure convergence to the global optimum and can also produce an efficient design for a split questionnaire.

Key words: Efficient design; respondent burden; sample allocation; simulated annealing; split questionnaire.

1. Introduction

In recent years, several national statistical agencies have explored the possibility of integrating social surveys, as a means of meeting growing demand for new and improved

[1] Athens University of Economics and Business, Patission str. 76, 10434 Athens, Greece. Email: eioannid@aueb.gr
[2] Athens University of Economics and Business, Patission str. 76, 10434 Athens, Greece. Email: merkouris@aueb.gr
[3] University of Southampton, Southampton, SO17 1BJ, United Kingdom and Statistics Norway. Email: l.zhang@soton.ac.uk
[4] European Commission (Eurostat), L-2920 Luxembourg. Email: Martin.KARLBERG@ec.europa.eu
[5] Agilis SA, Acadimias 96-100, 10677 Athens, Greece, Email: Michalis.Petrakos@agilis-sa.gr
[6] European Commission (Eurostat), L-2920 Luxembourg. Email: Fernando.REIS@ec.europa.eu
[7] Agilis SA, Acadimias 96-100, 10677 Athens, Greece. Email: Photis.Stavropoulos@agilis-sa.gr

Acknowledgments: The work presented in this article was part of a methodological consultancy project (contract number 61001.2011.005-2012.426) undertaken by Agilis SA (Athens, Greece) on behalf of the European Commission (Eurostat). The participation of the members of the Department of Statistics of the Athens University of Economics and Business in this project was coordinated by the University's Research Centre. We would like to thank the Associate Editor and the three referees for their valuable comments and suggestions, which have led to a considerable improvement of the manuscript.

Disclaimer: The views expressed here are those of the authors and do not necessarily reflect the official views of the European Commission (Eurostat).

statistics, while at the same time streamlining the survey operations in order to curb costs and limit respondent burden. See, for example, the experience of the UK Office for National Statistics (Smith 2009), the Australian Bureau of Statistics (2012), and the Dutch Central Bureau of Statistics (Cuppen et al. 2013).

In this article, we consider a general approach to survey integration that includes all the social surveys managed by a national statistical agency. This approach is designed to provide a flexible framework that will be able to accommodate the needs of social statistics as they change over time. The starting point for this research was the Eurostat project *Streamlining and Integration of the European Social Surveys* (see Reis 2013). This project was set up to support the implementation of the ESS Vision (European Commission 2009), the strategy for modernising the European Statistical System (ESS), in particular the European system of social statistics, in accordance with the commitments made in the Wiesbaden memorandum (European Statistical System Committee 2011).

The approach presented in this article is based on a modular design. The basic features are as follows. The existing ‘items’ (variables) in social surveys are restructured into a number of mutually exclusive groups, called modules, each of which consists of a small number of items. The modules are themselves distributed among an appropriate number of ‘instruments’, in such a way that each instrument consists of a fixed set of modules and each module is present in one or more instruments. These instruments will together replace the current set of survey questionnaires. Each instrument will be administered to a probability sample of population units within a specified time period. All the units in a sample will be asked to respond to all the items in the instrument assigned to them, except where routing determines otherwise. The modular design is therefore mainly characterised by: (i) the composition of the instruments; and (ii) their associated sample sizes.

The modularisation approach offers a number of potential advantages over the traditional ‘stovepipe’ survey programmes. Both the cost of carrying out the survey and the burden placed on respondents can be reduced, subject to precision requirements, for example by putting modules with lower precision requirements in instruments that are administered to fewer sample units, or by combining several instruments that all contain the same module. The modularisation can enhance the analytical potential of survey data by means of a suitable composition of the instruments, instead of by simply adding more modules to existing surveys. Likewise, new modules can be introduced with a short lead time and in a more cost-effective way, thus allowing emerging needs to be better met.

Designing and implementing a modular approach is by no means straightforward, however. There are specific methodological and technical problems that need to be addressed, and there are also a number of broader issues, some of which we will discuss here. In this article we focus on two questions: (1) how to develop a modular design that is sufficiently versatile to support an integrated social survey system; and (2) how to find a solution to the dual design problem of instrument composition and sample size allocation. We also refer to Karlberg et al. (2015) for a nontechnical description of this approach to modular design, and to Reis (2013) for a discussion of related survey-management and subject-matter issues.

The existing survey methodology of split questionnaire design (SQD) is most akin to the modular design. For a single survey the SQD allows different sets of items to be collected from different sample units; the full questionnaire consists of the union of all these *split* *questionnaires*. The SQD has attracted increasing attention in recent years. See, for example, Raghunathan and Grizzle (1995) for a Bayesian approach, or Chipperfield and Steel (2009) for a design-based approach. While the SQD is traditionally used to lessen the high respondent burden caused by a long questionnaire, it also provides a possible design for survey integration. For example, two existing separate and independent sample surveys can be considered jointly as a single survey with a two subsample SQD. Issues relating to efficient design and estimation for a single SQD have been studied recently in Chipperfield and Steel (2009; 2011). Efficient estimation in a broadened context of SQD has been studied more extensively in Merkouris (2010; 2015), building on earlier methods for composite estimation that combined data from different sample surveys, in order to increase precision and align estimates relating to the same item (see Renssen and Nieuwenbroek 1997 and Merkouris 2004; 2013).

There are, nevertheless, a number of difficulties that make it impossible to fully endorse an SQD model for the integration of social surveys. For example, it is not clear how to accommodate the difference in the timing and frequency of different surveys, or how to implement rotating panels overlapping over time. As we explain in Section 2, these situations can be accommodated under the approach proposed in this article.

To formulate the modular design problem, denote the set of all, say m, modules by $\{M_1, \ldots, M_m\}$, and let the variables in module M_i be $\{X_{i,j}, j = 1, \ldots, \nu_i\}$. Denote the set of all, say k, instruments by $\{I_1, \ldots, I_k\}$ and the associated samples of sizes $n_1, \ldots, n_k$ by $\{S_1, \ldots, S_k\}$. Further, denote the number of modules in instrument I_j by m_j and the set of instruments containing module M_a by $\Theta_a \subseteq \{I_1, \ldots, I_k\}$. Then, for a given set of modules and a predetermined number of instruments, the design problem is to determine the composition of instruments $I_1, \ldots, I_k$ and the corresponding sample sizes $n_1, \ldots, n_k$, in such a way as to minimise a given cost function while respecting multiple constraints. The cost function is a generic loss function determined by the sample sizes and a number of other features of the integrated survey. The constraints are primarily lower bounds on the precision of the estimation for individual modules, and for groups of modules that have to be placed in the same instrument. Such groups will be called 'mandatory crossings'. Other constraints that need to be taken into account relate to features of particular modules and to the choice of which instrument to put them in. For example, modules that have to be administered with a certain periodicity (e.g., quarterly or annually) should be put in instruments of the same periodicity. These constraints mean that the solution to the optimisation problem (i.e., constrained minimisation of the cost function) can only be sought among certain instrument compositions, which we will call 'admissible'. It should be noted that we do not consider the alternative setup of the optimisation problem — in which the information collected is maximised, subject to constraints on cost, precision and admissibility — because it is unclear how the objective function would need to be defined in this case.

Table 1 provides a generic illustration of the structure of instruments and modules. Suppose, initially, that there are two usual sample surveys. Survey A is carried out quarterly, with sample size of 1,000 per quarter, and contains modules M_1, M_2, and M_3. Survey B is conducted annually, with sample size of 10,000, and contains only module M_4. The precision requirements are given in terms of the required annual sample sizes for the respective modules. Assuming for simplicity that all modules contribute equally to the cost function, these two surveys together cost $4*3*1{,}000 + 10{,}000 = 22{,}000$ person

Table 1. Illustration of the composition of instruments in terms of modules

Possible optimal reorganisation of a quarterly survey of size 1,000 per quarter with modules M_1, M_2, M_3 and a further annual survey of size 10,000 with a single module M_4

Instruments (after optimisation)	Name	$I_{1.Q1}$	$I_{1.Q2}$	$I_{1.Q3}$	$I_{1.Q4}$	I_2
	Sample	S_1	S_2	S_3	S_4	S_5
	Sample size	1,000	1,000	1,000	1,000	8,000

Modules	Module variables	Required periodicity	Required Annual sample size	Annual sample size before optimisation	Annual sample size after optimisation	Instrument composition (modules) (after optimisation)				
M_1	$\{X_{1,1},\ldots,X_{1,v_1}\}$	Quarterly	4,000	4,000	4,000	**1**	**1**	**1**	**1**	0
M_2	$\{X_{2,1},\ldots,X_{2,v_2}\}$	Annual	2,000	4,000	2,000	**1**	0	**1**	0	0
M_3	$\{X_{3,1},\ldots,X_{3,v_3}\}$	Annual	1,000	4,000	1,000	0	**1**	0	0	0
M_4	$\{X_{4,1},\ldots,X_{4,v_4}\}$	Annual	10,000	10,000	10,000	**1**	**1**	0	0	**1**

Crossings	Crossing members	Required periodicity		Annual sample size before optimisation	Annual sample size after optimisation	Instrument composition (modules) (after optimisation)				
G_1	$\{M_2, M_4\}$	Annual	1,000	0	1,000	**1**	0	0	0	0
G_2	$\{M_3, M_4\}$	Annual	1,000	0	1,000	0	**1**	0	0	0

modules. Reorganising the modules into the four quarterly instruments $I_{1.Q1}$, $I_{1.Q2}$, $I_{1.Q3}$, $I_{1.Q4}$ and the annual instrument I_2, as illustrated in Table 1, reduces the cost to $(4+2+1+2) * 1{,}000 + 8{,}000 = 17{,}000$ person modules. Moreover, the reorganisation can be used to accommodate additional information needs, illustrated here by the crossings $G_1 = \{M_2, M_4\}$ and $G_2 = \{M_3, M_4\}$, with the required respective sample sizes. Normally, the required crossings could be obtained by including, for example M_2 and M_3 in Survey B, or M_4 in Survey A, thus further increasing the total cost of the survey.

In this article, we propose to perform the optimisation described for all admissible instrument compositions using a random search algorithm known as simulated annealing (see e.g., Kirkpatrick et al. 1983). We apply the simulated annealing algorithm in a manner adapted to this problem, and show that certain conditions ensuring convergence with probability 1 to the global optimum are satisfied. This means, in practice, that a sufficiently large number of iterations of this algorithm should yield an acceptable approximate solution to the optimisation problem. At each iteration of the search algorithm, we use the simplex algorithm to determine the optimal sample sizes for each instrument, subject to the given precision requirements. This allows us to evaluate the cost of each admissible instrument composition.

The rest of the article is organised as follows. In Section 2, we consider the various constraints that may be encountered in practice and describe how these are built into the optimisation procedure. To simplify the exposition, we start by assuming simple random sampling, and discuss the case of complex sampling designs later. In Section 3, we describe the proposed random search algorithm, which can be used to find approximate solutions to the optimisation problem in practice. The framework of Sections 2 and 3 is extended in Section 4 to allow for complex sampling designs. In Section 5, we give an illustration of the modular design approach using the EU Labour Force Survey as an example. Section 6 contains some concluding remarks.

2. Constraints on Modules and Instruments

In this section, we discuss a number of constraints on modules and instruments. We explain how they can be incorporated into the optimisation framework, so that the algorithm only visits admissible solutions. The main set of constraints arises from the requirement to achieve a given level of precision in estimating parameters (e.g., totals, means and proportions) for the variables of each module. Other constraints relate to the joint observation of modules and the periodicity of modules and instruments.

2.1. *Precision Requirements for Modules*

Precision requirements for a variable are often formulated in terms of the variance, denoted by $V(\hat{\theta})$, of an estimator $\hat{\theta}$ of some finite-population parameter θ (total, mean or proportion). The first step is to determine the sample size that is required to achieve a specified level of precision. We start by assuming that the samples for all instruments are selected using simple random sampling and independently from one another (in Section 4 we discuss departures from this assumption). Let us first consider the Horvitz-Thompson (HT) estimator $\hat{\theta}$ of θ based on the sample of a single instrument. Then the variance of $\hat{\theta}$ is $V(\hat{\theta}) = \sigma_\theta^2/n$, assuming a negligible sampling fraction n/N. Here N is the population size and $\sigma_\theta^2 = S^2$, the finite-population variance, if θ is a mean, and $\sigma_\theta^2 = N\theta\,(1-\theta)/(N-1)$ if θ is a proportion. Assuming that $\hat{\theta}$ is normally distributed, the precision requirement defined as the attainment of a certain margin of error $e = \left|\hat{\theta} - \theta\right|$ with probability $1-\alpha$ is satisfied if $V(\hat{\theta}) \leq \left(e/z_{1-\alpha}\right)^2$, where $z_{1-\alpha}$ is the $1-\alpha/2$ standard normal quantile. Thus, in the case of a single instrument, the precision requirement is satisfied by

$$n \geq n^*, \text{ where } n^* = z_{1-\alpha}^2 \sigma_\theta^2 / e^2. \qquad (1)$$

We could alternatively start with the coefficient of variation $CV(\hat{\theta}) = \sqrt{V(\hat{\theta})}/\theta$ and formulate the precision requirement as: $CV(\hat{\theta}) \leq c$ for some constant c. This then translates into $n \geq n^*$, where $n^* = \sigma_\theta^2/c^2\theta^2$.

If a variable belongs to a module that is present in more than one instrument, then θ may be estimated, using the combined data from the associated samples, as the weighted average $\hat{\theta} = \sum w_i \hat{\theta}_i$, where the summation extends over all instruments that contain the variable of interest, $w_i = n_i/\sum n_j$, and n_i is the size of the ith sample. It is easy to show that if these simple random samples are independent, this composite estimator will have a minimum variance given by $V(\hat{\theta}) = \sigma_\theta^2/\sum n_i$, for negligible sampling fractions n_i/N.

Where a variable is present in a number of instruments, the precision requirement can therefore be satisfied by

$$\sum n_j \geq n^*. \tag{2}$$

When considering all the variables in a module M_a, which is present in instruments $I_j \in \Theta_a$, this process could either be repeated for all variables in the module, with the maximum of the required sample sizes being chosen, or one of the variables could be considered as the 'main variable', and the sample size requirement of the module set in accordance with the precision requirement for this variable. In both cases, the module's precision requirement would be satisfied by

$$\sum_{\Theta_a} n_j \geq n_a^*, \tag{3}$$

for some appropriate n_a^*, which is obtained as above. Here, the summation extends over Θ_a, the set of all instruments containing module M_a. Thus, if, for example, $\Theta_a = \{I_3\}$, then we require $n_3 \geq n_a^*$, while if $\Theta_a = \{I_1, I_3, I_7\}$, then we require $n_1 + n_3 + n_7 \geq n_a^*$.

These m linear constraints, one for each module, on the sample sizes $n_1, \ldots, n_k$ of the various samples, can be represented in the $m \times k$ 'composition matrix' $\mathbf{A}$, with elements $a_{i,j}$, where $a_{i,j} = 1$ if the module in row i is present in the instrument in column j and $a_{i,j} = 0$ otherwise. The precision requirements for all modules may thus be expressed as the following set of constraints on the vector $\mathbf{n} = (n_1, \ldots, n_k)'$ of the sample sizes of all instruments:

$$\mathbf{An} \geq \mathbf{n}^*, \ \mathbf{A} \text{ is an } m \times k \text{ matrix}, \tag{4}$$

where the inequality is to be understood componentwise, and $\mathbf{n}^* = \left(n_1^*, \ldots, n_m^*\right)'$ is the vector of the minimal sample size required for each module.

2.2. Mandatory Crossings

In the previous section we discussed sample size requirements determined by the need for a certain level of precision in estimating the parameters of the marginal distribution of the variables in a module. In this section, we consider the sample size requirements relating to mandatory crossings, that is, the sample sizes needed to ensure that the requirement for a group of modules to be simultaneously present in at least one instrument is met. Mandatory crossings make it possible to carry out multivariate analyses as they ensure that all the relevant information is collected from the same sample units. Multivariate analysis can include, for example, estimating regression coefficients or estimating the average of a variable conditionally on the value of another variable (e.g., unemployment given the level of educational attainment). Clearly, for a mandatory crossing a suitable sample size requirement must be specified, so as to ensure estimation of desired precision for the parameters of the multivariate analysis.

Crossings are added as separate rows, $i = m + 1, \ldots, m'$ in the $\mathbf{A}$ matrix. If the crossing is, say, in the i_0th row, then the corresponding sample size requirement is given by $n_{i_0}^*$. The idea is that row i_0 will act as a 'constraint', forcing all modules that are members of the crossing in row i_0 to be jointly present in some instruments, which have a total sample size of at least $n^*_{i_0}$. If the crossing is included in the instrument corresponding to column j_0, we set $a_{i_0,j_0} = 1$, and $a_{i,j_0} = 1$ for all rows i corresponding to modules that are members of the crossing. This means that for a given column of the matrix **A** some rows will have to be set jointly due to mandatory crossings. Note, however, that each member M_i may be included in further instruments beyond the ones in which the crossing is included, for example, if M_i has a sample size requirement n^*_i greater than $n^*_{i_0}$.

Thus, the extended sample size constraints in modules and crossings can be expressed as

$$\mathbf{An} \geq \mathbf{n}^*, \ \mathbf{A} \text{ is an } m' \times k \text{ matrix.} \tag{5}$$

2.3. Periodicity and Concordance of Modules and Instruments

In addition to guaranteeing a certain sample size for individual modules and crossings, the choice of instruments must also take into account the periodicity of the modules: for example, certain modules may be administered on a quarterly basis, others on a yearly basis. In order for each module to be administered to a population sample at the required periodicity, the instruments will also be assigned a periodicity, at which they will be administered to sample units. The resulting instruments, one for each wave of data collection, belong together to a common ‘parent’ set of instruments. Thus, a quarterly parent instrument is sprouted into four quarterly ‘sibling’ instruments and a monthly parent instrument is sprouted into twelve monthly sibling instruments. We will label all sibling instruments using two indices, the first of which will indicate the common parent, and the second the siblings. Thus, for example, a parent set of instruments I_i with monthly period will consist of sibling instruments $\{I_{i.1}, \ldots, I_{i.12}\}$, one of which is to be administered to a population sample each month. It should be noted that although our approach does not require this, sibling instruments belonging to the same parent will usually have a similar core of modules, as each parent instrument would typically have been created to accommodate the repeated administration of a set of modules. Furthermore, although I_i is actually a set of instruments, we will use the same notation as for a single instrument when the context is clear. We will need twelve columns in the composition matrix, one for each sibling instrument, to represent the parent set of instruments $I_i := \{I_{i.1}, \ldots, I_{i.12}\}$ accurately.

In order to ensure that a module is administered with the appropriate periodicity and with the correct sample sizes, we will assume that all instruments belonging to the parent I_i are administered equally spaced in time and to samples of identical sizes. Moreover, a module may only belong to members of I_i if the period of the module (e.g., twelve for an annual module, three for a quarterly module) divided by the period of the parent instrument I_i is an integer, say r. The module must thus participate in a ‘complete periodic subset’ of the parent instrument, which is defined as a subset of the form $\{I_{i.s}, I_{i.(s+r)}, I_{i.(s+2r)}, \ldots\}$ of I_i or as a union of such subsets. For example, if a quarterly module is included in a monthly parent instrument I_i, then it must be included either in $\{I_{i.1}, I_{i.4}, I_{i.7}, I_{i.10}\}$, or in $\{I_{i.2}, I_{i.5}, I_{i.8}, I_{i.11}\}$, or in $\{I_{i.3}, I_{i.6}, I_{i.9}, I_{i.12}\}$, or in a union of such subsets.

This means that for a given row (i.e., a module) of the matrix **A**, its elements for the columns corresponding to a complete subset of instruments for the given module will have to be set jointly. See the ‘proposal mechanism’ discussed in Subsection 3.2 for an explanation of how the proposed random search algorithm deals with such constraints.

A module (or mandatory crossing) cannot be included in an instrument whose period is greater than that of the module. Thus, a quarterly module may be present in quarterly or monthly instruments, but not in annual instruments. If it is nevertheless desired to include a module in an instrument of a lower frequency, for example in order to cross it with a module appearing only in that instrument, then this module should be duplicated and should appear in the list of modules once with the greater period and once with the lower period. The same restriction holds for crossings, their period being defined as the lowest of those of the member modules. Indeed, the period of each member module must be an integer multiple of the period of the whole crossing.

The constraint for sibling instruments to be allocated equal sample sizes is represented by a balancing matrix **B**, with k columns. Each set of sibling instruments with p siblings corresponds to a set of $p - 1$ rows of the matrix **B**. Each such row, say i_0, will ensure that $n_j = n_{j'}$ for one of $j' = j + 1, \ldots, j + p$. We therefore set $b_{i_0,j} = 1$ and $b_{i_0,j'} = -1$, and $b_{i_0,j''} = 0$ for $j'' \notin \{j, j'\}$. We then require $\mathbf{Bn} = \mathbf{0}$. It should be noted that the balancing matrix **B** is a constant in the optimisation process, unlike the composition matrix **A**.

2.4. *Other Constraints on Modules and Instruments*

In this section we describe some additional constraints that may arise in practice and indicate how they could be expressed in terms of admissible instrument compositions.

2.4.1. Constraints on the Presence of Modules in Specific Instruments

There are certain modules that should be included either a) in all instruments or b) in specific instruments only:

a. Modules containing variables used as auxiliaries in weight calibration. These are usually categorical variables for which the population totals by category are known, such as demographic variables. As calibration increases the precision of estimates of correlated variables, and as we are interested in bringing together information from different instruments, it is advisable to require a core set of such modules to be included in all instruments. Similarly, it would be possible to include certain modules related to rare population groups or low prevalence characteristics of interest in all instruments, so as to gather information on these items from all units of the integrated sample.
b. Modules for which longitudinal information is required. These modules should only be included in appropriate instruments (e.g., where there is a specified sample overlap between successive waves).

It should therefore be possible, within the process of optimisation, to specify whether a certain module should be present in all instruments, or only in a specified group of instruments. This is achieved by specifying which instrument compositions are ‘admissible’ (those satisfying all necessary specifications) and by allowing the algorithm to visit only these compositions.

2.4.2. Constraints on the Joint Presence of Modules in Some Instruments

a. Dependencies between modules. It may only make sense for respondents to answer questions from one module if they have responded in a particular way to another module. For example, it only makes sense for an individual to answer a question on hours worked if he/she has responded positively to the question as to whether he/she works at all.
b. Many surveys currently being used may include groups of modules that relate to the same thematic block and that are thus logically related to one another. Breaking up such thematic blocks may be confusing for respondents, may make it difficult for the questionnaire to focus on certain issues and may make training of interviewers more difficult. It is therefore advisable to allow the grouping of modules into thematic blocks, and to either include or exclude whole thematic blocks from any instrument.

In view of the above, the optimisation algorithm should exclude from the admissible compositions any composition where an instrument includes a module, say M, but not all the modules on which M depends. The requirement that certain modules only appear as part of a specific thematic block could be imposed by building a mandatory crossing, the sample size requirement of which would be equal to the maximum of the sample size requirements of the modules contained in the block. Alternatively, if a group of modules needs to appear together in all instruments, then such a group could be treated as a single 'supermodule' rather than as a mandatory crossing. This would reduce the computational complexity of the optimisation. Finally, if a certain group of modules should not all appear in the same instrument under any circumstances, then any composition where an instrument contains all the modules from this group will be made inadmissible.

2.4.3. Putting Limits on the Questionnaire Size of An Instrument

There are two ways of ensuring that the questionnaire is of a reasonable size:

The first is the 'hard' or rule-based approach: instrument compositions are only considered admissible if the sum of the burden of the modules belonging to the instrument is below a certain upper bound.

The second is the 'soft', more flexible approach: the questionnaire size is incorporated into the cost function. The cost function — which we discuss in Subsection 3.1 below —is mainly determined by the sample size and the unit cost of a questionnaire, with the unit cost depending on the size of the questionnaire. In Subsection 3.1, we model unit cost as increasing linearly with the size of the questionnaire. Instead, it could be set to increase linearly up to a certain threshold, and then to increase at an accelerated rate (e.g., quadratically) above the threshold. This creates a pressure for the questionnaire size to stay below this threshold, while leaving some flexibility such that exceptions could be allowed, the 'price' of going above this threshold being a higher unit cost.

3. Optimisation over Instrument Compositions and Sample Sizes

In this section, we describe an approach to optimisation over instrument composition and sample size allocation that respects the constraints described above and yields an approximately optimal modular design.

3.1. *The Cost Function*

The cost function should ideally capture all the relevant contributing factors, including both the cost of producing the survey and the respondents' burden. In reality, this has proven to be impossible in any strict sense, not least because a national statistical agency will necessarily manage multiple surveys in parallel, and the methods and technical systems for conducting surveys are constantly evolving. For the sake of simplicity, we will follow the standard approach and use a linear cost function.

The overall cost, C, is assumed to be the sum of the cost of all instruments (samples) considered:

$$C = \sum\nolimits_{j=1}^{k} C_j.$$

The cost C_j of sample j is assumed to be a linear function of the number of individuals in the sample. This follows Groves (1989, 51) and Cochran (1977, 280). It is composed of a survey-specific fixed cost $C_j^{(f)}$ — the cost incurred regardless of the sample size chosen — and a variable cost, which is assumed (as an approximation) to increase linearly with the sample size,

$$C_j = C_j^{(f)} + \alpha_j n_j.$$

The variable cost coefficient α_j is the unit cost. Its dependence on j allows different unit costs to be specified for different instruments, reflecting, for example, different modes of data collection.

The linearity of C in the sample sizes n_j, as expressed by the two preceding equations, is an important assumption, and one which is necessary for the optimisation algorithm (Subsection 3.2.2, Step ii). A realistic calculation of the cost of each instrument is, of course, important to the design; see, for example UNSTATS (2005, 249–300) for an extensive discussion on the assessment of survey costs. In what follows, we further elaborate on a possible decomposition of the unit cost α_j. The proposed optimisation algorithm is, however, applicable regardless of the details of this decomposition.

Let α_j have a household and individual component,

$$\alpha_j = \overline{q}^{-1}\alpha_j^{(hh)} + \alpha_j^{(ind)},$$

where $\alpha_j^{(hh)}$ is the cost of including a household in the sample j, $\alpha_j^{(ind)}$ the cost of including an individual in the sample and $\overline{q}$ the average number of individuals in the household. The coefficients $\alpha_j^{(hh)}$ and $\alpha_j^{(ind)}$ may consist of a fixed cost β_0 (per person or household) and a variable cost, which increases with the respondent burden of the respective questionnaire.

The burden of the questionnaire can be measured in such a way as to appropriately reflect its cognitive and operational burden, taking into account, for example, whether the respondent needs to look at specific personal records or use complex recall processes in order to be able to answer the questions. The burden could, for example, be measured as the average amount of the interviewer's time required, as in Chipperfield et al. (2013). Alternatively, it could be approximated as the number of questions in the modules included in the instrument, if the burden created by each question is judged sufficiently similar. The unit cost α_j of a questionnaire may, in general, depend on the burden of the modules contained in the instrument. The form of this dependence may be arbitrary, suggested by the specific application. A simple linear specification can be expressed as

$$\alpha_j^{(hh)} = \beta_0^{(hh)} + \beta_1^{(hh)} \sum_{\Lambda_j^{(hh)}} l_\gamma \text{ and } \alpha_j^{(ind)} = \beta_0^{(ind)} + \beta_1^{(ind)} \sum_{\Lambda_j^{(ind)}} l_\gamma,$$

where l_γ is the respondent burden of module M_γ. The first summation extends over $\Lambda_j^{(hh)}$, the set of all modules concerning households contained in I_j, and the second extends over $\Lambda_j^{(ind)}$, the set of modules concerning individuals contained in I_j. Finally, $\beta_0^{(hh)}$ is the fixed cost per household visited (independently of the number of individuals in the household), $\beta_1^{(hh)}$ is the cost per unit of response burden for the household, $\beta_0^{(ind)}$ is the additional fixed cost per person in the sample (if any), and $\beta_1^{(ind)}$ is the cost per unit of response burden per person in the sample. Thus, we obtain that the unit cost α_j of instrument j has the form

$$\alpha_j = \bar{q}^{-1} \beta_0^{(hh)} + \bar{q}^{-1} \beta_1^{(hh)} \sum_{\Lambda_j^{(hh)}} l_\gamma + \beta_0^{(ind)} + \beta_1^{(ind)} \sum_{\Lambda_j^{(ind)}} l_\gamma.$$

A similar idea for defining the unit cost is used in Chipperfield and Steel (2009; 2011), where α_j is defined as the fixed cost per unit plus the sum of the marginal data collection cost across variables in pattern j.

3.2. *Dual Optimisation: Simulated Annealing and Simplex*

3.2.1. The Optimisation Problem

Assuming a predetermined and fixed number k of instruments, our objective function C will depend on the composition of these instruments and on the sample sizes $\mathbf{n}$. The composition of the instruments is determined by the $m' \times k$ matrix $\mathbf{A}$, where $[\mathbf{A}]_{i,j} = a_{i,j}$: the module i belongs to instrument j if and only if $a_{i,j} = 1$. Thus, we have $C = C(\mathbf{A}, \mathbf{n})$. Our aim is to find

$$\begin{aligned} &\min{}_{\mathbf{A},\mathbf{n}} C(\mathbf{A}, \mathbf{n}), \text{ under the conditions : } \mathbf{A} \text{ admissible}, (\mathbf{A}\mathbf{n})_i \geq n_i^*, \\ &i = 1, \ldots, m' \text{ and } \mathbf{B}\mathbf{n} = \mathbf{0} \end{aligned} \tag{6}$$

that is, over the dual-space of $\mathbf{A}$ and $\mathbf{n}$. Now, for any given composition of instruments, that is, for any given $\mathbf{A}$, it is easy to optimise an objective function that is linear in $\mathbf{n}$ under constraints that are also linear in $\mathbf{n}$ via the simplex algorithm. This yields an optimal sample size allocation $\mathbf{n}^{opt} = \mathbf{n}^{opt}(\mathbf{A})$. In this way, our problem can be translated into a minimisation problem in $\mathbf{A}$ alone, that is,

$$\begin{aligned} &\min{}_{\mathbf{A}} C(\mathbf{A}, \mathbf{n}^{opt}(\mathbf{A})) \text{ under the conditions : } \mathbf{A} \text{ admissible}, (\mathbf{A}\mathbf{n}^{opt}(\mathbf{A}))_i \geq n_i^*, \\ &i = 1, \ldots, m' \text{ and } \mathbf{B}\mathbf{n} = \mathbf{0} \end{aligned} \tag{7}$$

The space to which $\mathbf{A}$ may belong to is a subspace of $\{0,1\}^{km'}$, the space of all sequences of zeros (‘0’) and ones (‘1’) of length km', the number of instruments multiplied by the number of modules and crossings. The size of this space depends exponentially on km'. For this reason, the space cannot be searched exhaustively to find the minimum. Heuristic and metaheuristic methods have been proposed as a way of solving combinatorial optimisation problems, such as the one we are facing here. They yield approximate solutions while being computationally feasible (see e.g., Blum and Roli 2003). Algorithms of this type ‘search’ the state space in such a way that there is a good chance of finding an approximation to the minimum (even if the state space is not searched exhaustively). One of the most prominent such algorithms is simulated annealing. It was inspired by a physical process for growing crystals: a molten fluid is slowly cooled until crystals are formed, the slow cooling rate being crucial for crystal formation. The theoretical properties of simulated annealing have been studied extensively. In particular, the conditions under which the algorithm will converge to the optimum with probability 1 are given, for example in Mitra et al. (1986) and Hajek (1988). The main reason for favouring this algorithm is that we could come up with an idea of implementing it in such a way as to accommodate the complicated constraints of the problem (e.g., different periodicities of modules and instruments and the presence of crossings), while respecting the theoretical conditions known to guarantee convergence.

Simulated annealing is a probabilistic search method: it moves in the state space from one element to the next by applying small random changes to the current state in the search for the minimum of a certain cost function. If the new state has a lower value of the cost function, it is accepted. If it has a higher cost, it can still be accepted with some small probability. Allowing a higher cost to be accepted in this way allows the algorithm to climb out of local minima of the cost function, so that it can converge towards the global minimum. The probability of accepting a state with a higher cost than the previous one depends on the difference in cost and on a ‘cooling schedule’, given by the ‘temperature’ T_t. This temperature decreases towards zero at a specified rate. Accepting a state with a higher cost than the previous one becomes increasingly less likely as the temperature decreases. This enforces convergence to the optimum as the system ‘cools down’. Mitra et al. (1986) note that the condition $T_t \rightarrow 0$ is not sufficient for the resulting time-inhomogeneous Markov process to converge to the optimum. They prove that for discrete state spaces an assumption of T_t decreasing at a logarithmic rate, that is $T_t = \gamma / \log(t + t_0 + 1)$, with a suitable lower bound set for γ, is sufficient for this convergence. These results are in line with Hajek (1988), who gives a necessary and sufficient condition on the cooling schedules for convergence, and also suggests a logarithmic cooling schedule. For continuous state spaces, faster cooling rates may be sufficient for an appropriately chosen distribution of random changes from the current state, see, for example Tsallis and Stariollo (1996).

3.2.2. Implementing Simulated Annealing

Simulated annealing can be described as follows: starting with some arbitrary (but admissible) matrix $\mathbf{A}_0$, the following steps are repeated until no further cost reductions are achieved. Setting $C(\mathbf{A}) = C(\mathbf{A},\mathbf{n}^{opt}(\mathbf{A}))$:

i. *Proposal of a new state*. In each step t from the current $\mathbf{A}_t$, a new composition $\mathbf{A}_{t+1}$ is proposed by randomly perturbing $\mathbf{A}_t$. If $\mathbf{A}_{t+1}$ is not admissible, it is rejected by setting $\mathbf{A}_{t+1} = \mathbf{A}_t$

ii. *Sample size optimisation*. For the new candidate state $\mathbf{A}_{t+1}$, the simplex algorithm determines $\mathbf{n}_{t+1}^{opt} = \mathbf{n}^{opt}(\mathbf{A}_{t+1})$, which allows calculation of $C(\mathbf{A}_{t+1}) = C(\mathbf{A}_{t+1}, \mathbf{n}^{opt}(\mathbf{A}_{t+1}))$.

iii. *Acceptance of the proposal*. If the proposal does not increase the cost, that is $C(\mathbf{A}_{t+1}) \le C(\mathbf{A}_t)$, it is accepted. If the cost increases, that is $C(\mathbf{A}_{t+1}) \ge C(\mathbf{A}_t)$, then the step may still be accepted, with the probability of acceptance being given by $\exp(-[C(\mathbf{A}_{t+1}) - C(\mathbf{A}_t)]/T_t)$, where $T_t = \gamma/\log(t + t_0 + 1)$.

iv. If accepted, $\mathbf{A}_{t+1}$ becomes the new current state, and the new sample size allocation is given by $\mathbf{n}_{t+1}^{opt} = \mathbf{n}^{opt}(\mathbf{A}_{t+1})$.

A further assumption needed for convergence (with probability 1) to the optimum — in addition to the assumption of the logarithmic rate of the cooling schedule, concerns the proposal of $\mathbf{A}_{t+1}$ from the current state $\mathbf{A}_t$ (see point i above). This assumption ensures that, were the temperature frozen at some fixed value, the resulting time-homogeneous Markov process would have a limit distribution. Mitra et al. (1986) note that if, additionally, $T_t \rightarrow 0$ and $T_{t+1} < T_t$ are assumed, the limit distribution would be concentrated on the optimum. This would, however, require that at each fixed value of the temperature T_{t} a number of steps be undertaken large enough that this limit distribution is reached (see e.g., Angelis et al. 2001).

'Proposal-symmetry' condition: the conditional probability of composition $\mathbf{A}'$ *being proposed if the current state is* $\mathbf{A}$, $P(\mathbf{A}'|\mathbf{A})$), *equals the probability of composition* $\mathbf{A}$ *being proposed if the current state is* $\mathbf{A}'$, $P(\mathbf{A}|\mathbf{A}')$). Thus $P(\mathbf{A}'|\mathbf{A}) = P(\mathbf{A}|\mathbf{A}')$. *(See e.g., Mitra et al. 1986)*

We describe below a random perturbation mechanism that respects this symmetry condition and the constraints from Subsections 2.1–2.3.

We use the notation introduced in Section 2: single modules correspond to rows $\{1, \ldots, m\}$ of the matrix $\mathbf{A}$, and crossings correspond to rows $m+1, \ldots, m'$. We also make use of the concept of a 'complete periodic subset of instruments'. We denote as $J_{j_0}(i_0)$ a subset of a parent instrument that is complete for the module in i_0, and to which the instrument in j_0 belongs (see Subsection 2.3). (For example, if the module in row i_0 is of quarterly frequency and the instrument in column j_0 belongs to a monthly parent instrument, and if j_0 corresponds, say, to the February sibling, then $J_{j_0}(i_0)$ consists of the column indices of the siblings of the same parent instrument corresponding to February, May, August and November.) Finally, we say that the module in row i_0 is 'switched on' in the instrument in column j if $a_{i_0,j} = 1$, while if $a_{i_0,j} = 0$ we say that it is 'switched off'. We also say that we 'reverse a module's switch position' if we set $a_{i_0,j}$ to $1 - a_{i_0,j}$.

For admissibility we must respect the following properties:

a. If a mandatory crossing is switched on for some instrument, then it must also be switched on for all modules that are members of the crossing for the same instrument.

b. For each row of the matrix $\mathbf{A}$ corresponding to a module (or a crossing) of a certain periodicity, there is at least one complete periodic set of instruments for this module (or crossing) that is switched on. This is necessary in order to be able to allocate some sample size to this module.

c. Possible further constraints, such as those discussed in Subsection 2.4, are satisfied.

From a practical point of view, the value of the parameter γ of the cooling schedule should be chosen in such a way as to strike a good balance between an in-depth local search, that is exploring the area in the immediate vicinity of the current state space, and a global search, that is exploring areas of the state space that have not yet been sufficiently searched. For example, a value of $\gamma = 0$ would enforce quick convergence to the next local minimum, but would leave other areas of the state space unexplored. A quick apparent convergence is therefore not necessarily desirable. According to the exposition in Cicirello (2007), Lam and Delosme (1988) proposed a so called *"D-equilibrium"* as a trade-off between speed of convergence to an optimum and quality of this optimum in simulated annealing. Their analysis suggested that this is achieved when the acceptance rate, averaged across the previous 500 steps of the algorithm, is kept around 44 %. One might set the parameter γ in our algorithm so as to achieve this target. Reference is made to other approaches to nonmonotonic cooling schedules (alternating cooling and reheating) in Blum and Roli (2003).

A possible stopping rule for the algorithm would be one of the type: 'exit when the reduction in cost achieved during the last t_{crit} steps did not exceed eps_{crit}', for values of t_{crit} and eps_{crit} chosen by the user. The choice of the initial composition $\mathbf{A}_0$ may affect the number of steps required to achieve the optimum. A possible approach is to start with a composition in which each module participates in all siblings of all instruments that have a period required by the module, provided the constraints of Subsubsection 2.4.1 are also respected. If an admissible initial composition is found, this guarantees the existence of a solution to the optimisation problem: since the set of admissible compositions is nonempty and finite, the minimum will be attained. In more sizeable problems, that is those with a large number of instruments, modules and crossings, an alternative possible strategy would be to first find the optimum for a reduced problem, and then to use this to obtain a starting value for the full problem.

3.2.3. The Proposal of a New State: Simple Cases

The first step (i) in the implementation of simulated annealing, as set out in the previous section — the proposal of $\mathbf{A}_{t+1}$ from $\mathbf{A}_t$ at step t — can be defined as consisting of three phases:

1) Choose a row (module or crossing) i_0 at random from the uniform distribution on $\{1, \ldots, m'\}$ and a column (instrument) j_0 from the uniform distribution on all instruments that are of an appropriate periodicity for this module (i.e., have the same period or a period that is an integer fraction thereof).
2) Modify a_{i_0, j_0} and other related elements of $\mathbf{A}_t$ (see below for details).
3) If $\mathbf{A}_{t+1}$ is not admissible, discard it and set $\mathbf{A}_{t+1} = \mathbf{A}_t$.

Phase (2), the modification of a_{i_0, j_0} and other related elements, must be done in such a way that the 'proposal-symmetry' condition is satisfied.

To illustrate the idea, let us first consider the simple case where there is a unique frequency and there are no crossings. In this case, there are no further 'related elements' to be taken into account in Phase (2). It is only a_{i_0, j_0} that is modified by having its switch position reversed, that is by being set to $1 - a_{i_0,j_0}$. Then, starting at some state $\mathbf{A}$, the probability of proposing a neighbouring state (defined as those to which the probability of moving from $\mathbf{A}$ is positive), say $\mathbf{A}'$, is equal to $1/(mk)$. The same holds for the probability of proposing $\mathbf{A}$ when starting at $\mathbf{A}'$. Thus, the condition is fulfilled.

Let us now consider a slightly more complicated case, where there is one crossing. The constraint created by the presence of the crossing is that when the crossing is switched on, the modules belonging to that crossing must also all be switched on. This is not true the other way around: modules have different constraints on their sample sizes and may, moreover, be members of many crossings. They should therefore be given the freedom to participate in an instrument, even if a crossing to which they belong is not in this instrument. Crossings require the joint presence of modules, but do not set any constraints on their joint absence.

Assuming there is a crossing, and that the row selected in Phase (1) corresponds to a module, that is if $i_0 \leq m$, Phase (2) may be defined in exactly the same way as in the case considered above; the argument on symmetry still applies, as may be easily verified. We now consider the case where in Phase (1) a crossing, rather than an individual module, is selected (i.e., if $i_0 > m$). To illustrate this situation, let us assume that M_1 in Table 1 is also of annual periodicity and that $i_0 = 5$, which means G_1 was selected. Let us also assume that an instrument is selected that is switched off, that is $a_{5,j_0} = 0$, for example $j_0 = 3$. The idea is to decide in a randomised way whether the crossing will be switched on: we will switch it on (by setting $a_{5,3} = 1$) with some probability p_0, or leave everything unchanged (the reason for doing so will become apparent at the end of this paragraph). If we switch it on, then this obviously implies that the members of G_1 should be switched on as well, that is we would also set $a_{2,3} = a_{4.3} = 1$. Ignoring rows and columns that are not involved, we denote this state by $\mathbf{A}'$. Note, however, that, depending on which of the members ($i = 2$ and $i = 4$) are already switched on at the current step t, there are different states $\mathbf{A}$ that may lead to the same state $\mathbf{A}'$. In fact, if the crossing has L members, then there are 2^L different states $\mathbf{A}$ that could all lead to $\mathbf{A}'$. In order to satisfy the symmetry condition, we need to be able to get back to any state $\mathbf{A}$ that may have yielded $\mathbf{A}'$. Assuming that each one of these states is to be proposed from $\mathbf{A}'$ with equal probability, this probability should be $1/2^L$. This can be achieved by switching off each member of the crossing independently with probability $1/2$. This implies that, in order to satisfy the symmetry condition, we should also choose $p_0 = 1/2^L$ in our initial randomised decision on whether to have the crossing switched on or off.

Greater caution is required if there are two or more crossings: if at the step at which, say, the first crossing will be switched off (in a column j_0), some of its members are also members of some other crossing that is currently switched on in j_0, then these modules cannot be switched off, as this would lead to an inadmissible composition. Thus, they should not be counted in the different states $\mathbf{A}$, from which $\mathbf{A}'$ may be obtained. Such modules should therefore not be counted in L (which will thus be a random variable).

If there are a number of different frequencies but no crossings, the situation is simple again: for each selected module i_0 and instrument j_0, reverse the switch position of the module in all instruments in a complete periodic set for module i_0, the one to which j_0 belongs, that is, change the setting $a_{i_0,j}$ to $1 - a_{i_0,j}$ for all $j \in J_{j_0}(i_0)$.

3.2.4. The Proposal of a New State: The General Case

In the general case, where there are both different frequencies and crossings, the approach follows the same principle but is significantly more complicated, due to the fact that a crossing and its members may have different frequencies. Only Phase (2) from the previous subsection needs to be modified to accommodate this situation.

Phase (2) of the first step (i) of the implementation of simulated annealing can be defined by distinguishing between two separate cases:

I) If a single module was chosen (i.e., if $i_0 \leq m$), then all instruments in the complete subset to which the module belongs should have their switch position reversed, that is, if $a_{i_0,j_0} = 0$ then set $a_{i_0,j} = 1$ for all $j \in J_{j_0}(i_0)$. Similarly, if $a_{i_0,j_0} = 1$ then set $a_{i_0,j} = 0$ for all $j \in J_{j_0}(i_0)$.

II) If a crossing of period p_{i_0} was chosen (i.e., if $i_0 > m$), then for all its members M_i that have a period p_i, set $r_{i|i_0} = p_i / p_{i_0}$. Then for each M_i there is a partition of $J_{j_0}(i_0)$ into $r_{i|i_0}$ subsets, each of which is complete for M_i. Let us call these subsets $J_{j_0}(i,1), \ldots, J_{j_0}(i, r_{i|i_0})$. The number of these subsets equals $r_{i|i_0}$. Let $\rho_{i|i_0}$ be the number of subsets excluding those subsets $J_{j_0}(i,k)$ for which M_i is a member of some further crossing $G_{i_1}(i_1 \neq i_0)$ that is currently switched on, that is, $a_{i_1,j} = 1$ for $j \in J_{j_0}(i, k)$. Thus

$$\rho_{i|i_0} = \#\{J_{j_0}(i,k), k = 1, \ldots, r_{i|i_0} | a_{i_1,j} = 0, \forall i_1 > m, i_1 \neq i_0 : M_i \in G_{i_1}, j \in J_{j_0}(i,k)\},$$

where # denotes cardinality. Note that if the module in row i is a member of another crossing that is currently switched on for all $j \in J_{j_0}(i_0)$, then $\rho_{i|i_0} = 0$.

i. If $a_{i_0,j_0} = 0$, then let $L = \sum \rho_{i|i_0}$, where the summation extends over all members of the crossing. Then, with probability $(1/2)^L$, for all $j \in J_{j_0}(i_0)$, set $a_{i_0,j} = 1$ and $a_{i,j} = 1$ for all rows i containing modules that are members of the crossing; else (with probability $1 - (1/2)^L$), set $\mathbf{A}_{t+1} = \mathbf{A}_t$.

ii. If $a_{i_0,j_0} = 1$, then set $a_{i_0,j} = 0$, for all $j \in J_{j_0}(i_0)$. Also, independently for all rows i containing modules that are members of the crossing considered (in row i_0), and independently for each of the subsets $J_{j_0}(i,k)$ of the partition $J_{j_0}(i, 1), \ldots, J_{j_0}(i, r_{i|i_0})$ of $J_{j_0}(i_0)$, for which the module in row i is not a member of another crossing (beyond the one in i_0), which (other crossing) is currently "switched on", with probability equal to $1/2$ set $a_{i,j} = 0$ for all $j \in J_{j_0}(i,k)$.

Note that the proposal mechanism is such that the constraints set out in Subsections 2.2 and 2.3 are already incorporated in Phase (2), while the constraints from Subsection 2.4 are only checked *a posteriori* in Phase (3), once a new proposal has been made.

Different ways of updating the composition matrix **A** can of course affect the numerical efficiency of the search algorithm. We do not have an optimal search algorithm at our disposal.

On the basis of the above, it is possible to prove the following special case. We are not currently able to provide a proof for cases where other constraints, such as dependencies between modules, are taken into account.

Proposition. *The proposal mechanism described in Steps 1–3 above respects the proposal-symmetry condition, subject to the constraints described in Subsections 2.1–2.3.* The proof of this proposition may be found in the Appendix.

3.3. Appropriateness of the Simulated Annealing Algorithm

The algorithm, consisting of simulated annealing and simplex, can also be used in the design of an efficient SQD, thus resolving a problem to which feasible solutions exist in the literature only for the case of a very limited number of variables (or modules).

For the single survey SQD, and under simple random sampling, Chipperfield and Steel (2009; 2011) consider the optimal allocation of sample size to minimise cost, subject to precision constraints. The most efficient design is determined on the basis of best linear unbiased estimation (possible under simple random sampling and given correlations between all variables), and therefore comprises all $2^m - 1$ possible patterns (instruments). For the case being studied here — an integrated survey with a complex and modular sampling design and a baseline estimation approach involving HT estimators for single instruments (samples), and composite HT estimators combining data from different instruments — we are also aiming to minimise cost, under similar precision constraints. However, our minimisation is over all possible $c_k = \binom{2^m - 1}{k}$ combinations of k instruments, and over the associated sample size allocations. Noting that the use of $2^m - 1$ samples is not practical even for moderate m, Chipperfield and Steel (2009) also propose the choice of a limited number of k best patterns according to a ranking of all patterns based on their relative estimation efficiency. They thus circumvent the more difficult question of optimisation over all c_k combinations, but at the expense of a loss of efficiency.

Our specification of a fixed number of instruments, and the optimisation over all c_k combinations, coincides with the approach adopted by Adiguzel and Wedel (2008). They, however, consider the distinct problem of finding the single survey SQD that minimises the Kullback-Leibler distance (see Kullback and Leibler 1951) to the full questionnaire, while assuming equal sample sizes for all k instruments. Without the assumption of equal sample sizes, the minimum is attained by the full questionnaire. Thus, their approach does not include optimal sample size allocation. A random search algorithm, a modification of the Fedorov algorithm (see e.g., Cook and Nachtsheim 1980), is then used to find the optimal instrument composition.

Applying this algorithm to our problem, including the simultaneous optimisation over sample sizes, would mean that a single full step of the algorithm should examine each of the $(2^m - 1)$ possible compositions for each of the k instruments. It would thus require the same computation time as $k(2^m - 1)$ steps of the annealing algorithm. For example, with 30 modules and three instruments, this figure is around three billion. The number of modules in an integrated social survey system could range from 50 to 200, making it impractical to apply this algorithm.

There are alternative ways of implementing simulated annealing. In particular, future research may draw on the extensive literature on the application of simulated annealing in the optimisation of experimental designs (see Meyer and Nachtsheim 1988). Two examples of work in this area are the discussions on sequential and nonsequential exchange methods in Lejeune (2003) and on exchange and interchange steps in Jansen et al. (1992), which may lead to improvements to the current proposal mechanism. Cooling rates declining faster to zero, as discussed in Tsallis and Stariollo (1996), and the analysis of the impact of the values of the parameters involved (see e.g., Angelis et al. 2001), are also interesting topics for further research.

Finally, the use of other heuristic and metaheuristic algorithms, such as genetic algorithms and other population-based methods (see e.g., Blum and Roli 2003 for a brief discussion of and further references to work on these types of algorithms) could be explored.

4. Modular Design with Complex Sampling

In Section 2, we assumed the use of simple random sampling. In this section, we extend the approach to include complex sampling designs. We first set out how the precision requirements for estimating θ are translated into constraints on the sample sizes, and then explain how the optimisation framework can be modified accordingly.

4.1. *Adjusting Sample Size Requirements*

4.1.1. Complex Sampling Design for Independent Samples

Estimating θ from a single sample under some complex design, which, for example, involves stratification and/or multistage sampling, leads to a different variance $V'(\hat{\theta})$. Assuming normality for $\hat{\theta}$, the precision constraint $V'(\hat{\theta}) \leq \left(e/z_{1-\alpha}\right)^2$ is then satisfied by the constraint $n/d \geq n^*$, where d is the customary design effect (defined as $d = V'(\hat{\theta})/V(\hat{\theta})$, with $V(\hat{\theta})$ being the variance of the same estimator under simple random sampling), and n^* is as in (1).

When θ is estimated by a weighted average of estimates from different instruments, as in Subsection 2.1, then the weights should be modified to $w_i = \left(n_i/d_i\right)/\sum\left(n_j/d_j\right)$, where d_i is the design effect for the ith sample. It can be shown that for independent simple random samples and for this choice of weights, the composite estimator $\hat{\theta} = \sum w_i \hat{\theta}_i$ has minimum variance $V'(\hat{\theta}) = \sigma_\theta^2/\sum\left(n_i/d_i\right)$, provided that sampling fractions n_i/N are negligible. The precision requirement $V'(\hat{\theta}) \leq \left(e/z_{1-\alpha}\right)^2$ is then satisfied if

$$\sum\left(n_j/d_j\right) \geq n^*.$$

In the case where the design effects for the different samples are identical, the weights of the composite estimator are $w_i = n_i/\sum n_j$, and the precision requirement is satisfied by $\left(\sum n_i\right)/d \geq n^*$, where d is the common design effect.

4.1.2. Complex Sampling Design for Coordinated Samples

If the samples of the various instruments are negatively coordinated, for example by splitting a sample into subsamples, one for each instrument, then $V'(\hat{\theta}) \leq \sigma_\theta^2/\sum\left(n_i/d_i\right)$, due to the negative covariance term between the estimates $\hat{\theta}_i$. The precision requirement is, therefore, again satisfied if

$$\sum \left(n_j/d_j\right) \geq n^*.$$

Let us finally consider an example where there is positive coordination between the samples and show how this case may also be embedded in the current framework. We assume that there is some overlap between L samples $S_1, \ldots, S_L$, of sizes $n_1, \ldots, n_L$, as is the case in surveys with rotating samples. Let us denote by $n_{i\cap j}$ the number of sample units in $S_i \cap S_j$, let $n_{i\cap j} = \tau_{ij} n_i$ for some appropriate τ_{ij}, and assume that all subsets reflecting common and noncommon parts of these samples are drawn independently from each other by simple random sampling. It is then logical to pool estimates over time, as is done for example when estimating a yearly average from four quarterly estimates. Putting $\hat{\theta} = \sum w_i \hat{\theta}_i$, where $w_i = n_i/n_+$, $n_+ = \sum n_i$, $\tau_{i+} = \sum_j \tau_{ij}$ and $\tau_{avg} = \sum_i \tau_{i+} w_i$, we obtain

$$V(\hat{\theta}) \leq \frac{\sigma_\theta^2}{n_+}\left(n_+^{-1}\sum_{i,j} n_{i\cap j}\right) = \frac{\sigma_\theta^2}{n_+}\left(\sum_i \frac{n_i}{n_+}\tau_{i+}\right) = \frac{\sigma_\theta^2}{n_+}\tau_{avg}.$$

In the specific case of a longitudinal survey with L wave, for which the sample sizes $n_1, \ldots, n_L$ are restricted to be equal to each other, we have $\tau_{avg} = L^{-1}\sum \tau_{i+}$. It is then possible to adopt a 'design' effect $d_j = \tau_{avg}$, such that the precision requirement is satisfied if $\sum \left(n_j/d_j\right) \geq n^*$.

4.2. Adapting the Optimisation Framework

In all the cases described in Subsection 4.1, the precision requirements for a module M_a are satisfied by

$$\sum_{\Theta_a} \left(n_j/d_{a,j}\right) \geq n_a^*, \tag{8}$$

where $d_{a,j}$ is the module's design effect for instrument j, n_a^* is obtained as in (3), and the summation extends over Θ_a, the set of all instruments containing the module M_a.

A way of representing these m' linear constraints (one for each module and each crossing), on the sample sizes $n_1, \ldots, n_k$ of the various samples, is by introducing the $m' \times k$ 'design-composition matrix' $\mathbf{R}$, with elements $[\mathbf{R}]_{i,j} = a_{i,j} d_{i,j}^{-1}$, where $a_{i,j}$ are the elements of the composition matrix $\mathbf{A}$ and $d_{i,j}$ those of the design-effect matrix $[\mathbf{D}]_{i,j}$. For a crossing in row i_0, we set $d_{i_0 j_0} = \max_i\{d_{ij_0}\}$, where the maximum is over all members of the crossing. In the special case where all samples are independent and drawn using simple random sampling, and a weighted average of HT estimators is used, all $d_{i,j} = 1$, and we obtain $\mathbf{R} = \mathbf{A}$.

Generally, the constraints in (8) may now be expressed as

$$\mathbf{Rn} \geq \mathbf{n}^*,\ \mathbf{R}\text{ is an } m' \times k \text{ matrix}, \tag{9}$$

where the inequality is to be understood componentwise, $\mathbf{n} = (n_1, \ldots, n_k)'$ is the vector of the actual sizes of the samples associated with the k instruments, and $\mathbf{n}^* = (n_1^*, \ldots, n_{m'}^*)'$ is the vector of the minimum sample sizes for the m' modules and crossings required under simple random sampling, that is, those given in (1). This relation now replaces (5).

Furthermore, (7) is now replaced by

$$\min{}_{\mathbf{A}} C(\mathbf{A}, \mathbf{n}^{opt}(\mathbf{A}, \mathbf{D})) \text{ under the conditions : } \mathbf{A} \text{ admissible, } (\mathbf{Rn}^{opt}(\mathbf{A}, \mathbf{D}))_i \geq n_i^*, i = 1, \ldots, m' \text{ and } \mathbf{Bn} = \mathbf{0} \tag{10}$$

where $\mathbf{n}^{opt}(\mathbf{A}, \mathbf{D})$ is the admissible vector of optimal instrument sample sizes and $\mathbf{n}^{opt}(\mathbf{A}, \mathbf{D}) \neq \mathbf{n}^{opt}(\mathbf{A})$.

In other words, the algorithm is defined exactly as previously on the space of all composition matrices **A**, with the only modification being that the sample size constraints are now formulated using the design-composition matrix **R** instead of the composition matrix **A**.

5. An Illustrative Example

As an illustrative example of how the approach presented above could be used, we consider a hypothetical reorganisation of the Labour Force Survey (LFS) using three quarterly instruments, in place of the current single questionnaire, which is administered quarterly.

The blocking of the LFS variables into 30 modules was prepared by Eurostat (see Table 2). The first six of these modules represent demographic and household characteristics, and will be present in all available instruments. A further group of ten modules contains the structural variables, that is, those on which information must be collected on an annual basis (in accordance with EU regulations). Of the modules containing quarterly variables, five are needed for the definition of "ILO-Unemployment" (unemployment as defined by the International Labour Organization). We therefore group these into a single crossing. For the purposes of this illustration, we split the remaining nine quarterly modules into two crossings: 'Employment conditions 1' ('Empl_cond1'), containing six modules, and 'Employment conditions 2' ('Empl_cond2') with three modules, as indicated in Table 2.

For each module, Eurostat defined the dependencies on other modules and calculated, based on historically observed proportions of respondents, the average respondent burden l_γ as the sum (across individuals in the sample) of the number of questions in M_γ answered by each individual, divided by the total number of individuals in the sample. For the purposes of this illustration, we assume the fix costs $C_j^{(f)}$, $\beta_0^{(hh)}$ and $\beta_0^{(ind)}$ to be 0, and the coefficients $\beta_1^{(hh)}$ and $\beta_1^{(ind)}$ to be equal to 1. The calculations are based on the data for Portugal. The reason for choosing Portugal was that it has a population size close to the EU median.

The required sample sizes are calculated so as to satisfy EU regulations that apply to Portugal, and on the assumption that all instruments are administered to independent samples, drawn using simple random sampling. According to Commission Regulation (EC) No 377/2008, for structural variables, the relative standard error (assuming simple random sampling) of any yearly estimate representing one percent or more of the working-age population must not exceed nine percent for countries with a population of between one million and 20 million inhabitants. Note that, if simple random sampling is assumed, the sample sizes required to satisfy these precision requirements may be derived from the variance of estimating proportions of a specified magnitude, and, therefore, do not need to be estimated by means of pilot samples or other similar single surveys. This implies a

Table 2. *Optimal instrument composition, burden and sample sizes for the illustrative example from Section 5*

Instrument													
	Name	I1.1	I1.2	I1.3	I1.4	I2.1	I2.2	I2.3	I2.4	I3.1	I3.2	I3.3	I3.4
	Final sample size	22,050	22,050	22,050	22,050	7,367	7,367	7,367	7,367	12,585	12,585	12,585	12,585
	Burden	35	35	35	35	37	37	37	37	33	33	38	41

Module							Instrument composition (modules)											
Name	Description	Depends on	Burden	Period	Sample size													
					Required	Final												
GEO	Core geographic module		1.25			168,008	1	1	1	1	1	1	1	1	1	1	1	1
DEMO	Core demographic module		4.62			168,008	1	1	1	1	1	1	1	1	1	1	1	1
DEMOHAR	Common harmonised demographic module (migration)		4.00			168,008	1	1	1	1	1	1	1	1	1	1	1	1
HHOLD	Household composition		3.75			168,008	1	1	1	1	1	1	1	1	1	1	1	1
GEN	Inter-generational information		2.00			168,008	1	1	1	1	1	1	1	1	1	1	1	1
GENHAR	Common harmonised inter-generational module		9.81			168,008	1	1	1	1	1	1	1	1	1	1	1	1
ATWORK	Employment status - At work	DOME	1.75	3	168,005	168,008	1	1	1	1	1	1	1	1	1	1	1	1
ABSWORK	Employment status - Absences from work	ATWORK, STAPRO	0.03	3	168,005	168,008	1	1	1	1	1	1	1	1	1	1	1	1
STAPRO	Contract - Professional status	AKROWT	0.74	3	168,005	168,008	1	1	1	1	1	1	1	1	1	1	1	1
WORKSEEK	Looking for work - Search for work	ABSWORK, ATWORK, DEMO	2.20	3	168,005	168,008	1	1	1	1	1	1	1	1	1	1	1	1
WORKSEEKAVAIL	Looking for work - Availability	WORKSEEK	0.25	3	168,005	168,008	1	1	1	1	1	1	1	1	1	1	1	1
OCCUP	Characteristics of the workplace - Occupation	ATWORK, STAPRO	0.90	3	117,665	117,668	1	1	1	1	1	1	1	1	0	0	0	0
WORKHOURUSUAL	Working hours - Usual hours	ATWORK	0.48	3	117,665	117,668	1	1	1	1	1	1	1	1	0	0	0	0
WORKHOURACTUAL	Working hours - Actual hours	ATWORK, STAPRO, WORKHOURUSUAL	1.39	3	117,665	117,668	1	1	1	1	1	1	1	1	0	0	0	0
WORKMORE	Wished hours	ATWORK	0.52	3	117,665	117,668	1	1	1	1	1	1	1	1	0	0	0	0
ECONACT	Characteristics of the workplace - Economic activity	ATWORK	0.67	3	117,665	117,668	1	1	1	1	1	1	1	1	0	0	0	0
SIZEFIRM	Characteristics of the workplace - Enterprise characteristics	STAPRO	0.33	3	117,665	117,668	1	1	1	1	1	1	1	1	0	0	0	0
CONTRACT	Contract - Characteristics of contract	STAPRO	0.74	3	79,805	79,808	0	0	0	0	1	1	1	1	1	1	1	1
WORKPLACE	Characteristics of the workplace - Workplace	ATWORK	1.23	3	79,805	79,808	0	0	0	0	1	1	1	1	1	1	1	1
JOBSTART	Start of job	ATWORK, STAPRO	0.12	3	79,805	79,808	0	0	0	0	1	1	1	1	1	1	1	1
MAINEMPLSTAT	Main employment status	DEMO	1.00	12	12,205	12,585	0	0	0	0	0	0	0	0	1	0	0	0
WORKTIME	Working times	ATWORK, STAPRO	1.95	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	0	1
JOBSECOND	Second job	ATWORK	0.45	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	0	1
JOBLOOKING	Looking for other job	ATWORK, DEMO	0.44	12	12,205	12,585	0	0	0	0	0	0	0	0	0	1	0	0
WORKEXP	Previous work experience	ATWORK	1.24	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	1	0
EMPLOY1Y	Employment Situation 1 year before	DEMO	3.72	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	0	1
EDUCATT_LEVEL	Educational attainment level	DEMO	1.11	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	0	1
EDUCATT_CHAR	Educational attainment main characteristics	DEMO, EDUCATT_LEVEL	1.21	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	0	1
EDUCPART4W	Participation in education and training (4 weeks)	DEMO	1.83	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	1	0
EDUCPART4W_CHAR	Characteristics of participation in education and training (4 weeks)	EDUCPART4W	2.10	12	12,205	12,585	0	0	0	0	0	0	0	0	0	0	1	0

Crossing					Instrument composition (crossings)											
Name	Members	Period	Sample size													
			Required	Final												
Demogr./Housh.	GEO,DEMO,DEMOHAR,HHOLD,GEN,GENHAR			168,008	1	1	1	1	1	1	1	1	1	1	1	1
ILO_Unempl	ATWORK,ABSWORK,STAPRO,WORKSEEK,WORKSEEKAVAIL	3	168,005	168,008	1	1	1	1	1	1	1	1	1	1	1	1
Empl_cond1	OCCUP,WORKHOURUSUAL,WORKHOURACTUAL,WORKMORE,ECONACT,SIZEFIRM	3	117,665	117,668	1	1	1	1	1	1	1	1	0	0	0	0
Empl_cond2	CONTRACT,WORKPLACE,JOBSTART	3	79,805	79,808	0	0	0	0	1	1	1	1	1	1	1	1

required total sample size of 12,205. For the variables in the ILO-Unemployment crossing, according to Council Regulation (EC) No 577/98, the relative standard error when estimating the change in subpopulations representing five percent of the working-age population between two successive quarters must not exceed three percent for countries with a population of between one million and 20 million inhabitants. This implies a required yearly sample size of 168,005 (i.e., 42,001 per quarter). This regulation dominates the others, which require smaller sample sizes. For the other two crossings containing quarterly variables, and in order to introduce some further precision differentiation, we assume sample sizes corresponding to the same level of precision in estimating the change between quarters (3 %) but for subpopulations of seven percent and ten percent of the working-age population respectively. This implies annual sample sizes of 117,665 and 79,805.

The benchmark composition used as the standard against which to compare the cost reduction achieved by the reorganisation is the composition where all modules (including structural modules) are present in all instruments. This is equivalent to the traditional approach of administering a single questionnaire to all sample units. As there is no differentiation between instruments, the total yearly sample size equals the maximum of the required sample sizes, that is, 168,005, and the total yearly data collection cost (taking into account the modules' respondent burden) is around 8.70 million person-questions. This benchmark composition was also used as the initial composition $\mathbf{A}_0$ for the annealing algorithm.

The optimisation algorithm is programmed in R (R Core Team 2013) and used R-package *lpSolve* for the simplex optimisation (Berkelaar et al. 2013). In this example, it was run for 50,000 steps, using the benchmark composition as the initial state. The cost reduction achieved by the optimisation is illustrated in Figure 1 below. It can be seen that most of the reduction has already been achieved after around 4,000 steps. The minimum is

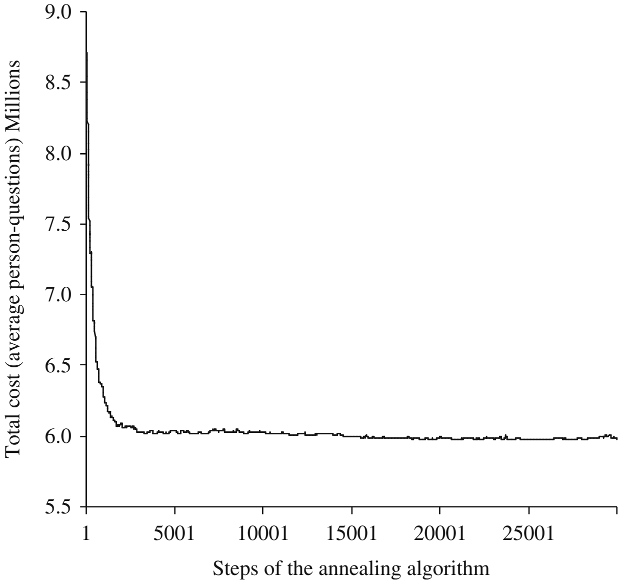


Fig. 1. Cost evolution (first 30,000 out of 50,000 steps) for compositions visited by the simulated annealing algorithm when searching for the optimal composition for the illustrative example described in Section 5.

reached after approximately 28,000 steps, with a cost of 5.97 million person-questions, a reduction of around 31 % from the cost of using the benchmark composition. Had we included a component representing the fixed cost for each instrument, increasing the number of instruments would, of course, have cancelled out some of this cost reduction. The purpose of this article is, however, to illustrate the modular approach rather than to propose a definitive alternative design.

At this ‘approximate optimum’, the ‘ILO-Unemployment’ modules are included in all three instruments, and allocated a total sample size (168,008) that is almost identical to the required one. The ‘Empl_cond1’ crossing participates in the first two instruments, and the ‘Empl_cond2’ crossing in the last two. For both crossings, the total sample sizes achieved in the optimum (117,668 and 79,808, respectively) are almost identical to the sample sizes required for these modules. The annual modules are distributed across the quarters of the instrument which has the sample size (12,585 per quarter) closest to what is required for these modules (12,205), in such a way as to respect module dependencies. These modules are therefore slightly ‘oversampled’. This oversampling would, of course, have been avoided, resulting in a further cost reduction, had one more annual instrument been introduced. Note that with the distribution of modules to instruments described in Table 2, the annual modules and the modules in Empl_cond1 are not jointly administered to any sample units, implying that there is no information on the interaction between them. This could have been avoided by introducing a further crossing comprising the modules in Empl_cond1 and some or all of the annual modules, with a required sample size chosen accordingly. Satisfying this additional constraint would inevitably have meant giving up some of the cost reduction achieved by the design shown in Table 2.

6. Conclusion

In this article, we have formulated a modular design and have proposed an algorithm that could be used to restructure the questionnaires currently used for social surveys into instruments. A better balance could thus be achieved between precision requirements on the one hand, and the need to reduce the cost and burden of the survey on the other. In view of the extremely large number of possible instrument compositions, a random search algorithm has been developed to help find the optimal composition.

There are several methodological issues that should be acknowledged in relation to the modular design. It should be noted that these are all perennial difficulties associated with survey sampling, which apply to any design approach, and a more detailed investigation of these issues is beyond the scope of this study. Nevertheless, it is helpful to be clear about the tacit assumptions that we necessarily have to make as a result, for which we have generally adopted the standard practice.

When determining required sample sizes, the requisite design parameters σ_θ^2 for all θ (means and proportions) and all modules involved and the required design effects are likely to be unknown, and would need to be estimated. This is typically done by means of pilot samples or similar single surveys. For the estimation of design effects under complex survey designs, see, in particular, Gambino (2009) and Park and Lee (2004).

Practical discussions of the modular approach often note that certain features of the modular design may have an effect on various nonsampling errors. The typical errors include those related to the cognitive aspects of the instrument composition, the mode effects and the nonresponse. The cognitive effects relate to the respondents as well as to the data collection personnel. The possible consequences include confusion, increased training costs, measurement errors, and potential nonresponse. Mode effects exist if the obtained measurements differ according to the mode of data collection, with possible modes being, for example, face-to-face interview, computer-assisted telephone interview, and self-administered postal questionnaire. The nonresponse rate, both for unit and partial nonresponse, may vary depending on the items or the combination of the modules.

We do not explicitly account for nonsampling errors while addressing the two design questions (1) and (2) that we posed in Section 1, due to the lack of carefully studied empirical evidence. For example, suppose the nonresponse rate in a forthcoming survey is expected to be around 50% based on past experience. What is then the difference in cost and precision for example between (a) planning a sample that is twice as large as the required net sample, and (b) planning a sample that is three times as large and aiming at a 33% response rate in fieldwork? Meanwhile, we do implicitly assume that the modular design is based on the best current knowledge as to how to structure the items in the different modules. We assume that the constraints on instrument composition reflect choices made as to whether some modules should or should not be present simultaneously in the same instrument, or whether a module should only be assigned to a certain mode of collection. As was demonstrated in Section 2, the modular design approach can incorporate all such constraints.

We would like to emphasise that, in practice, a potentially large error in approximating the optimal design is unlikely to be the most critical concern, not least because the optimum is only such with respect to a certain design parameterisation, which can itself be challenged and certainly will be revised over time. Moreover, attention should be given to addressing and reducing the potential nonsampling errors that may come with the new system in future. In this respect, it is important to use a suitable benchmark in order to ensure a fair assessment under which potential shortcomings are balanced against gains. It is unrealistic to require the new system to meet certain unattainable ideals which today's standalone surveys also fail to achieve.

Moreover, a number of considerable changes would have to take place at national statistical institutes for it to become possible to implement the modular approach. These involve considerable investment in resources, and the complexity and cost of these organisational changes should not be underestimated. The changes fall into three main areas: (i) Information structure and production systems: data collection and processing require a greater degree of conceptual harmonisation and operational standardisation in order to make them comparable, shareable and reusable. Standardised database and warehouse solutions and metadata systems will need to replace end-to-end statistics-specific processes and management, in order to make the data and metadata accessible across the different statistical domains. (ii) Fieldwork management systems: these are needed both as a way of managing the numerous modules and instruments required in fieldwork and so as to be able to update or replace existing modules and instruments over time. (iii) Staff: staff need to adapt to a new situation where the data are collected in multiple instruments and stored, managed and shared using standard system solutions.

Finally, we would like to draw attention to the fact that while the approach presented here could allow efficiency gains to be made compared to the system currently in use (as illustrated in Section 5), and while it is reasonable to explore this (since today's system of social surveys would be the point of departure from which a gradual phasing in of a modular system would take place), a main strength of the modular design lies in its flexibility to meet new information requirements. The system could easily incorporate new modules (or eliminate redundant ones), scale up (or reduce) the precision requirements for individual modules or cross certain modules with each other. Thus, the potential of the new system is not limited to meeting today's information needs. One of its main attractions is that it can accommodate new information requests from policymakers in a much more flexible way, without the constraints of the present survey system.

Appendix

Proof of the Proposition on the Fulfilment of the Proposal-Symmetry Condition

Let us denote by $\pi(i_0,j_0)$ the probability of selecting an i_0 and a j_0 in one of the available complete sets of instruments. In the case that $\mathbf{A}'$ was generated from $\mathbf{A}$ in the case described in Subsubsection 3.2.4, Point I) above (i.e., the switch position of a module in row $i_0 \leq m$ was reversed), we have $P(\mathbf{A}'|\mathbf{A}) = P(\mathbf{A}|\mathbf{A}') = \pi(i_0,j_0)$, and the proposal mechanism respects the symmetry condition.

Similarly, let us now consider the case where $\mathbf{A}'$ was generated from $\mathbf{A}$ in the case described in Subsubsection 3.2.4 Point II) (i.e., a crossing in row $i_0 > m$ was selected). First, assume that all the modules contained in the crossing have the same periodicity, implying $r_{i|i_0} = 1$ for all its members. Moreover, for illustration purposes, assume the crossing has only $L = 2$ members, which are assumed not to be members of any other crossings that are currently switched on. Then, in the state, say $\mathbf{A}'$, in which the crossing is switched on, all members are also switched on, while, if the crossing is switched off, there are four possible states in relation to its two members, say $\mathbf{A}_1, \ldots, \mathbf{A}_4$, corresponding to the situations where none, both, or one of the two members are switched on (for all $j \in J_{j_0}(i_0)$). Then, in the case presented in Subsubsection 3.2.4 Point II.i), we get $P(\mathbf{A}'|\mathbf{A}_i) = \left(1/2^2\right)\pi(i_0,j_0)$, while in the case presented in point II.ii), we get $\mathrm{P}(\mathbf{A}_i|A') = \left(1/2\right)\left(1/2\right)\pi(i_0,j_0)$, and the proposal mechanism respects the symmetry condition. The argument is similar for $L > 2$: the number of states from which it is possible to generate $\mathbf{A}'$ equals 2^L, and the transition probabilities to and from $\mathbf{A}'$ are all equal to $1/2^L$. Note that modules that are currently members of other crossings that are switched on are fixed in their current switch position during the transitions considered.

In the case where a crossing may contain modules of different periodicities, again, the number of states from which it is possible to generate $\mathbf{A}'$ equals 2^L, and the transition probabilities to and from S' are all equal to $1/2^L$, which concludes the proof.

7. References

Adiguzel, F. and M. Wedel. 2008. "Split Questionnaire Design for Massive Surveys." *Journal of Marketing Research* 45: 608–617. Doi: http://dx.doi.org/10.1509/jmkr.45.5.608.

Angelis, L., E. Bora-Senta, and C. Moyssiadis. 2001. "Optimal Exact Experimental Designs with Correlated Errors Through a Simulated Annealing Algorithm." *Computational Statistics & Data Analysis* 37: 275–296. Doi: http://dx.doi.org/doi:10.1016/S0167-9473(01)00011-1.

Australian Bureau of Statistics. 2012. *Household Expenditure Survey and Survey of Income and Housing, User Guide*. Australia, 2009-10 (cat. no. 6503.0). Available at: http://www.abs.gov.au/ausstats/abs@.nsf/mf/6503.0 (accessed April 2016).

Berkelaar, M., et al. 2013. *lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs. R package version 5.6.10*. Available at: http://CRAN.R-project.org/package=lpSolve (accessed August 2013).

Blum, C. and A. Roli. 2003. "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison." *ACM Computing Surveys* 35: 268–308. Doi:http://dx.doi.org/10.1145/937503.937505.

Chipperfield, J.O. and D.G. Steel. 2009. "Design and Estimation for Split Questionnaire Designs." *Journal of Official Statistics* 25: 227–244.

Chipperfield, J.O. and D.G. Steel. 2011. "Efficiency of Split Questionnaire Surveys." *Journal of Statistical Planning and Inference* 141: 1925–1925. Doi: http://dx.doi.org/10.1016/j.jspi.2010.12.003.

Chipperfield, J.O., M. Barr, and D.G. Steel. 2013. "Split Questionnaire Designs: Are They an Efficient Design Choice?" In Proceedings of the 59th ISI World Statistics Congress, 25–30 August 2013, Hong Kong. 311–316. Available at: http://2013.isiproceedings.org/Files/IPS033-P1-S.pdf (accessed June 2015).

Cicirello, V.A. 2007. "On the Design of an Adaptive Simulated Annealing Algorithm." In First Workshop on Autonomous Search, in conjunction with CP'2007, Providence, Rhode Island, USA.

Cochran, W.G. 1977. *Sampling Techniques*. New York: Wiley.

Cook, R.D. and C.J. Nachtsheim. 1980. "A Comparison of Algorithms for Constructing Exact D-optimal Designs." *Technometrics* 22: 315–324. Doi: http://dx.doi.org/10.1080/00401706.1980.10486162.

Cuppen, M.D.J., P. van der Laan, and W. van Nunspeet. 2013. "Reengineering Dutch Social Surveys: From Single-Purpose Surveys to an Integrated Design." *Statistical Journal of the International Association for Official Statistics* 29: 21–29. Doi: http://dx.doi.org/10.3233/SJI-130762.

European Commission. 2009. *Communication on the Production Method of EU Statistics: A Vision for the Next Decade*. Brussels: European Commission (COM(2009)/404). Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF (accessed January 2015).

European Statistical System Committee (ESSC). 2011. *New Conceptual Design for Household and Social Statistics*. Wiesbaden, Germany. Available at: https://www.destatis.de/EN/AboutUs/Events/DGINS/Document_Memorandum.pdf (accessed January 2015).

Gambino, J.G. 2009. "Design Effects Caveats." *The American Statistician* 63: 141–146.

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

Hajek, B. 1988. "Cooling Schedules for Optimal Annealing." *Mathematics of Operations Research* 13: 311–329. Doi: http://dx.doi.org/10.1287/moor.13.2.311.

Jansen, J., R.C.M.H. Douven, and E.E.M. van Berkum. 1992. "An Annealing Algorithm for Searching Optimal Block Designs." *Biometrical Journal* 34: 529–538. Doi: http://dx.doi.org/10.1002/bimj.4710340503.

Karlberg, M., F. Reis, C. Calizzani, and F. Gras. 2015. "A Toolbox for a Modular Design and Pooled Analysis of Sample Survey Programmes." *Statistical Journal of the International Association for Official Statistics* 31: 447–462. Doi: http://dx.doi.org/10.3233/SJI-150913.

Kirkpatrick, S., C.D. Gelatt, Jr., and M.P. Vecchi. 1983. "Optimization by Simulated Annealing." *Science* 220: 671–680.

Kullback, S. and L.A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22: 79–86. Available at: http://www.jstor.org/stable/2236703.

Lam, J. and J. Delosme. 1988. "Performance of a New Annealing Schedule." In Proceedings of the 25th ACM/IEEE Design Automation Conference. 12–15 June 1988, Anaheim, CA, U.S.A. 306–311.

Lejeune, M.A. 2003. "Heuristic Optimization of Experimental Designs." *European Journal of Operational Research* 147: 484–498. Doi: http://dx.doi.org/10.1016/S0377-2217(02)00292-8.

Merkouris, T. 2004. "Combining Independent Regression Estimators from Multiple Surveys." *Journal of the American Statistical Association* 99: 1131–1139. Doi: http://dx.doi.org/10.1198/016214504000000601.

Merkouris, T. 2010. "An Estimation Method for Matrix Survey Sampling." In Proceedings of the Section on Survey Research Methods: American Statistical Association, July 31 to August 5, 2010, Vancouver, Canada. 4880–4886. Available at: https://www.amstat.org/Sections/Srms/Proceedings/y2010/ Files/308769_61580.pdf (accessed March 2016).

Merkouris, T. 2013. "Composite Calibration Estimation Integrating Data from Different Surveys." In Proceedings of the 59th ISI World Statistics Congress, 25–30 August 2013, Hong Kong. 205–210. Available at: http://2013.isiproceedings.org/Files/IPS020-P3-S.pdf (accessed March 2016).

Merkouris, T. 2015. "An Efficient Estimation Method for Matrix Survey Sampling." *Survey Methodology* 41: 237–262.

Meyer, R.K. and C.J. Nachtsheim. 1988. "Constructing Exact D-optimal Experimental Designs by Simulated Annealing." *American Journal of Mathematical and Management Sciences* 8: 329–359. Doi: http://dx.doi.org/10.1080/01966324.1988.10737244.

Mitra, D., F. Romeo, and A. Sangiovanni-Vincentelli. 1986. "Convergence and Finite-Time Behavior of Simulated Annealing." *Advances in Applied Probability* 18: 747–771.

Park, I. and H. Lee. 2004. "Design Effects for the Weighted Mean and Total Estimators under Complex Survey Sampling." *Survey Methodology* 30: 183–193.

R Core Team. 2013. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/ (accessed March 2014).

Raghunathan, T.E. and J.E. Grizzle. 1995. "A Split Questionnaire Survey Design." *Journal of the American Statistical Association* 90: 54–63.

Reis, F. 2013. "Links Between Centralisation of Data Collection and Survey Integration in the Context of the Industrialisation of Statistical Production." Working paper presented at the UNECE Seminar on Statistical Data Collection, 25–27 September, 2013, Geneva, Switzerland. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP2.pdf (accessed January 2015).

Renssen, R.H. and N.J. Nieuwenbroek. 1997. "Aligning Estimates for Common Variables in Two or More Sample Surveys." *Journal of the American Statistical Association* 92: 368–374. Doi: http://dx.doi.org/10.1080/01621459.1997.10473635.

Smith, P. 2009. "Survey Harmonization in Official Household Surveys in the United Kingdom." In Proceedings of the ISI World Statistical Congresses, 16–22 August 2009, Durban, South Africa.

Tsallis, C. and D.A. Stariollo. 1996. "Generalized Simulated Annealing." *Physica A* 233: 395–406. Doi: http://dx.doi.org/10.1016/S0378-4371(96)00271-3.

UNSTATS 2005. *Household Sample Surveys in Developing and Transition Countries*. New York: United Nations (ST/ESA/STAT/SER.F/96.). Available at: http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf (accessed January 2015).

Received February 2015
Revised July 2015
Accepted October 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 287–289, http://dx.doi.org/10.1515/JOS-2016-0014

Discussion

James O. Chipperfield[1]

The traditional survey paradigm has been to collect *all* variables from *all* respondents. This paradigm, which results in the well-known Single Phase Design (SPD), is being challenged by recent trends. These trends include: a decrease in response rates; a demand for more information to be collected, as analysts become more sophisticated; availability of inexpensive, but perhaps less-reliable, secondary sources of data (e.g., administrative data, the internet or Big Data) which in turn may be used as a substitute for survey data; an increase in the burden that may be imposed on respondents (e.g., require medical procedures, such as taking blood samples); and widespread use of computer-assisted interviewing that allows flexible sequencing of respondents through a questionnaire (e.g., household screening in order to target units of special interest or rare subpopulations); and increase in costs of administering surveys.

In response to these trends, survey organisations are looking to alternative, more flexible and efficient survey paradigms such as Split Questionnaire Designs (SQDs). SQDs relax the constraint of "collecting *all* variables from *all* respondents", which in turn allows more flexible ways of redressing these current trends. The article of Ioannidis et al. is therefore a timely and welcome contribution. I enjoyed reading it and I hope many more like it will follow.

While current trends do encourage the use of SQDs, the idea of SQDs is not new. For about the last 15 years, reviews by statistical agencies of their survey data collection strategies have recommended use of SQDs in some form, citing many of the benefits noted in the introduction of Ioannidis et al. So why are SQDs still not standard practice? Perhaps it is because SQDs present new and difficult methodological problems that can significantly increase the complexity of the 'survey cycle' and require substantial investment in new systems. The intricacy of the SQD design problem discussed by Ioannidis et al. is a case in point! So perhaps much more methodological work, some of which is discussed below, is required before SQDs are standard practice.

I now turn to comment on Ioannidis et al. The article is about optimal sample design. The objectives of the sample design are described in the traditional way of balancing accuracy and cost. However, what is far from traditional is that the authors consider the optimal sample design for an SQD rather than an SPD. The optimal SQD is defined in terms of $\mathbf{n} = (n_1, \ldots n_j \ldots, n_k)$, where n_j is the number of respondents assigned instrument j, an instrument is made up of a selection of questionnaire modules, and k is the

[1] Associate Professor (Adjunct), National Institute for Applied Statistics Research, University of Wollongong NSW 2522, Australia. Email: james.chipperfield@abs.gov.au

total number of instruments used in the design. I would like to make a few comments about the set-up of the sample design problem.

First, the set-up assumes that the Horvitz-Thompson (HT) estimator is used for estimation. This estimator does not exploit correlations between variables collected in different modules. Exploiting these correlations to improve the accuracy of estimates, whether using a model-based likelihood approach (Rubin and Little 2002) or by using a finite sampling model-assisted approach (Merkouris 2004), could perhaps be factored into the design problem. For example, consider the situation whereby Modules A and B contain variables that have a known and high correlation. Collecting Module A but not Module B from a respondent would contribute, due to the correlations, would also contribute, a non-zero amount to the effective sample size of Module B.

Second, traditional survey designs have almost exclusively been designed for estimating means or totals. Analysts interested in model-fitting are often called *secondary analysts*, because they are not the primary consideration during the survey design process. This is perhaps because, given the wide variety of possible analyses, designing for analysts' requirements is difficult. Nevertheless, traditional survey designs have historically met the needs of analysts for two reasons: (1) all modules are collected from all respondents meaning there is no loss of information about interactions between variables collected by different modules; and (2) the sample size for accurate estimates of subpopulation means is sufficient for accurate estimates of model parameters, where subpopulation is often treated as a marginal effect. However, in the case of SQD these reasons may not apply. For example, if an SQD only collects two out of five modules from any respondent, then information about two-way interactions would be available but no information about three-way (or higher) interactions would be available. So the SQD design problem may need to explicitly take into account the needs of analysts. While Ioannidis et al acknowledge the needs of analysts via 'enforcing crossings', I wonder whether measures of accuracy for a broad class of analysis could be incorporated into the design, as they are for population means.

Third, instruments are assigned to respondents independently of their characteristics. This means data not collected by the SQD are Missing Completely At Random (MCAR). We could instead assign instruments to respondents with a probability that depends upon the respondent's characteristics. This means the data not collected by the SQD are Missing at Random (MAR). Chipperfield et al. (2013) considered assigning instruments to respondents with a probability that depended upon the respondent's diabetes status collected during the interview (diabetes effects about 5% of people in the Australian state of NSW). In a logistic model with diabetes as the outcome variable, a person *with* diabetes contributes about the same amount of *information* (in a likelihood sense) as 400 people *without* diabetes. So given diabetes status, collecting the model's covariates from people *with* diabetes is much more efficient than collecting them from people *without* diabetes.

It is also worth mentioning search algorithms for the optimal SQD. When Chipperfield and Steel (2009, 2011) and Chipperfield et al. (2013) search for the optimal SQD they do not impose a constraint on the set of instruments (i.e., combination of modules). In other words, they allow *all* $k = 2^m - 1$ possible instruments to be used in the optimal design, where m is the number of modules. However, this is computationally infeasible even for moderate m. Ioannidis et al avoids this computational problem by considering only a limited set of instruments, denoted by the matrix **A**, at each iteration (i.e., $k \ll 2^m - 1$). Across iterations, the algorithm searches for the optimal set of instruments. So Ioannidis et al. optimises over both **A** and **n**, and allows k to be set at the design stage rather than determined by the value m. This is a very useful development for moderate and large m.

In conclusion, it is hard to ignore that administrative data and Big Data will shape the way official agencies collect data in the future. I can see a role for an MAR-SQD whereby a respondent is assigned each module with a particular probability, where this probability depends on the information that is known about them from an administrative source. For example, if a person's health record shows that they are an unusually high user of medicines given their demographic characteristics, they may be more likely to be given a 'health' module. Their response values to the health module may affect the probability that they are given an 'education' module, and so on. These probabilities could be set to improve the efficiency of the SQD.

References

Chipperfield, J.O. and D.G. Steel. 2009. "Design and Estimation for Split Questionnaire Designs." *Journal of Official Statistics* 25: 227–244.

Chipperfield, J.O. and D.G. Steel. 2011. "Efficiency of Split Questionnaire Surveys." *Journal of Statistical Planning and Inference* 141: 1925–1932. Doi: http://dx.doi.org/10.1016/j.jspi.2010.12.003.

Chipperfield, J.O., M. Barr, and D.G. Steel. 2013. "Split Questionnaire Designs: Are They an efficient Design Choice?". In Proceedings of the 59th ISI World Statistics Congress, 25–30 August 2013. 311–316. Hong Kong. Available at: http://2013.isiproceedings.org/Files/IPS033-P1-S.pdf (accessed February 2016).

Merkouris, T. 2004. "Combining Independent Regression Estimators from Multiple Surveys." *Journal of the American Statistical Association* 99: 1131–1139. Doi: http://dx.doi.org/10.1198/016214504000000601.

Rubin, D.B. and R.J.A. Little. 2002. *Statistical Analysis of Missing Data* (2nd Edition). John Wiley and Sons.

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 291–294, http://dx.doi.org/10.1515/JOS-2016-0015

Discussion

David Dolson[1]

Amongst other things National Statistical Offices (NSO) and their coordinating bodies strive for information production strategies that are fit for their intended uses, cost efficient and can control respondent burden to an acceptable level. A common theme is survey integration (High-Level Group for the Modernisation of Official Statistics 2014; Priest, G. 2010; and United Nations Statistics Division and the Statistical Office of the European Union 2015), which in addition has the desirable effect of leading towards improved coherence of statistical information across survey programs and between countries. These objectives are often characterized in terms of standardized production processes, commonly accepted concepts, and harmonized survey questionnaire content.

Reis (2013) describes a strategy being taken for the improved integration of European social surveys. The essence of the approach is an integrated system of surveys based upon a set of standardized modules of survey questions which can be combined as needed into a set of different survey instruments to meet different needs in terms of frequency and precision of estimates.

The current article by Ioannidis et al. addresses a general approach to survey integration for social surveys conducted by a NSO. The situation described by Reis would be a specific instance of this challenge. The article describes how, compared to the status quo of more stove piped approaches, such a strategy has the potential to reduce both respondent burden and costs while continuing to produce estimates to meet precision requirements. The main focus of the article is development of a solution to the difficult dual optimization problem of 1) determining the survey modules to be included in each instrument and 2) the related sample sizes given various constraints on the modules and instruments (such as required precision of estimates, required joint observations of modules and periodicities of modules and instruments).

Their solution is an appealing and powerful if perhaps somewhat complex (for me!) method using simulated annealing and the simplex algorithm. It provides advances relative to split questionnaire design approaches that make it a superior solution for this survey integration challenge. Its power and flexibility provide a valuable method for use by any organization contemplating the kind of large scale survey integration considered here and in Reis (2013).

However, as the authors quite appropriately point out, an initiative towards this kind of integration across an entire program of social surveys would comprise several other major elements.

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

Although not currently contemplating this kind of large scale survey integration, Statistics Canada has made significant headway on several of the fronts that would be precursors. The unifying strategic direction is use of solutions that are globally optimal rather than locally optimal for each individual survey program. All of these approaches are directed towards reduced cost and risk associated with development, maintenance and use of multiple survey methods and systems as well as improved coherence of the information produced by the social survey program (Statistics Canada 2005). Through the development and use of generalized methods and systems for data collection, data processing, sampling, estimation and dissemination, risks and costs associated with using and maintaining multiple systems are reduced and better controlled (Brisebois and Dufour 2015).

Through use of our Address Register, we have implemented a common household survey frame service (MacNabb et al. 2011). It provides efficiencies in survey development and operation and allows for both negative and positive, as appropriate, sample coordination strategies to manage respondent burden while also facilitating sample overlap between surveys as needed.

In an initiative started several years ago, harmonized content modules with standardized concepts and question wording have been developed for use in paper questionnaires and for electronic questionnaires (Nadwodny and Best 2011). This work continues and has been particularly long and challenging as various surveys, each according to its particular redesign cycle, have adapted to possibly slightly different concepts, changed questionnaire wording and possible mode effect changes associated with updated data collection modalities. Creating common questionnaire modules takes time and careful testing to ensure acceptable usability across the surveys that will use them.

An Integrated Collection Operations System (Statistics Canada 2012) has been implemented that provides standardized questionnaire development techniques using the standardized modules, a consistent look and feel for all electronic questionnaires and common data collection operation and management tools.

In a related development, social and household surveys now use a common social survey processing environment that implements consistent methods and systems for steps such as data capture, editing, coding and imputation (Nadwody and Best 2011). Over many years a suite of generalized systems (Deguire et al. 2011), applicable for both household (social) and business surveys, has been developed, implementing commonly used statistical methods for sample design, edit and imputation, estimation, disclosure avoidance and tabulation.

Standardized methods, systems and strategies are to be used and strong justification is required for any exceptions.

In undertaking this at Statistics Canada, surveys have been able to continue to produce data of similar accuracy as previously with the benefit of improved coherence of statistical products. Use of generalized strategies and systems has been successful in reducing risks associated with multiple survey specific systems. Financially though, important investments must be made up front in harmonization of concepts and questionnaire modules as well as development of generalized systems. Even without the kind of large scale integration contemplated in this article, implementing these generalized methods and tools in a program of surveys is a substantial undertaking over many years with the effect that expected savings are only gradually realized.

Summarizing, strategies and tools such I have just outlined are useful and valuable in their own right to meet the strategic goals I noted above, providing important benefits for a statistical office or set of such offices. They are also an important element in moving towards improved and expanded survey integration, including helping to provide a flexible framework to adapt to changing needs in the social survey program. The development of the methods described in the current article provides a useful addition to the set of methods available in moving forward beyond what I have outlined towards effective large scale survey integration. Although powerful, the technique is also a complex one and the additional challenges in the kind of large scale integration being considered are substantial. It will be interesting to observe statistical offices in their assessment of the benefits of large scale integration and using this technique and their choices in whether and how best to proceed.

References

Brisebois, F. and J. Dufour. 2015. “Enquêtes santé au Canada : S’adapter aux nouvelles réalités de la société canadienne.” Paper presented at the Journées de Méthodologie Statistique de l’INSEE, Paris, France, March 2015. Available at: http://jms.insee.fr/files/documents/2015/S13_5_ACTE_V2_BRISEBOIS_JMS2015.PDF (accessed March 29, 2016).

Deguire, Y., L. Reedman, and M. Wenzowski. 2011. “Generalized Systems: The Statistics Canada Experience.” Proceedings of Statistics Canada Methodology Symposium 2011.

High-Level Group for the Modernisation of Official Statistics. 2014. “Strategic Vision of the HLG.” Available at: http://www1.unece.org/stat/platform/display/hlgbas/Strategic+vision+of+the+HLG (accessed March 29, 2016).

MacNabb, L., M. St-Pierre, and M. Grenier. 2011. “Development of a Common Frame for Household Surveys at Statistics Canada.” Proceedings of Statistics Canada Methodology Symposium 2011.

Priest, G. 2010. “The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency – A Case Study of Statistics Canada.” International Household Survey Network working paper No. 4. Available at: http://www.ihsn.org/home/integrating (accessed April, 2016).

Nadwodny, R. and P. Best. 2011. “Harmonized Content and Common Processing Tools: The New Paradigm in Developing Surveys at Statistics Canada.” In Proceedings of Statistics Canada Methodology Symposium 2011.

Reis, F. 2013. “Links Between Centralisation of Data Collection and Survey Integration in the Context of the Industrialisation of Statistical Production.” Working paper presented at the UNECE Seminar on Statistical Data Collection, Geneva, Switzerland. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2013/mgt1/WP2.pdf (accessed April, 2016).

Statistics Canada. 2005. “New Household Survey Strategy: Summary Report.” Internal Working Paper.

Statistics Canada. 2012. “Integrated Collection and Operations System (ICOS) Summary Presentation.” Internal presentation from Collection Planning and Research Division.

United Nations Statistics Division and the Statistical Office of the European Union. 2015. "Proceedings of the Global Conference on a Transformative Agenda for Official Statistics: Outcomes and Summaries of Sessions." Background document from the Global Conference on a Transformative Agenda for Official Statistics, New York, USA. Available at: http://unstats.un.org/unsd/nationalaccount/workshops/2015/NewYork/lod.asp (accessed March 29, 2016).

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 301–305, http://dx.doi.org/10.1515/JOS-2016-0017

Rejoinder

Evangelos Ioannidis, Takis Merkouris, Li-Chun Zhang, Martin Karlberg, Michalis Petrakos, Fernando Reis, and Photis Stavropoulos

We would like to thank all the discussants for their thoughtful and encouraging comments. Not surprisingly, there is some overlap among the issues raised. Below we organise our response in three parts, in correspondence with the discussion, and note the connections and overlaps across the parts when and where we consider appropriate.

1. Chipperfield

In his first comment, the author points out that the estimation method that we assumed in designing an integrated survey does not exploit correlations between variables collected in different modules to improve the accuracy of estimates. Thus he suggests that this improvement, achievable by suitable estimation methods, could be factored into the design problem, to make the design more efficient. Gonzalez and Eltinge make a similar comment in the last paragraph of (their) Section 2.

Our approach to designing an integrated survey assumes a baseline estimation procedure, involving standard Horvitz-Thompson (HT), estimators for items surveyed in a single instrument and simple but efficient composite HT estimators combining data on common items surveyed in different instruments. This general approach, requiring only estimates of design effects, is applicable to any setting of an integrated survey. As is the norm in the case of a single survey, we do not factor into the design of an integrated survey the effect of a regression or calibration estimator which might be used. Design-based estimation methods (cited in our article) that exploit correlations between variables for improved accuracy are in fact special calibration procedures, whereby estimates of the same totals from different instruments are aligned. Such methods can be used profitably in our setting, but the design we propose is not contingent on the use of any of them for the following reasons.

Firstly, an exact theoretical quantification of the correlation effects is intractable for the type of surveys under consideration, involving complex sampling designs plus multiple instruments of varying composition and periodicity of modules, different production timetable for various statistics, etc. Furthermore, it is not at all obvious how to factor such (variable-dependent) effects into the sampling design, in a manner analogous to the design-effect scalar adjustments. This would essentially require a measure of compound design-regression-correlation effect that accounts for the interaction between the three components, ideally for a number of important items. Devising such a measure seems to be an extremely challenging task.

Secondly, factoring correlations into the design depends on the particular process of exploiting them, and its interaction with customary calibration. But such estimation/

calibration procedures may be optional or subject to revision over time, and thus it is not sensible to embed their effect into the fixed survey design.

In his second comment, the author raises the issue of using data from a survey for analytic purposes, in addition to the descriptive purposes served by traditional survey designs. While he points out that our design for an integrated survey can accommodate the needs of analysts via ‘enforcing crossings’, which provide the necessary information on various interactions between variables, he wonders “whether measures of accuracy for a broad class of analysis could be incorporated into the design, as they are for population means”.

For the typical setting of SQD discussed by the author, with specialized survey requirements, an explicit incorporation of such measures of accuracy into the design, within an analytic framework of modelling methods, may well be considered (see Chipperfield and Steel 2011). But for an integrated survey with wide-ranging and primarily descriptive requirements, and an already complex optimization algorithm involving multiple constraints, such an expanded design encompassing modelling considerations might not be practical.

In his third point, the author suggests assigning instruments to respondents with a probability that depends upon the respondent’s characteristics, and cites a particular application of this idea (Chipperfield et al. 2013). Gonzalez and Eltinge (2008) also proposed an adaptive assignment of subsampling probabilities based on data (e.g., demographic) from the first interview in a panel consumer expenditure survey with a split-questionnaire design. See also the second “adaptive design” option suggested by Gonzalez and Eltinge in Section 3 of their discussion, in the context of rare populations and low prevalence characteristics.

Such a procedure of assigning instruments to respondents could be well adopted in our setting for increased design efficiency, if the instruments are administered to subsamples from an initial sample that collects the necessary information: a process resembling two-phase sampling. As a design feature, the resulting increase in design efficiency would be factored into the design effect, although at increased design complexity. Akin to this theme is the author’s consideration, in his concluding remarks, of the possibility of using administrative data to determine a respondent’s assignment to a particular module, to enhance the efficiency of the design. This possibility is worth considering when designing an integrated survey. See also below our discussion of “adaptive design options” as considered by Gonzalez and Eltinge.

2. Gonzalez and Eltinge

Gonzalez and Eltinge discuss a number of potential “complements, and possible extensions” to our approach. We agree that many of them seem worth studying in greater detail in the future. Here, we comment specifically on two of them.

Firstly, Gonzalez and Eltinge express a very relevant concern for estimation in “multiple” or even “a large number of domains” (Section 2, second last paragraph). Insofar as the *same* instrument is to be administered to every sample unit, a practical solution could be to use in our Equations (3) and (4) the overall sample size, corresponding to a sampling design that appropriately balances between the national and multi-domain estimation purposes, and to control for the desired domain sample sizes in sample selection.

Secondly, an interesting feature of the "schematic" model (1.1) and (1.2) in Section 1 is the "deviation terms" e_Q and e_C. The survey Quality and Cost are thereby made random, instead of being completely determined at the design stage. We find this a plausible and potentially useful perspective, in order to accommodate the "adaptive design options" (Section 3), in the spirit of the MAR-SQD approach considered by Chipperfield in his third comment.

Let us consider **n*** on the right-hand side of our Equation (4) as 'the minimum required number of ideal (i.e., complete and error-free) observation units' for each module. On the left-hand side, instead of the fixed sample sizes of all the instruments, let **n** be a matrix of the same dimension as **A**, where n_{ij} is in general a *random* number of ideal observation units for module *I* arising from the sample of instrument *j*. These can be random because of the presence of adaptive design options, such as two-phase design subject to screening, possible substitution of survey questionnaire by administrative data, adaptive assignment of proxy/backup modules (Section 3), or MAR-SQD (Chipperfield, 3rd comment), etc. We can now for example replace Equation (4) with

$$\mathrm{E}\{\mathrm{Diag}(\mathbf{A}\mathbf{n}^{\mathbf{T}})\} \geq \mathbf{n}^{*},$$

in which case one requires that the survey accuracy satisfy, *in expectation,* the minimum requirement. Or, we can for example, use instead

$$\Pr\{\mathrm{Diag}(\mathbf{A}\mathbf{n}^{\mathbf{T}}) \geq \mathbf{n}^{*}\} \geq \alpha_{m\times 1},$$

for chosen threshold α-values, provided it is possible to calculate these probabilities.

Similarly, the cost function can be made stochastic and the minimisation could be with respect to the *expected* cost instead of the fully deterministic one. Together, they could provide the starting point for a modular design approach that allows for adaptive design options.

3. Dolson

In our article, we focus on the dual design problem of determining the optimal instrument composition and appropriate sample sizes given a certain instrument composition. However, as pointed out by us and confirmed by Dolson, the application of these methods requires several other major elements.

Karlberg et al. (2015) provides a synoptic overview of the "Streamlining and integration of the European social surveys" project, and enumerates many such challenges. It is encouraging to see in Dolson's description how Statistics Canada has made headway regarding many of these components, such as (using the terminology of Karlberg et al. 2015) *harmonization of variables*, *definition of modules*, *harmonisation of sampling frames* and *IT infrastructure issues*. Still, as noted by Dolson, the additional challenges in the large-scale integration are substantial. In this connection, Karlberg et al. (2015) bring up regulatory and governance issues, user relations (eliciting user needs in terms of required precision rather than sample sizes) as well as issues triggered by the increase in the number and internal heterogeneity of instruments (going from a "one instrument – one survey" situation to a situation with multiple, multi-thematic instruments, and the challenges for interviewers that this would pose).

We can only agree with Dolson's conclusion, that "It will be interesting to observe statistical offices in their assessment of the benefits of large scale integration and using this technique and their choices in whether and how best to proceed." This will, to a large extent, depend on the political will to integrate and the path chosen towards integration. In developing countries building a statistical system from scratch, it would of course make sense to deploy an integrated system right away, but for advanced statistical systems, such as the one in Canada, it would not be advisable to go for a "big bang" approach. Still, as discussed by Gonzalez and Eltinge, "integration of certain groups of surveys might be feasible" (Section 4).

Karlberg et al. (2015) propose a gradual roll-out, in which the focus would be precisely on the issues where developments at Statistics Canada have already taken place. The first objective would be to achieve "**pooling maturity**", that is, a system that allows data to be pooled across surveys to provide more precise estimates. The key requirement here is that variables are harmonized across surveys and that the sampling frames are aligned; some attention also has to be given to complex indicators (such as the poverty rate). In all likelihood, Statistics Canada could already conduct pooling for surveys where concepts and frame have been harmonized (thereby obtaining increased precision "for free") – perhaps this is already done on an experimental basis, or even in a production setting?

Only then would one proceed to actually modify the design of the surveys. The subsequent step would be to reach "**reallocation maturity**", that is, a system which would allow the application of the simplex algorithm, as described in our article, to find a solution that is globally optimal taking into account that data would be pooled – with the major constraint that the existing survey instruments would remain unchanged. Technically, this step is trivial, as it mainly requires that the way the precision requirement is specified is harmonised between surveys. However, it could generate controversy in a "stovepipe setting", since surveys would need to accept a reduced sample size and to rely on other surveys in order to reach the total sample sizes needed for their required precision. This second step might only yield quite marginal gains in terms of cost, since excessive sampling would still take place for variables with low precision requirements. This would be the case when variables are administered in the same survey as variables with high precision requirements. As this step combines potential controversy with presumably marginal gains, the gains should be assessed before it is practically implemented. If the gains are marginal, it might be better to refrain from taking this step in isolation, and instead strive to achieve "**recomposition maturity**", that is, a system in which all technical, organisational and methodological challenges have been addressed, so that current survey questionnaires can be recomposed into modular instruments through the application of the optimization algorithms presented in our article.

4. References

Chipperfield, J.O. and D.G. Steel. 2011. "Efficiency of Split Questionnaire Surveys." *Journal of Statistical Planning and Inference* 141: 1925–1932. Doi: http://dx.doi.org/10.1016/j.jspi.2010.12.003.

Chipperfield, J.O., M. Barr, and D.G. Steel. 2013. "Split Questionnaire Designs: Are They an Efficient Design Choice?" In Proceedings of the 59th ISI World Statistics Congress, 25–30 August 2013, Hong Kong. 311–316. Available at: http://2013.isiproceedings.org/Files/IPS033-P1-S.pdf (accessed June 2015).

Gonzalez, J.M. and J.L. Eltinge. 2008. "Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey." American Statistical Association, In Proceedings of the Section on Survey Research Methods, Denver, Colorado, August 6, 2008. 3069–3075. Available at: http://www.amstat.org/sections/srms/proceedings/y2008/Files/301351.pdf (accessed April 2016).

Karlberg, M., R. Reis, C. Calizzani, and F. Gras. 2015. "A Toolbox for a Modular Design and Pooled Analysis of Sample Survey Programmes." *Statistical Journal of the International Association for Official Statistics* 31: 447–462.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 307–328, http://dx.doi.org/10.1515/JOS-2016-0018

The Impact of Question Format, Context, and Content on Survey Answers in Early and Late Adolescence

Nadine Diersch[1] *and Eva Walther*[2]

Self-reports in surveys are often influenced by the presented question format and question context. Much less is known about how these effects influence the answers of younger survey respondents. The present study investigated how variations in response format, answer scale frequency, and question order influence self-reports of two age groups: younger (11–13 years old) and older (16–18 years old) adolescents. In addition, the impact of the respondents' level of familiarity with the question content was taken into account. Results indicated that younger adolescents are more strongly influenced by the presented question format and context than older adolescents. This, however, was dependent on the particular question content, implying that response effects are more pronounced when questions deal with issues that lie outside of the respondents' field of experience. Implications of these findings in survey research with younger respondents are discussed.

Key words: Attitude judgments; question order effects; social influence; survey methodology; younger survey respondents.

1. Introduction

Self-reports are often the method of choice in the social and behavioral sciences to collect information about people's attitudes, beliefs, intentions, and behaviors (Krosnick 1999; Schwarz 1999). The age group between 5 and 15 years of age is of growing importance in public opinion and consumer research due to their increasing purchasing power and influence on families' purchasing decisions (De Leeuw et al. 2004; Heinzel 2000; McNeal 1992). In large-scale surveys like the international "EU Kids Online" survey, for example, thousands of children and adolescents aged 9 to 16 are regularly asked about their living situation, values, behaviors, and future plans (cf., Livingstone et al. 2012). Given the challenges that the inclusion of younger age groups poses to survey research, Scott (2008) concluded in her review on children as respondents that "improved data about children are essential in a society where children's role as consumers and citizens is being taken increasingly seriously in the economy, in law, and in social policy" (Scott 2008, 103). Thus, a better understanding of whether younger respondents interpret the questions as

[1] Aging & Cognition Research Group, German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany, and Department of Psychology, Chemnitz University of Technology, Chemnitz, Germany. Email: nadine.diersch@dzne.de
[2] Department of Psychology, University of Trier, Universitätsring 15, 54296 Trier, Germany. Email: walther@uni-trier.de

Acknowledgments: The authors thank the pupils of the "Diesterweg-Gymnasium" in Plauen, Germany, for donating their time to take part in the study, Karin Zahradnik and Rüdiger Stumm for their support in conducting the survey at this school, and Norbert Schwarz for helpful comments on an earlier version of this article.

intended by the (adult) researcher is needed in order to collect meaningful data in these age groups and interpret the results accordingly.

In general, answering a survey question is a complex cognitive process consisting of several tasks, ranging from understanding the intended meaning of the question to retrieving relevant information from memory, forming an appropriate judgment, and finally translating it into an answer provided by the questionnaire (Tourangeau and Rasinski 1988). It has been suggested that this answer process might be understood as a form of conversation (see Schwarz 1999; Strack and Schwarz 2007 for reviews). In contrast to a natural conversation, however, direct feedback from the conversation partner (i.e., the researcher) is lacking in standardized surveys, which in turn might lead to a greater reliance on contextual cues in order to infer the intended meaning of a question. Thus, social or contextual influences can affect any stage of the answer process, as indicated by a vast amount of research (e.g., Schuman and Presser 1981; Schwarz 1999; Schwarz and Bohner 2001; Schwarz and Hippler 1995; Schwarz and Oyserman 2001; Strack and Schwarz 2007). The results of these studies suggest that the interpretation and categorization of a question may be influenced strongly by contextually provided cues such as the respective question wording, format, and order (cf. Gawronski and Cesario 2013). However, the sensitivity to the context may be modulated by the particular topic in question and the respondents' level of familiarity and knowledge about it (cf. Festinger 1954). In line with this, Bickart (1992) provided evidence that brand evaluations are influenced more strongly by the presented question context in respondents who possess low levels of category-specific knowledge as compared to highly experienced respondents.

Although these response effects are well documented in adult respondents, much less is known regarding such effects in younger survey respondents. Because the answer process in a survey draws on several cognitive resources such as attention and memory, age-related differences in cognitive functioning are likely to result in different answers to the very same question as a function of age (see Knäuper et al. 2007; Schwarz 2003 for examples of survey research with respondents aged 60 and older). In accordance with cognitive-developmental theories (see Goswami 2010 for an overview), it is usually assumed that data reliability increases with age and cognitive development (cf. Borgers et al. 2000; Fallon Jr. and Schwab-Stone 1994; Reynolds 1993). For example, Borgers et al. (2000) concluded in their secondary analysis of existing data with different age groups that children around the age of eleven are able to give relatively consistent answers although they are very context sensitive and may invest less effort into the question-answer process if they are not interested in the topic in question or if the meaning of the question is somewhat unclear to them. In their review of empirical evidence about developmental influences on the ability to answer survey questions, De Leeuw et al. (2004) pointed out that cognitive functioning (e.g., memory) is comparable to that of adult respondents at the end of middle childhood (7 to 12 years of age), but that suggestibility resulting from social and motivational factors might influence self-reports of this age group. Children, more so than adults, tend towards answers they perceive as socially desirable or that reflect their own subjective wishes (Kränzl-Nagl and Wilk 2000; Reynolds 1993). Children below the age of twelve are generally more susceptible to social influence than adults, particularly when they receive information from authorities (Ceci et al. 1987; Warren and Lane 1995).

Adolescents at the age of 16, by contrast, are assumed to be able to answer survey questions similarly to adult respondents (De Leeuw et al. 2004). Thus, younger respondents might also take the presented question format and context into account when answering a survey, especially when they are not very familiar with the topic in question. Below the age of 16, social desirability concerns and satisficing, that is, the reliance on simple heuristics to give an answer due to a lack of motivation, might additionally bias the answers in a survey (Borgers et al. 2000; De Leeuw et al. 2004).

In the following, four types of response effects will be outlined that are well documented in the literature on survey research with adult respondents. Next, studies will be reviewed providing a first indication of how younger survey respondents might be influenced by the presented question format and context. Finally, an overview of the present study will be provided.

1.1. Effects of Question Format and Context on Self-Reports in a Survey

Adult respondents often react differently when they are offered a specified set of response alternatives compared with a situation in which no response alternatives are given (e.g., Kane and Schuman 1991; Schuman and Presser 1981; Schwarz 1999; Schwarz and Oyserman 2001; Strack and Schwarz 2007). In an open response format, relevant information has to be retrieved from memory in a first step. In the next step, a decision has to be made as to which information is relevant in the context of the question. In a closed response format, the first step can be skipped, resulting in more responses on average. Moreover, predefined response alternatives might indicate which level or type of response is considered appropriate in the context of the survey. In addition, several studies have shown that the frequency of answer scales and their reference points may be used as a frame of reference to infer the intended meaning of a question, which in turn might influence judgments about the frequency of own behaviors (e.g., Schwarz 1999; Schwarz and Bienias 1990; Schwarz et al. 1985; Schwarz and Oyserman 2001; Strack and Schwarz 2007).

Not only reports about actual behaviors are highly context-dependent; the same is true of reports about personal opinions, interests, and attitudes. This implies that attitudes are often “constructed on the spot” (Schwarz and Bohner 2001, 442). Numerous studies have shown that preceding questions influence the answers to subsequent questions (e.g., Bishop 1987; Krosnick 1991; Krosnick and Alwin 1987; Schuman and Presser 1981; Schwarz 1999; Schwarz and Hippler 1995, Strack 1992; Strack and Schwarz 2007; Tourangeau et al. 1989). According to the norm of evenhandedness, respondents answer two questions that are similar in some important aspect in an evenhanded way in order to appear consistent (Schuman and Presser 1981). Thus, a personal preference might be adjusted to an answer on a preceding question if this norm is evoked. Comparable mechanisms operate when respondents have to answer certain filter questions before or after they report personal attitudes and opinions on the same topic (Strack 1994). These questions, whether they are related to own behaviors or knowledge about the topic in question, also make information available that might shape the answers to following questions (cf., Bishop 1987; Bishop et al. 1983; Schwarz et al. 1991). Self-perception and self-evaluation may be altered by answering a filter question, resulting in changes in the focus of attention afterwards in order to reaffirm the self-concept (Martin and Harlow 1992).

1.2. Effects of Question Format and Context on Self-Reports of Younger Survey Respondents

To date, there are only very few studies that investigate the impact of question format and context on the answers of younger respondents in a survey. For example, Borgers et al. (2004) provided evidence that the reliability of self-reported self-esteem and well-being in children and younger adolescents between 8 and 16 years of age is modulated by the number of presented response options, the usage of a neutral midpoint, and negatively formulated questions. It was shown that the internal consistency of the self-reports increases with increasing numbers of response options on a rating scale but decreases if more than six response options are presented. The consistency over time, however, showed the reverse pattern of results. Negatively formulated questions did not influence the reliability, although the answers generally differed between negatively and positively formulated questions. Possible answer differences within the sample that encompassed a relatively large age range, however, were not taken into account.

Fuchs (2005) examined the influence of question format and context on self-reports of different younger age groups in several independent studies. The respondents' cognitive-developmental status was additionally assessed by means of self-reported educational achievement. The results showed that the impact of response order, response scale, and numeric values associated with response alternatives was higher in children and younger adolescents (aged 10 to 13 in Study 1, aged 13 to 15 in Study 2) compared to older adolescents (aged 14 to 17 in Study 1, aged 16 to 17 in Study 2) and adults (aged 18 and older in all studies). In addition, the effects tended to be more pronounced in respondents with poor or intermediate educational achievement. At the same time, younger respondents (aged 13 to 15 and aged 16 to 17) were less affected by the presented question order than older respondents (aged 18 and older), indicating that information from preceding questions was not taken into account in the formation of a judgment on following questions. However, correlations between the different questions, the order of which was manipulated, were not reported. The analysis focused on relative differences in answers to the single questions as a function of question order. In general, it was concluded that limited cognitive skills may result in an incomplete understanding of a given question and an insufficient retrieval of relevant information, resulting in a less sophisticated question-answer process in younger survey respondents. Notably, the composition of the single age groups as well as the content of the critical questions varied across the single experiments, making it difficult to compare the data across experiments. In a follow-up study, standard tests of cognitive functioning (e.g., vocabulary and working memory tests) were applied in a sample of children between 8 and 14 years of age to examine the relation between age, cognitive functioning, and respondent problems (e.g., inadequate answers, uncertainty) in a face-to-face interview setting (Fuchs 2009). The results confirmed that increasing age leads to fewer problems in the question-answer process. This, however, was largely independent from differences in cognitive functioning. Thus, one might speculate that other factors such as social and/or motivational factors might contribute to the observed answer differences in younger survey respondents. As outlined above, children and younger adolescents might be more likely to construct certain attitudes "on the spot", resulting in answers that are based to a larger extent on the information available at the time they are asked, due to their higher suggestibility and satisficing strategies (De Leeuw et al. 2004).

1.3. Overview of the Present Study

The present study aimed at investigating how variations in question format and context affect the answers of two age groups, younger (11–13 years old) and older (16–18 years old) adolescents, when they are interviewed in a survey setting. The two age groups were chosen because they represent the lower and upper end of adolescence. Based on previous research, we assumed that social desirability concerns and motivational factors would become more evident in the younger age group, whereas cognitive functioning would be relatively comparable between the two groups (cf. De Leeuw et al. 2004). Going beyond previous research, we also addressed the impact of the particular question content on the answers of the two age groups. More specifically, we examined to what extent the subjective experience with a certain question domain would shape the potentially biasing influences of contextual cues provided by the questionnaire. In addition, by using a relatively short questionnaire that incorporated a number of aspects, we aimed to resemble a typical situation in survey research, where many data need to be gathered while keeping questionnaire length at a reasonable level in order to reduce dropouts. We developed two versions of a questionnaire in which four response effects (i.e., open vs. closed response formats, frequency of answer scales, question order in attitude judgments, and question order of filter questions in combination with attitude judgments) were varied within two different topics: environmental protection and sports activities. The two topics were selected because we assumed that respondents have less experience with the former as compared to the latter. As a consequence, reports about environmental protection activities should be more affected by social desirability concerns than sports activities, and if younger adolescents are more susceptible to social influence than older adolescents this should be become evident within this topic.

We hypothesized that the difference between the open and closed response format is stronger in younger than in older adolescents. Younger adolescents might invest less effort in the question-answer process and consequently rely more on predefined response sets than older adolescents, especially if the topic in question is rather unfamiliar and less interesting to them (cf. Borgers et al. 2000). In addition, we expected that younger adolescents would be more strongly influenced by the frequency of answer scale than older adolescents because they tend more towards socially desirable answers due to a higher suggestibility (cf., De Leeuw et al. 2004; Känzl-Nagl and Wilk 2000). Preceding questions might shape subsequent attitude judgments more in younger than in older adolescents because they are more context sensitive, especially when they have not already formed a firm attitude about the topic in question (cf. Borgers et al. 2000). Similarly, attitude judgments might be more strongly influenced in younger than in older adolescents by the success or failure answering a filter question beforehand due to information about own relevant behaviors that is brought to mind.

2. Method

2.1. Participants

The survey was completed by 188 pupils from a German secondary school following the survey's approval by the local school board. The sample consisted of two age groups: younger adolescents ($n = 104$, 54 female, $M_{age} = 12.2$ (mean), SD = 0.72 (standard deviation), age range: 11–13 years) and older adolescents ($n = 84$, 39 female, $M_{age} = 16.9$, SD = 0.61, age range: 16–18 years), and was largely homogeneous in terms of academic performance. All of them aspired to the highest educational qualification available in the German school system and when asked which school grade they achieved most frequently, 88.8% of the pupils reported to receive good or satisfactory grades on average. For respondents under the age of 18 ($n = 179$), informed consent was obtained from their parents. Three pupils whose parents refused permission were excluded from participation. Additional written informed consent was obtained from the respondents themselves.

2.2. Design and Material

Each age group received one of two questionnaire versions in which question format, frequency of answer scales, question order in attitude judgments, and order of filter questions in combination with attitude judgments were manipulated in a balanced order. Half of the questions in each questionnaire version referred to environmental protection activities and the other half to sports activities. In the section about environmental protection activities in one questionnaire version (Questionnaire A), for example, a certain question was presented in an open response format, whereas in the section about sports activities, a certain question was presented in a closed response format. In the other questionnaire version (Questionnaire B), the same questions were presented in the reverse format, that is, in a closed response format in the section about environmental protection activities and in an open response format in the section about sports activities. The remaining response effects were varied accordingly (see Table 1 for an overview of the questionnaire design). Special care was taken to word the question in an easy and understandable manner and to capture the characteristics of the two topics in an ecologically valid way. In the following, the implementation of each response effect within the two questionnaire sections will be outlined in detail.

2.2.1. Open vs. Closed Response Format

In order to investigate the influence of different response formats on the number and type of retrieved answers, a question in each thematic section was presented either in a closed response format or in an open response format. In the section about environmental protection activities, the question *"Are you actively involved in saving the environment?", if yes, "How do you do that?"* was presented with no answer alternatives in one questionnaire version, and with a list of twelve different answer alternatives (e.g., *"I am a member of an environmental protection organization"*) in the other questionnaire version. With regard to sports activities, the question *"Do you engage in sports?", if yes, "Which sports do you do?"* was varied correspondingly with a list of 31 different sports activities or with no answer alternatives.

Table 1. Variation of four response effects within two thematic sections in two questionnaire versions

	Questionnaire A	Questionnaire B
Response format		
Environ. protection activities	Open	Closed
Sports activities	Closed	Open
Frequency of answer scale		
Environ. protection activities	Low frequency	High frequency
Sports activities	High frequency	Low frequency
Question order in attitude judgments		
Environ. protection activities	1st pair: Specific – general 2nd pair: General – specific 3rd pair: Specific – general 4th pair: General – specific	1st pair: General – specific 2nd pair: Specific – general 3rd pair: General – specific 4th pair: Specific – general
Sports activities	4 items about sports – 4 items about soccer	4 items about soccer – 4 items about sports
Order of filter questions		
Environ. protection activities	After attitude judgments	Before attitude judgments
Sports activities	Before attitude judgments	After attitude judgments

2.2.2. Frequency of Answer Scales

The influence of answer scale frequencies on reported behavioral frequencies was tested in the section about environmental protection activities in that participants were asked whether they sought information about issues of environmental protection, and if so, how often they did that on average. In one questionnaire version, this was accompanied by a low-frequency scale with five answer alternatives ranging from *"Several times a week"* to *"Less than once a month"*. The high-frequency answer scale in the second questionnaire version comprised a reference period from *"Several times a day"* to *"Less than once a week"* with five answer alternatives (see Table 3 for item wording). Within the section about sports activities, the same two answer scales were used for the question "*Do you engage in sports?", if yes, "How often do you do that?"*.

2.2.3. Question Order in Attitude Judgments

To test the impact of question order in attitude judgments, different items consisting of statements that either referred to a rather general aspect of the topic in question (i.e., how everybody should behave or the general attitude towards the topic) were used together with more specific statements referring to subordinate and personal aspects of the respective topic. In the section about environmental protection activities, four item pairs were presented in a varying order across the two questionnaire versions (see Table 2 for item wording). The overall order of the item pairs was kept constant. Every item pair was separated from the others by a "buffer" item about animal protection issues to reduce spillovers between the item pairs. In the section about sports activities, the question order was varied in blocks such that in one questionnaire version four items about sports in general preceded four items about soccer, whereas in the other questionnaire version the general items followed the specific items. Soccer was chosen because we assumed that soccer is a sports activity performed by a considerable portion of the respondents and, even

Table 2. Wording of the attitude statements in each thematic section

		Attitude statements
Environmental protection activities		
1st item pair:		
	Specific	“I pay attention to act and to live environmentally friendly, even though it is sometimes quite difficult for me“
	General	“Everybody has to be absolutely concerned about the protection of the environment”
2nd item pair:		
	Specific	“For me personally, it is rather difficult to do something to save the environment” (reversed)
	General	“Everybody is responsible for the causes of environmental problems on our planet”
3rd item pair		
	Specific	“I can imagine well participating in projects that deal with environmental protection issues (e.g., volunteering)”
	General	“Environmental protection is an important topic to me”
4th item pair		
	Specific	“I am actively involved in saving the environment and, thus, make my personal contribution to the environment and its protection”
	General	“People behave in general very environmentally friendly”
Sports activities		
	Specific	“Soccer is a very trendy sport” “I think soccer is one of the most popular sports all over the world” “Soccer is my favorite sport” “I can imagine well participating in a soccer contest”
	General	“Sport is fun” “Living without sports is unhealthy” “People do not exercise enough” “Sport plays a central role in my life”

if this is not the case, they would encounter it on a regular basis (e.g., on television) due to its popularity in Germany. This should allow them to report a personal attitude about the topic that is relatively stable. Respondents had to answer these statements on a five-point answer scale ranging from *“I strongly agree”* to *“I strongly disagree”*.

2.2.4. Order of Filter Questions in Combination with Attitude Judgments

The impact of questions asking about actual behavior on attitude statements was tested by varying their order as a function of questionnaire version. For example, questions such as *“Are you actively involved in saving the environment?”, if yes, “How do you do that?”*, which were presented with different response alternatives (see above), were presented either before or after the attitude statements in each thematic section of the questionnaire.

2.3. Procedure

The study was administered via paper-and-pencil questionnaires. Each participant received a questionnaire that had to be filled out individually under supervision of a teacher. Participants were told that we were interested in their personal opinions and behaviors concerning two different topics, environmental protection and sports activities, to gain better insight into their individual attitudes. Special care was taken to ensure that participants sitting next to each other received the same questionnaire version to ensure that they did not realize that two versions of one questionnaire were being applied. The whole procedure took approximately 15 minutes. All of the respondents received some sweets after completion of the survey and were informed about the aim of the study and the usage of two questionnaire versions in which question format and context were varied. None of them reported that they had become aware of the two different versions while filling out the questionnaires.

3. Results

All data were first analyzed for main effects of question format and context across both age groups. In addition, we looked for answer differences as a function of age group. Due to the differences between the questions across the two thematic sections, data were analyzed separately for each topic.

3.1. Open vs. Closed Response Format

In the section about environmental protection activities, the question that was presented either in a closed or in an open response format was preceded by a filter question that asked about active engagement in these kinds of activities. The data showed that a considerable number of respondents from both age groups indicated in the filter question that they were *not* actively involved in saving the environment (younger adolescents: $n = 41$; older adolescents: $n = 40$). Thus the number of participants who answered the following question, in which the response format was varied, was significantly reduced. More importantly, in younger adolescents, $\chi^2(1, n = 104) = 14.54, p < .001$, as well as in older adolescents, $\chi^2(1, n = 84) = 9.36$, $p = .002$, this decrease in the number of respondents was significantly more pronounced in the questionnaire version in which an open response format was presented (younger adolescents: $n = 22$; older adolescents: $n = 15$) compared to the other questionnaire version (younger adolescents: $n = 41$; older adolescents: $n = 29$). In order to take the unequal sample sizes into account, nonparametric tests were used to determine the impact of the respective response format on the answers of the two age groups in this section of the questionnaire.

A Mann-Whitney U-Test on the number of retrieved answers across the whole sample confirmed a significant effect of response format, $U = 414.00$, $Z = 5.89$, $p < .001$, $r = .57$. In line with our hypothesis, significantly more answers were given in a closed response format ($Mdn = 4$ (median)) compared to an open response format ($Mdn = 2$). For example, when participants were asked whether they were actively involved in saving the environment and how they did that, 58.6% of them reported *"paying attention to saving energy and water at home"* when it was included in a list of answer alternatives, whereas only 2.7% gave an equivalent answer in the open response format.

Older adolescents ($Mdn = 2$) provided more relevant answers in an open response format than younger adolescents ($Mdn = 1$), $U = 85.00$, $Z = 2.68$, $p = .007$, $r = .44$. Younger adolescents ($Mdn = 4$), in contrast, showed the tendency to report more activities than older adolescents ($Mdn = 3$) when a list of alternatives was provided, $U = 443.50$, $Z = 1.83$, $p = .067$, $r = .22$. Thus, if they answered this question at all, younger adolescents seemed to be more strongly influenced by the presented question format than older adolescents within this section of the questionnaire.

In the section about sports activities, only a small fraction of respondents from both age groups denied the filter question that asked whether they engage in sports (younger adolescents: $n = 4$; older adolescents: $n = 9$). In addition, in both age groups, the answers did not differ as a function of the presented questionnaire version, all $\chi^2 \leq 1.24$, all $p \geq .724$. Thus, a 2 (response format: open vs. closed) × 2 (age group: younger adolescents vs. older adolescents) analysis of variance (ANOVA) on the number of retrieved answers was used in order to examine answer differences as a function of presented response format and age group. A significant main effect of response format, $F(1,171) = 33.23, p < .001$, $\eta_p^2 = .163$, showed that a higher number of sports activities was reported when a list of answer alternatives was provided ($M = 4.84$, SD = 3.31) compared to an open response format ($M = 2.69$, SD = 1.48). The biggest differences were found for sports activities that might also be categorized as leisure activities. For example, in an open response format only 9.2% of respondents reported skateboarding, while 28.4% picked this option when it was part of the response list. Reports differed most notably for the answer alternative *"School sports in general"*: 60.2% of respondents chose this in the closed response format compared to 10.3% in the open response format.

Concerning answer differences between younger and older adolescents, no significant interaction between response format and age group was found, $F(1,171) = 2.90, p = .091$, $\eta_p^2 = .017$. This indicates that the presented response format affected the answers of both younger (open response format: $M = 2.72$, SD = 1.49; closed response format: $M = 4.30$, SD = 3.01) and older adolescents (open response format: $M = 2.65$, SD = 1.50; closed response format: $M = 5.55$, SD = 3.58) to a similar degree. Thus no answer differences between the age groups were evident if the questions were related to sports activities, despite the fact that a very high number of response alternatives was presented in the closed response format (31 in total).

3.2. Frequency of Answer Scales

In order to analyze the influence of the presented answer scale and its frequency, answers from both questionnaire versions in each thematic section were coded on one dimension ranging from 1 – "*Several times a day*" to 7 – "*Less than once a month*". Within this combined scale, "*Several times a week*" and "*Once a week*" that appeared in both questionnaire versions were equally coded as well as "*Less than once a week*" from the high-frequency scale together with "*2–3 times a month*" from the low-frequency scale. In the section about environmental protection activities, the question that was either accompanied by a low- or a high-frequency scale was preceded by a filter question that asked whether participants sought out information about environmental protection issues. Approximately half of the respondents in each age group answered this filter question in the negative (younger adolescents: $n = 44$; older adolescents: $n = 51$). The presented questionnaire version did not affect the agreement with the filter question, neither in younger (high-frequency scale: $n = 30$; low-frequency scale: $n = 30$) nor in older adolescents (high-frequency scale: $n = 15$; low-frequency scale: $n = 18$), all $\chi^2 \leq 0.45$, all $p \geq .503$. Younger adolescents, however, agreed with the filter question more often than older adolescents overall, $\chi^2(1, n = 188) = 6.30$, $p = .012$. Due to the resulting unequal sample sizes and the ordinal nature of the responses, nonparametric tests were used in order to determine the impact of the respective answer scale frequency on the answers of the two age groups.

A Mann-Whitney U-Test on the reported behavioral frequencies confirmed a significant difference in the expected direction as a function of answer scale frequency, $U = 512.50$, $Z = 4.52$, $p < .001$, $r = .47$. Respondents reported informing themselves about once a week when the high-frequency scale was presented ($Mdn = 4$), whereas along the low-frequency answer scale they reported informing themselves only about 2–3 times a month on average ($Mdn = 5$). In other words, 53.4% of respondents reported that they sought information about environmental protection issues at least once a week when a high-frequency scale was presented, whereas only 27.1% gave an equivalent answer along the low-frequency answer scale. This confirms that an answer scale with high-frequency response alternatives resulted in higher estimates about behavioral frequencies than a low-frequency answer scale.

This answer difference was highly significant for younger adolescents, $U = 171.50$, $Z = 4.26$, $p < .001$, $r = .55$, whereas the difference among older adolescents only approached significance, $U = 85.50$, $Z = 1.86$, $p = .062$, $r = .22$ (see Table 3, upper panel). More specifically, 56.7% of the younger adolescents and 46.7% of the older adolescents reported that they informed themselves about environmental protection issues at least once a week when a high-frequency scale was presented. Only 20.0% of the younger adolescents and 38.9% of the older adolescents gave an equivalent answer along the low-frequency answer scale. This indicates that younger adolescents relied more than older adolescents on the frequency of the presented answer scale in order to estimate their behavioral frequency in this section of the questionnaire.

Within the section about sports activities, the relevant question in which the answer scale frequency was varied was preceded by the same filter question as the open versus closed response format question. As outlined above, only a small fraction of the respondents did not agree with this filter question (two additional respondents did not report their behavioral frequencies despite agreeing with the filter question). A Mann-Whitney U-Test on the reported behavioral frequencies as a function of presented answer scale did not show a significant difference between the two conditions, $U = 3290.00$, $Z = 1.52$, $p = .128$, $r = .12$. Respondents reported engaging in sports several times a week, regardless of which answer scale was presented (high-frequency scale: $Mdn = 3$; low-frequency scale: $Mdn = 3$). 90.7% of respondents reported doing sports at least once a week when a high-frequency answer scale was used, compared to 86.2% when low- frequency answer alternatives were presented.

No significant answer differences were found, neither for younger adolescents, $U = 1100.50$, $Z = 0.97$, $p = .331$, $r = .10$, nor for older adolescents, $U = 583.00$, $Z = 1.21$, $p = .225$, $r = .14$ (see Table 3, lower panel). Thus younger adolescents – much like older adolescents – seemed to estimate their behavioral frequencies relatively independently of the presented answer scale in this section of the questionnaire.

Table 3. Frequency of reported behaviors (in percent) as a function of answer scale for each thematic section and age group

High-frequency scale			Low-frequency scale		
	Younger adolescents	Older adolescents		Younger adolescents	Older adolescents
Environmental protection activities					
Several times a day	**0.0**	**0.0**	Several times a week	**6.7**	**11.1**
Daily	**20.0**	**6.7**	Once a week	**13.3**	**27.8**
Several times a week	**10.0**	**33.3**	2–3 times a month	30.0	27.8
Once a week	**26.7**	**6.7**	Once a month	33.3	16.7
Less than once a week	43.3	53.3	Less than once a month	16.7	16.7
Sports activities					
Several times a day	**6.1**	**5.4**	Several times a week	**74.0**	**59.5**
Daily	**18.4**	**8.1**	Once a week	**16.0**	**21.6**
Several times a week	**38.8**	**51.4**	2–3 times a month	6.0	10.8
Once a week	**32.7**	**18.9**	Once a month	4.0	8.1
Less than once a week	4.1	16.2	Less than once a month	0.0	0.0

Note. Within the section about environmental protection activities, answers refer to the question "*Do you seek inform yourself seek information about the environment and how to protect it?", if yes, "How often do you do that*?". For the sports activities section, results refer to the question "*Do you engage in sports?", if yes, "How often do you do that*?".

3.3. Question Order in Attitude Judgments

The reports about personal attitudes towards the respective topic in question were first analyzed by means of 2 (question order: specific-general vs. general-specific) × 2 (age group: younger vs. older adolescents) ANOVAs to examine the respondents' level of agreement with the statements depending on their position in the questionnaire. In addition, Pearson's product-moment correlations between the item pairs were calculated for each age group in order to determine to which extent their answers were related to each other in the different conditions. Fisher r-to-z transformations were used to compare the size of the correlations (two-tailed) as a function of question order (Preacher 2002).

In the section about environmental protection activities, the agreement with the attitude statements was not influenced by the presented question order (all $F \leq 3.85$, $p \geq .051$, $\eta_p^2 \leq .020$). However, for seven of the eight attitude statements (except the specific item of the second item pair), significant main effects of age group were obtained, irrespective of question order or item specificity, all $F \geq 4.25$, $p \leq .041$, $\eta_p^2 \geq .023$. An interaction between question order and age group was not found for any of the items, all $F \leq 2.90$, $p \geq .091$, $\eta_p^2 \leq .015$. Younger adolescents generally agreed more strongly than older adolescents with the attitude statements in this section of the questionnaire. This is particularly interesting in light of the behavioral reports implying that younger adolescents did not possess much direct experience of this topic on average.

The correlational analysis yielded the following results (see Table 4, upper panel): in younger adolescents, correlations between the answers in three of the four item pairs were significantly higher when the general item had to be evaluated first. For example, the items of the third item pair correlated with $r = .28$ when the specific item *"I can imagine well participating in projects that deal with environmental protection issues (e.g., volunteering)"* preceded the general item *"Environmental protection is an important topic to me"*. The correlation increased to $r = .63$ when the general item had to be evaluated first. These results show that younger adolescents adapted their answers on statements referring to a rather subordinate/personal aspect of the topic to the more general statements in most of the cases. This pattern was not found for the second item pair, possibly due to the respective question wording. The specific item of this item pair was reversed. In older adolescents, the direction of the relationship between the item pairs was not as consistent, and the differences between the correlations as a function of presented question order were not significant. Together, the results suggest that younger adolescents, in contrast to older adolescents, provided answers that were considerably influenced by the presented question context when the topic in question was rather unfamiliar to them.

In the section about sports activities, a 2 (question order: specific-general vs. general vs. specific) × 2 (age group: younger vs. older adolescents) ANOVA on the answers in each item block revealed a significant main effect of question order only for the items about sports in general, $F(1,184) = 7.43$, $p = .007$, $\eta_p^2 = .039$. Respondents agreed more strongly with these items when they were presented after the soccer items. In addition, a significant main effect of age group was obtained for the general statements, $F(1,184) = 6.95$, $p = .009$, $\eta_p^2 = .036$. In line with findings in the other section of the questionnaire, younger adolescents agreed more strongly with the general statements than older adolescents.

Table 4. Pearson's Product-Moment Correlation Coefficients of the attitude statements as a function of question order for each thematic section and age group

		Specific - general		General - specific			
		r	*n*	*r*	*n*	*z*	*p*
Environmental protection activities							
1st item pair	Younger adolescents	.39**	51	.73***	52	**2.55**	**.011**
	Older adolescents	.58***	42	.52***	42	0.38	.704
2nd item pair	Younger adolescents	−.15	52	−.21	51	0.31	.757
	Older adolescents	.02	42	−.19	42	0.94	.347
3rd item pair	Younger adolescents	.28*	52	.63***	52	**2.25**	**.024**
	Older adolescents	.61***	42	.36*	42	1.47	.142
4th item pair	Younger adolescents	−.03	52	.42**	52	**2.36**	**.009**
	Older adolescents	−.03	42	.00	42	0.13	.897
Sports activities							
Question blocks	Younger adolescents	.38**	52	.19	52	1.03	.303
	Older adolescents	−.03	42	.19	42	0.89	.327

Note. *** = $p < .001$; ** = $p < .01$; * = $p < .05$. The results of the Fisher r-to-z transformations are shown that compare the differences between the correlation coefficients.

The correlations between the item blocks about soccer and sports in general were rather small and did not differ significantly as a function of question order, neither in younger nor in older adolescents (see Table 4, lower panel). This suggests that both age groups answered these statements fairly independently of the presented question order.

3.4. *Order of Filter Questions in Combination with Attitude Judgments*

To examine the impact of answering or failing to answer filter questions asking about actual engagement in environmental protection activities depending on their position in the questionnaire, only those respondents who answered the two filter questions consistently (i.e., agreed or disagreed with both filter questions) were included in the analysis. In addition, answers on each of the eight attitude statements were averaged into a composite judgment score ranging from 1 – *"I strongly disagree"* to 5 – *"I strongly agree"*. In general, respondents agreed more often with the filter questions when they preceded the attitude statements (agreed: $n = 41$; disagreed: $n = 20$) compared to the reversed question order (agreed: $n = 28$; disagreed: $n = 37$), $\chi^2(1,\ n = 126) = 7.40$, $p = .007$. This seemed to be mainly due to the younger adolescents, who agreed more often with the filter questions when they were presented in first (agreed: $n = 27$; disagreed: $n = 8$), but did not show any answer preference if they had already reported their attitudes about the topic in question (agreed: $n = 18$; disagreed: $n = 18$), $\chi^2(1,\ n = 71) = 5.63$, $p = .018$. This might have been related to the respective response format (i.e., closed response format and a high-frequency answer scale) that was used for the two filter questions in this questionnaire version. Older adolescents, in contrast, appeared to be less affected by the presented order of the filter questions (agreed before: $n = 14$; disagreed before: $n = 12$; agreed afterwards: $n = 10$; disagreed afterwards: $n = 19$), $\chi^2(1,\ n = 55) = 2.09$, $p = .148$. Due to the unequal sample sizes in the different conditions, nonparametric tests were used in order to determine the relation between answering or failing to answer the filter questions and the attitude reports.

A Mann-Whitney U-Test confirmed that subsequent attitude reports were influenced by answering or failing to answer the filter questions beforehand, $U = 168.00$, $Z = 3.73$, $p < .001$, $r = .48$. Respondents who failed to answer the filter questions successfully beforehand evaluated the attitude statements more neutrally ($Mdn = 3.00$) than respondents who were able to report some relevant behaviors ($Mdn = 3.38$). This result indicates that information about own behaviors that became accessible through these filter questions shaped the following attitude judgments accordingly. Moreover, attitude reports were less positive when the filter questions were denied beforehand than when the attitude statements preceded the filter questions ($Mdn = 3.50$), $U = 337.00$, $Z = 3.25$, $p = .001$, $r = .35$. No significant answer differences were found between respondents who agreed with the filter questions beforehand and those who answered them afterwards, $U = 1300.00$, $Z = 0.21$, $p = .832$, $r = .02$.

When analyzed separately for each age group, the difference between attitude reports as a function of answering or failing to answer the filter questions beforehand was significant for younger adolescents, $U = 48.00$, $Z = 2.38$, $p = .017$, $r = .40$, as well as for older adolescents, $U = 40.50$, $Z = 2.25$, $p = .025$, $r = .44$. Younger adolescents' ($Mdn = 3.06$) as well as older adolescents' ($Mdn = 2.94$) attitude reports were less positive when they failed to answer the filter questions successfully compared to the situation when they were able to report some relevant behaviors beforehand (younger adolescents: $Mdn = 3.38$; older adolescents: $Mdn = 3.31$). The attitude reports of younger adolescents who answered the filter questions afterwards ($Mdn = 3.63$) were more positive than the reports of younger adolescents who failed to answer the filter questions beforehand, $U = 60.50$, $Z = 2.55$, $p = .011$, $r = .39$, but did not differ from the reports of younger adolescents who were able to answer them before, $U = 427.50$, $Z = 0.82$, $p = .414$, $r = .10$. In older adolescents, attitude reports did not differ in respondents who answered the filter question positively or negatively in the first place compared to respondents who answered them afterwards ($Mdn = 3.25$), all $U \geq 117.00$, $Z \leq 1.64$, $p \geq .102$, $r \leq .26$. Thus, younger adolescents' attitude reports seemed to be more strongly influenced by the presented question order and the information brought to mind by the failure to report any relevant behaviors.

Corresponding analyses in the section about sports activities were not conducted because only 6 of 94 respondents gave a negative answer to the filter questions about their engagement in sports activities when these questions were presented first.

4. Discussion

This is one of the first studies investigating the impact of question format and question context on self-reports of younger and older adolescents within different topics. Results show that contextually provided cues related to certain question characteristics, response scales, and question order may exert a considerable influence on self-reports of these age groups. Additionally, our findings indicate that in most cases, the answers of adolescents aged between 11 and 13 years were more strongly affected by the presented question format and context than those of adolescents aged between 16 and 18 years. However, response effects were generally less pronounced when the questions referred to sports activities, a topic of which the majority of respondents had direct experience, but were stronger in the section about environmental protection activities, a topic that seemed to be more "remote" from the respondents' field of experience. As a consequence, answer differences between younger and older adolescents were mainly found in the section about environmental protection activities.

4.1. The Impact of Question Format and Context

A closed response format resulted in more and different answers compared to an open response format. As expected, generating answers in an open response format appeared to be more difficult than picking an answer from a list of answer alternatives, especially for younger adolescents in the section about environmental protection activities. Results further implied that the respondents were more likely to skip this question completely in this section of the questionnaire if an open response format was presented. Thus, if possible in the context of the questionnaire, younger survey respondents might avoid questions that require more effort in order to generate an appropriate answer (cf. Borgers et al. 2000). In these cases, they may use a satisficing strategy and rely more strongly on contextual cues provided by the questionnaire (cf. De Leeuw et al. 2004). This is further supported by the results in the closed response format suggesting that younger adolescents are more likely than older adolescents to check a number of activities that appear meaningful to them. In the section about sports activities, reports differed most notably for the answer alternative *"School sports in general"*. In line with previous research, even younger survey respondents appeared to exclude activities that the researcher may take for granted as long as this was not explicitly requested within the presented question format (cf. Schwarz 1999).

In addition, higher behavioral frequencies were reported when a high-frequency answer scale was presented as compared to a low-frequency answer scale in the section about environmental protection activities. Respondents seemed to use the presented answer scale as a frame of reference for estimating the "typical" behavioral frequency within the topic in question and classifying their own activities accordingly. This pattern was more evident in younger adolescents, who might have had difficulties in remembering relevant instances of somewhat infrequent and poorly represented behavior due to their less developed cognitive skills (cf. Fuchs 2005). However, one might also argue that younger adolescents tried to make a positive impression and more than older adolescents tended towards answers they considered socially desirable in this section of the questionnaire (cf. De Leeuw et al. 2004; Kränzl-Nagl and Wilk 2000; Reynolds 1993). In addition, younger adolescents agreed with the preceding filter question more often than older adolescents in both questionnaire versions. This might have been related to the general question format, which included a list of answer alternatives in both cases, or, in other words, contextual cues that could be used to generate an answer. Younger adolescents might rely more on such cues in order to infer the intended meaning of a given question and to form an appropriate answer in a survey.

We also found that younger but not older adolescents judged statements targeting personal opinions and attitudes differently as a function of presented question order in the section about environmental protection activities. In line with the norm of evenhandedness (Schuman and Presser 1981), they seemed to adjust their personal preference to normative expectations that might have been implied by the general items (containing phrases like "*everybody has*" or "*people behave*"). Thus, a lack of a stable personal attitude about the topic in question might have led to an orientation towards the general statements to infer the "appropriate" attitude in the presented context and to construct their own personal attitude accordingly. Younger adolescents generally agreed more strongly with the attitude statements than older adolescents. Whether this might be due to satisficing or whether they simply possessed a more positive attitude towards these topics cannot be answered completely within the present study. It appears more likely, however, that this might be linked to the former due to a stronger tendency to behave in a socially desirable way, especially when information from authorities is received (cf. Ceci et al. 1987; Kränzl-Nagl and Wilk 2000; Reynolds 1993; Scott 2008; Warren and Lane 1995). The results further suggested that younger adolescents had some problems with answering an item that was reversed (i.e., requiring disagreement in order to express a positive attitude) in contrast to the other items (cf. Borgers et al. 2004).

In addition, we showed that the attitude statements were judged differently depending on whether they were presented before or after certain filter questions about personal behaviors and answering or and failing to answer these questions. The awareness that no relevant behaviors could be reported obviously shaped the subsequent attitude statements. This was especially evident in younger adolescents. This age group also showed a stronger tendency to agree with the filter questions when they were presented first, in contrast to the reversed question order. Given that both filter questions incorporated a list of answer alternatives when they preceded the attitude statements, one might speculate that a list of answer alternatives encouraged younger more than older adolescents to agree with the filter questions. If they were not able to answer them despite these contextual cues, however, their agreement with the following attitude statements decreased considerably. Thus younger survey respondents might try to convey a consistent picture and reaffirm their self-concept after certain information becomes accessible in the context of a survey (cf. Martin and Harlow 1992).

In sum, the results on the influence of question order on attitude reports speak against the assumption that younger survey respondents may not recognize preceding questions as a relevant context for the interpretation of following questions due to limitations in cognitive resources (cf. Fuchs 2005). As Fuchs (2009) noted, age-related differences in self-reports cannot be completely accounted for by differences in cognitive status. According to Scott (2008), children's performance in memory tasks is similar to that of adults by the age of eleven, suggesting that the younger adolescents in the present study were able to keep in mind the information that became accessible due to preceding questions. Instead, problems of literacy, confidentiality, and context seem to play an important role when conducting survey research with these age groups (Scott 2008). The participation in a scientific survey conducted in the classroom might already have shaped their interpretation of the presented questions and their perceptions about social desirability.

One limitation of the study concerns the general questionnaire design that varied four response effects within two versions of a questionnaire. For example, some of the behavioral questions could be skipped when a filter question was answered negatively beforehand (e.g., *"Are you actively involved in saving the environment?"*). Given that these skipped questions varied another effect (e.g., open vs. closed response format), the sample size was sometimes considerably reduced. Moreover, the usage of paper-and-pencil questionnaires enabled respondents to see forthcoming questions. The results indicate that the characteristics of these questions might bias responses to preceding questions in order to avoid questions that require more effort to answer (e.g., when presented in an open response format). In survey research, however, many topics are often incorporated within as few questions as possible, resulting in questionnaires similar to those used in the present study. Therefore, it may be considered an important step to elucidate possible effects of the presented question characteristics on the answers of younger respondents in a typical research survey.

4.2. The Impact of Question Content

Younger adolescents were more influenced by the presented question format and context than older adolescents for the majority of the tested effects in the section about environmental protection activities, but not in the section about sports activities. This suggests once more that possible differences in cognitive development are unlikely to account for the present results, also given that the sample was rather homogenous in terms of academic performance. Instead, younger adolescents might be more susceptible to these kinds of effects compared to older adolescents when they are not particularly experienced with the topic in question. In these cases, suggestibility resulting from social and motivational factors might bias their responses. For the majority of respondents from both age groups, sports activities seemed to play an important role in their lives, whereas direct experiences with environmental protection activities appeared to be much less prevalent. In line with the findings from Bickart (1992), this indicates that frequent and personally relevant behaviors are better represented in memory, which enables even younger respondents to retrieve relevant information from memory rather than construct it based on contextual cues. By contrast, if experience with the question content is not given, younger age groups in particular may be highly context sensitive, susceptible towards normative expectations, and tend towards answers they may perceive as socially desirable (cf. Borgers et al. 2000; Kränzl-Nagl and Wilk 2000; Reynolds 1993). One should note, however, that the very high frequency of relevant behaviors in the section about sports activities may have attenuated possible response effects to some extent. Thus further research is warranted in future in order to examine the influence of these response effects within topics that are regularly encountered in these age groups but differ in their degree of social desirability (however, see Fuchs 2005, who found large response scale effects in younger adolescents aged 13 to 15 when asking them to report their daily TV consumption). Due to differences between the questions in the two sections of the questionnaire that resulted from the fact that they were designed to capture the characteristics of each topic in an ecologically valid way, direct comparisons between the topics could not be implemented in the present study. Nevertheless, it still provides important insights into the way younger respondents deal with questions asked in a survey and may have important applications for improving the design of questionnaires targeted at younger age groups and the interpretation of their results.

4.3. Implications for Survey Research With Younger Age Groups

The results of the present study confirm the assumptions of Lipski (2000) and Scott (2008) that for questions related to their direct field of experience, younger age groups are capable of giving answers that are less influenced by the presented question format and context. It should be noted, however, that younger respondents are likely to be "novices" in many domains. Because they are still in the process of acquiring knowledge and forming firm attitudes, they might be more susceptible to context effects on many occasions. Researchers are well advised to consider this when developing and conducting a survey in order to minimize erroneous conclusions from the obtained results. Systematic pretesting could provide the researcher with an informative basis about the way young age groups think, their level of experience with the topic, and their understanding of the question (see De Leeuw et al. 2004 and Scott 2008 for different methods). Essential cues can then be derived about the most suitable design for a particular question, for example, when to favor an open over a closed response format. One further possibility might be the usage of computer-assisted or online surveys in which only one question is visible at a time to reduce potential negative effects of open response formats on the willingness to answer a question at all. Additionally, self-reports should be validated on actual behavior or by employing multiple methods, if possible. Examining bigger samples, comparing groups which differ to a larger extent in age and developmental status, as well as testing the effects within a wider range of topics, are therefore important avenues for future research.

5. References

Bickart, B.A. 1992. "Question-Order Effects and Brand Evaluations: The Moderating Role of Consumer Knowledge." In *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. 63–80. New York: Springer.

Bishop, G.F., R.W. Oldendick, and A.J. Tuchfarber. 1983. "Effects of Filter Questions in Public Opinion Surveys." *Public Opinion Quarterly* 47: 528–546. Doi: http://dx.doi.org/10.1086/268810.

Bishop, G.F. 1987. "Context Effects on Self-Perceptions of Interest in Government and Public Affairs." In *Social Information Processing and Survey Methodology*, edited by H.-J. Hippler, N. Schwarz, and S. Sudman. 179–199. New York: Springer.

Borgers, N., E. de Leeuw, and J. Hox. 2000. "Children as Respondents in Survey Research: Cognitive Development and Response Quality." *Bulletin de Méthodologie Sociologique* 66: 60–75. Doi: http://dx.doi.org/10.1177/075910630006600106.

Borgers, N., D. Sikkel, and J. Hox. 2004. "Response Effects in Surveys on Children and Adolescents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point." *Quality & Quantity* 38: 17–33. Doi: http://dx.doi.org/10.1023/B:QUQU.0000013236.29205.a6.

Ceci, S.J., D.F. Ross, and M.P. Toglia. 1987. "Suggestibility of Children's Memory: Psycholegal Implications." *Journal of Experimental Psychology: General* 116: 38–49. Doi: http://dx.doi.org/10.1037/0096-3445.116.1.38.

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

Fallon, Jr., T. and M. Schwab-Stone. 1994. "Determinants of Reliability in Psychiatric Surveys of Children Aged 6–12." *Journal of Child Psychology and Psychiatry* 35: 1391–1408. Doi: http://dx.doi.org/10.1111/j.1469-7610.1994.tb01282.x.

Festinger, L. 1954. "A Theory of Social Comparison Processes." *Human Relations* 7: 117–140. Doi: http://dx.doi.org/10.1177/001872675400700202.

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

Fuchs, M. 2009. "The Reliability of Children's Survey Responses: The Impact of Cognitive Functioning on Respondent Behavior." In Proceedings of Statistics Canada Symposium 2008. Data Collection: Challenges, Achievements and New Directions. Ottawa: StatCan. Available at: http://www.statcan.gc.ca/pub/11-522-x/2008000/article/10961-eng.pdf (accessed June 2013).

Gawronski, B. and J. Cesario. 2013. "Of Mice and Men: What Animal Research Can Tell Us About Context Effects on Automatic Responses in Humans." *Personality and Social Psychology Review* 17: 187–215. Doi: http://dx.doi.org/10.1177/1088868313480096.

Goswami, U. 2010. *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, 2nd ed. Oxford: Wiley-Blackwell.

Heinzel, F. 2000. "Methoden und Zugänge der Kindheitsforschung im Überblick." In *Methoden der Kindheitsforschung: Ein Überblick über Forschungszugänge zur kindlichen Perspektive*, edited by F. Heinzel. 21–35. Weinheim: Juventa.

Kane, E. and H. Schuman. 1991. "Open Survey Questions as Measures of Personal Concern with Issues: A Reanalysis of Stouffer's Communism, Conformity, and Civil Liberties." In *Sociological Methodology*, edited by P. Marsden. 81–96. Oxford: Basil Blackwell.

Knäuper, B., N. Schwarz, D. Park, and A. Fritsch. 2007. "The Perils of Interpreting Age Differences in Attitude Reports: Question Order Effects Decrease with Age." *Journal of Official Statistics* 23: 515–528.

Kränzl-Nagl, R. and L. Wilk. 2000. "Möglichkeiten und Grenzen standardisierter Befragungen unter besonderer Berücksichtigung der Faktoren soziale und personale Wünschbarkeit." In *Methoden der Kindheitsforschung: Ein Überblick über Forschungszugänge zur kindlichen Perspektive*, edited by F. Heinzel. 59–75. Weinheim: Juventa.

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

Krosnick, J.A. 1999. "Survey Research." *Annual Review of Psychology* 50: 537–567. Doi: http://dx.doi.org/10.1146/annurev.psych.50.1.537.

Krosnick, J.A. and D.F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects." *Public Opinion Quarterly* 51: 201–219. DOI: http://dx.doi.org/10.1086/269029.

Lipski, J. 2000. "Zur Verlässlichkeit der Angaben von Kindern bei standardisierten Befragungen." In *Methoden der Kindheitsforschung: Ein Überblick über*

Forschungszugänge zur kindlichen Perspektive, edited by F. Heinzel. 77–86. Weinheim: Juventa.

Livingstone, S., L. Haddon, and A. Gorzig. 2012. *Children, Risk and Safety Online: Research and Policy Challenges in Comparative Perspective*. Bristol: The Policy Press.

Martin, L.L. and T.F. Harlow. 1992. "Basking and Brooding: The Motivating Effects of Filter Questions in Surveys." In *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. 81–96. New York: Springer.

McNeal, J.U. 1992. *Kids as Customers: A Handbook of Marketing to Children*. New York: Lexington Books.

Preacher, K.J. 2002. Calculation for the Test of the Difference Between Two Independent Correlation Coefficients [Computer software]. Available at: http://quantpsy.org. (accessed June 2013).

Reynolds, W.M. 1993. "Self Report Methodology." In *Handbook of Child and Adolescent Assessment*, edited by T.H. Ollendick and M. Hersen. 98–120. Boston: Allyn & Bacon.

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

Schwarz, N. 1999. "Self-Reports: How the Questions Shape the Answers." *American Psychologist* 54: 93–105. Doi: http://dx.doi.org/10.1037/0003-066X.54.2.93.

Schwarz, N. 2003. "Self-Reports in Consumer Research: The Challenge of Comparing Cohorts and Cultures." *Journal of Consumer Research* 29: 588–594. Doi: http://dx.doi.org/10.1086/346253.

Schwarz, N. and J. Bienias. 1990. "What Mediates the Impact of Response Alternatives on Frequency Reports of Mundane Behaviors?" *Applied Cognitive Psychology* 4: 61–72. Doi: http://dx.doi.org/10.1002/acp.2350040106.

Schwarz, N., H. Bless, F. Strack, G. Klumpp, H. Rittenauer-Schatka, and A. Simons. 1991. "Ease of Retrieval as Information: Another Look at the Availability Heuristic." *Journal of Personality and Social Psychology* 61: 195–202. Doi: http://dx.doi.org/10.1037/0022-3514.61.2.195.

Schwarz, N. and G. Bohner. 2001. "The Construction of Attitudes." In *Blackwell Handbook of Social Psychology: Intraindividual Processes*, edited by A. Tesser and N. Schwarz. 436–457. Oxford: Blackwell.

Schwarz, N. and H.J. Hippler. 1995. "Subsequent Questions May Influence Answers to Preceding Questions in Mail Surveys." *Public Opinion Quarterly* 59: 93–97. Doi: http://dx.doi.org/10.1086/269460.

Schwarz, N., H.J. Hippler, B. Deutsch, and F. Strack. 1985. "Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments." *Public Opinion Quarterly* 49: 388–395. Doi: http://dx.doi.org/10.1086/268936.

Schwarz, N. and D. Oyserman. 2001. "Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction." *American Journal of Evaluation* 22: 127–161. Doi: http://dx.doi.org/10.1177/109821400102200202.

Scott, J. 2008. "Children as Respondents: The Challenge for Quantitative Methods." In *Research with Children: Perspectives and Practices*, edited by P. Christensen and A. James. 87–108. London: Routledge.

Strack, F. 1992. "'Order Effects' in Survey Research: Activation and Information Functions of Preceding Questions." In *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman. 23–34. New York: Springer.

Strack, F. 1994. *Zur Psychologie der standardisierten Befragung: Kognitive und Kommunikative Prozesse*. Berlin: Springer.

Strack, F. and N. Schwarz. 2007. "Asking Questions: Measurement in the Social Sciences." In *Psychology's Territories: Historical and Contemporary Perspectives from Different Disciplines*, edited by M. Ash and T. Sturm. 225–250. Mahwah, NJ: Erlbaum.

Tourangeau, R. and K.A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103: 299–314.

Tourangeau, R., K.A. Rasinski, N. Bradburn, and R. D'Andrade. 1989. "Carryover Effects in Attitude Surveys." *Public Opinion Quarterly* 53: 495–524. Doi: http://dx.doi.org/10.1086/269169.

Warren, A.R. and P. Lane. 1995. "Effects of Timing and Type of Questioning on Eyewitness Accuracy and Suggestibility." In *Memory and Testimony in the Child Witness*, edited by M.S. Zaragoza, J.R. Graham, G.C.N. Hall, R. Hirschman and Y.S. Ben-Porath. 44–60. Thousand Oaks: Sage Publications.

Received April 2014
Revised April 2015
Accepted August 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 329–348, http://dx.doi.org/10.1515/JOS-2016-0019

End User Licence to Open Government Data? A Simulated Penetration Attack on Two Social Survey Datasets

Mark Elliot[1], *Elaine Mackey*[2], *Susan O'Shea*[3], *Caroline Tudor*[4], *and Keith Spicer*[5]

In the UK, the transparency agenda is forcing data stewardship organisations to review their dissemination policies and to consider whether to release data that is currently only available to a restricted community of researchers under licence as open data. Here we describe the results of a study providing evidence about the risks of such an approach via a simulated attack on two social survey datasets. This is also the first systematic attempt to simulate a jigsaw identification attack (one using a mashup of multiple data sources) on an anonymised dataset. The information that we draw on is collected from multiple online data sources and purchasable commercial data. The results indicate that such an attack against anonymised end user licence (EUL) datasets, if converted into open datasets, is possible and therefore we would recommend that penetration tests should be factored into any decision to make datasets (that are about people) open.

1. Introduction

The UK's Office for National Statistics (ONS) currently disseminates large numbers of datasets under end user licence (EUL). This is a restricted dissemination of the data to researchers who agree to a set of sixteen licence conditions and specifically agree not to attempt to reidentify individuals. Under the transparency agenda, ONS has considered whether some of these could be released under an Open Government Data licence. This is effectively unrestricted publication on the Internet. This is clearly a very different level of dissemination and therefore careful conceptual and disclosure risk analyses was necessary in order to understand the marginal increase in disclosure risk (if any) associated with this change in dissemination policy.

The work took place in two phases. During Phase 1 we considered the interplay of legal and statistical definitions of confidentiality, developing a detailed understanding of the differences in the licences and associated documents. This was essentially a socio-legal piece of work, which in turn allowed us to generate a set of feasible scenarios that extended beyond the orthodox intruder scenarios. Orthodox scenarios come in two basic forms: (i) *database cross match*, where an intruder attempts to link records in an *identification file* and a de-identified *target file* but does not know for certain for any given record in the identification file whether there is any corresponding record in the

[1,2,3] School of Social Sciences, University of Manchester, Manchester M13 9PL, UK. Emails: mark.elliot@manchester.ac.uk, laine.mackey@manchester.ac.uk, and susan.o'shea@manchester.ac.uk.
[4,5] Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hampshire, PO15 5RR, UK. Emails: caroline.tudor@ons.gsi.gov.uk and keith.spicer@ons.gsi.gov.uk.

target database, and (ii) *fishing*, where the intruder selects records from the target database and attempts to find the corresponding person in the population. There are other variants – see Elliot and Dale (1999) for a discussion – however, the critical point is that response knowledge is not assumed. In general, it is held that, with EUL licenses, the potential costs to a researcher of attempting reidentification (e.g., career damage) outweigh the benefits of doing so. Therefore, even though it is possible that a researcher might know that a third party is in the data, such intrusions are unlikely. Under the Open Government Data (OGD) licence, the mere act of identification would not break any rules and therefore the costs of such reidentification are simply the effort required to carry it out. On top of this, the fact that OGD means effectively universal access implies the very strong possibility that somebody exists who would know that some other person was in the data (and probably for many respondents there would be somebody with such knowledge). It is generally accepted that with response knowledge, reidentification is considerably easier (i.e., the effort required is considerably less). The combination of these factors makes response knowledge scenarios far more plausible with OGD.

After a review of the report on Phase 1, ONS commissioned a Phase 2 study: a simulated attack based on an intruder who had response knowledge that an individual was in a dataset and then used publically available information (either openly available or available for a fee) in order to identify that individual in the dataset. For this stage, which we report upon here, two UK datasets were focused on: the Labour Force Survey (LFS) and the Living Costs and Food Survey (LCF). These are both microdata samples that are smaller than one percent of the UK population and contain information on individuals and their households. Some disclosure control has been applied, such as banding age and ethnic group. This study builds on previous attack simulations (e.g., Müller et al. 1995; Elliot 2009) but adds an additional step of trawling for and combining available public information (rather than simply matching two fixed datasets).

This article consists of the following sections. Section 2 reviews the existing literature on reidentification. Section 3 summarises the Phase 1 study, which describes the motivation for the attack scenario. Section 4 describes the methodological approach to the penetration test. Sections 5 and 6 describe the matching process and the results of the consequential reidentification attempts. Section 7 is the general discussion.

2. Review of Reidentification Studies

Reidentification studies come in three different forms: (i) defensive studies carried out by or on behalf of data custodians, often called penetration tests, where the goal is to assess disclosure risk associated with a proposed data release; (ii) academic studies exploring new attack forms or new potential anonymisation techniques; and (iii) demonstrative studies usually carried out by data journalists and/or academics, where the point of the study is to demonstrate that a given release is unsafe.

The earlier studies were largely of the second type. Müller et al. (1995) tested whether it was possible to link records in the 1987 German microcensus file to an administrative register. Results varied depending on the scenario assumed, but in general they concluded that “although identification is not impossible, only under special circumstances are the chances of a successful identification larger than virtually zero"; p.149. Similarly, Elliot and Dale (1998) showed through a study linking UK census microdata to a sample survey that it was possible to reidentify some people by cross matching databases but that the correct matches were effectively hidden amongst many false positive matches. Later Elliot (2009) demonstrated that by focusing on unusual records (using the so-called fishing attack) it was possible to achieve a higher hit rate, but he also found that the anonymisation methods that ONS had employed in the test file (2001 census microdata) did effectively stymie that attack.

El Emam et al. (2011) carried out a systematic review of reidentification attacks on health data to (i) compute the overall proportion of correctly identified records, and (ii) assess whether it indicated weakness in current anonymisation methods as used with health data. On average, approximately a quarter of the records were reidentified across all studies. They concluded the evidence showed a high reidentification rate, but that this was mostly based on small-scale studies on data that were not anonymised according to existing standards. This evidence is insufficient to draw general conclusions about the efficacy of anonymisation methods.

Recent academic work has focused on new forms of data. For example, genomics data (Malin and Sweeney 2004; Gymrek et al. 2013) and social network data (Backstrom et al. 2007; Narayanan and Shmatikov 2009) have both come under the spotlight; the general conclusion drawn is that the more complex the form of data, the more vulnerable those data are to reidentification attacks.

The practical importance of these studies has been to show that care is required before data is released, particularly if it is to be released as open data. Examples where such care has not been exercised have led to the third (demonstrative) type of reidentification study.

A particularly notorious example of a demonstrative study arose from the release of a database of supposedly anonymised movie ratings by Netflix. The data were released in an attempt to improve its movie recommendations algorithm through crowdsourcing the problem, offering a $1 million prize for the best solution. For each case, a unique subscriber ID, the movie title, year of release and the date in which the subscriber rated the movie were given.

Narayanan and Shmatikov (2008) showed in their study how Netflix users could be reidentified. They were able to identify (some) users by matching their Netflix reviews with data from other sites like IMDb (http://www.imdb.com accessed 14/7/15). Furthermore, they found that if you knew a few movies a Netflix subscriber had rented in a given time period, you could reverse engineer the data and find out the rest of their viewing history. They concluded that very little auxiliary information is needed to de-anonymise an average subscriber record from the Netflix Prize dataset. With eight movie ratings (of which two may be completely wrong) and dates that may have a 14-day error, 99% of records can be uniquely identified in the dataset. For a 68% hit rate, two ratings and dates (with a three-day error) are sufficient.

The Netflix example and similar demonstrative studies involving AOL search data (in 2006), the New York taxi cab dataset (in 2014) and Transport for London bike journey data (in 2014) demonstrate the difficulties of releasing datasets that have not been thoroughly tested for reidentification risk as open data, and in particular they demonstrate the value of defensive reidentification tests.

3. Motivation for the Response Knowledge Scenario

The initial focus of the Phase 1 work was to consider a range of different materials including:

1) The Data Protection Act (1998).
2) The OGD licence.
3) The EUL licence.
4) A document provided by ONS detailing their view of the differences between the OGD and EUL licences.
5) The Anonymisation Code of Practice produced by the UK Information Commissioner's Office (ICO).
6) The standard confidentiality pledge provided by ONS to respondents.
7) The UK Government Statistica Service (GSS) Disclosure Control policy for Microdata Produced from Social Surveys.

Analysis focused on differences in the data environment in which data would exist under the two different licence forms. This has been embedded in developments of our thinking about both the relationship between statistical and legal confidentiality and the conceptualisation of the data environment (see Mackey and Elliot 2013; Elliot and Mackey 2014).

The analysis presented here is somewhat different from an orthodox disclosure risk analysis. During this phase we were trying to build a well-grounded description of the problem, its attributes and the likely and plausible consequences of a decision to change the licensing for the current EUL datasets.

3.1. Understanding the Differences Between the Licences

There are considerable differences between the EUL and OGD licences. An analysis of the licences led us to the conclusion that the following differences impact on the disclosure risk either directly or indirectly.

Restrictions on use: Clause 2 of the standard EUL provides a fairly tight definition of how the data may be processed. The OGD licence provides no restriction. As we shall see later, this is a critical factor in creating new disclosure scenarios.

Restrictions on sharing: EUL Clause 6 restricts sharing to other EUL holders (which in effect means that data can only be shared with those who already have access to it). The OGD licence (of course) places no restrictions on sharing.

Preservation of confidentiality: EUL Clause 8 imposes a specific responsibility on users to preserve respondent confidentiality. It also specifically prohibits deliberate statistical disclosure. The OGD licence does not provide any such responsibility, relying only on the Data Protection Act. The OGD license does refer to compliance with the European Directive 2002/58 on Privacy and Electronic Communications. However, this Directive does not concern individual data of the type that is in question here and therefore it is ignored henceforth.

It should be noted that the EUL does not explicitly prohibit identification (of oneself or others). However, it is hard to construct a reasonable use case where identification (of oneself or others) does not breach Clause 2 and it is hard to construct a plausible scenario of identification of others which does not breach Clause 8. We consider therefore that identification is, for practical purposes, implicitly prohibited by the EUL.

3.2. Key Points of Interpretation of the Data Protection Act

The Data Protection Act (DPA) concerns personal data. In general, the processes of anonymisation and statistical disclosure control are designed to render the data non-personal. Personal data are defined in the DPA as data which relate to a living individual who can be identified either:

a. from that data, or
b. from that data and other information which is in the possession of, or is likely to come into the possession of, the data controller.

De-identified data – where the formal identifiers have been removed or masked – is no longer *identified* but may still be *identifiable*. The first clause clearly does not apply to de-identified data and therefore Clause b is our concern. *The key point about Clause b is that it is contextual. One cannot make judgements about whether data is personal or not simply by considering the data itself.* Whether the data is personal or not will depend upon the environment in which it resides. Each data environment has attributes which affect the personal/non-personal nature of the individual-level data contained within it. These include:

- Other data. It is explicit in the definition of "personal data" that other data is relevant.
- Data users. Data users have identifying knowledge of other individuals. They also move between data environments and carry information with them as they do so. Users have varying levels of expertise that make them more or less able to carry out the necessary data processes to enact identification.
- Data security. The better the security with which data is kept, the stronger the partition between the data environment and other data environments.
- Governance structures and processes (including licences). Governance processes create expectations about what may be reasonably done with the data. Some processing of data will make it more likely that the data will become personal.
- The intra-environment ethos. The prevailing ethos within a data environment will affect the practice of interacting with data. Behaviour and attitudes which are not necessarily precisely specified in licences come into play here. Does the prevailing ethos specify that one should *look after* data?

It is fairly clear that for all of the above attributes, an OGD licence increases the likelihood that a given dataset which relates to living individuals will be regarded as personal data because the probability of identification will be higher:

- Other data. The OG data effectively exist in the global data environment and therefore in principle could be linked to any other data. Under EUL the data could only possibly be linked to data that the researchers who are using it have access to.
- Data users: The user base will increase massively under an OGD licence compared to the EUL licence (this is the point of OGD).

- Data security. By definition there is no security with OG data – the data is open. EUL data exist in relatively controlled research environments (although these are not high security).
- Governance structures and processes (including licences). With the OGD licence, the data is unregulated as the licence is deliberately permissive; there are very few restrictions on data processes. The EUL places restrictions on what a researcher may or may not do.
- The intra-environment ethos. In the global data environment, it is probably fair to say that there is no one coherent ethos and the full range of behaviours in respect of data can be expected. In the EUL environment there is a prevailing expectation that researchers will look after data.

In essence, it is thus clear from this surface analysis that for survey microdata the OGD licence can only increase the risk that data is personal, which leaves us with the question of: *by how much*?

3.3. The OGD User Base

The overarching principle of OG data is to increase the accessibility of government data. Increased accessibility means a larger user base; larger in number with a greater diversity in types of users. Ignoring any other impacts of the OGD licence, this must increase the risk of a disclosure event. Assuming the risk is non-zero with the population of EUL users and assuming that all users are equal in terms of their risk impact, then the probability of an attempt will increase in proportion to the increase in the size of the user base. It is certainly open to discussion whether the OGD user groups are more or less risky than the EUL ones but, given that the latter is essentially a subset of the former, it is difficult to argue against the proposition that the risk would increase with the increase in size of the user base.

If this were the only problem then the increase in risk attributable to users could be managed through the orthodox trade-off mechanism – by applying more stringent disclosure control to the data. However, the problem is far more difficult than this. The OGD licence changes the nature of the disclosure risk problem, creating a whole new type of disclosure scenario.

3.4. Open Government Data Disclosure Scenarios

Work on attack scenarios for survey data includes Paass (1988) and Elliot and Dale (1999). Orthodox scenarios include *database cross match* (where two databases are linked) and *fishing* (where the intruder identifiers outliers in the data and then attempts to find those individuals in the population). When considering what additional disclosure scenarios are in scope under OGD but are not in scope under EUL, we are not concerned with technical possibility. Technically, there is nothing that a user could not do under EUL that one could do under the OGD licence. However, the licences do create formal (quasi-legal) restrictions on activity and this restriction significantly affects the shape of the data environment.

We considered six different scenarios in our analysis at this stage:

1. Self-identification
2. Spontaneous recognition
3. Spontaneous recognition augmented by subsequent response knowledge
4. Commercial data augmentation
5. Response knowledge with collusion
6. Response knowledge without collusion

There is not sufficient space to consider all of these here; suffice it to say that we considered the sixth to be the most problematic.

3.4.1. Response Knowledge of Others

Response knowledge attacks are where users who know that a particular respondent is in the OGD dataset and also have other knowledge about them use the combination of that knowledge to identify the respondent. Here we are considering a data environment that is effectively just the user themselves. In this environment and scenario, the respondent's data is personal for the user, because they have data that allow them to identify the respondent.

Given this, there is a question of whether the identification process itself constitutes a breach of the DPA. Specifically, can a user be deemed to be unfairly processing data by identifying a record in a dataset? Such an identification process does not change the status of the record concerned – it was personal and it still is. The only difference is that the user now knows precisely which record is the data for the respondent. They may not have even learnt anything new about the respondent. The scope of what constitutes "processing" within the DPA is broad and includes "alignment" and "combination", which identification processes could be said to constitute, but it is unclear about whether identification in this case would be deemed unfair or not.

The wording of the OGD licence confuses things further. It states that "you are free to exploit the information commercially by for example combining it with other information". Although we are not primarily concerned with commercial enterprise in this scenario description, the user could, with some justification, argue that s/he is not doing any more than the licence says that s/he is allowed to do.

At this stage in the process we consulted the UK Information Commissioner's Office and the ensuing discussion led to the fairly clear interpretation that the mere act of identification did not in itself constitute a breach of the DPA and did not in itself mean the user becomes a data controller. In the UK Data Protection Act (1998) the term *data controller* means "a person who – either alone or jointly or in common with other persons – determines the purposes for which and the manner in which any personal data are, or are to be processed". Similarly, EU Directive 95/46/EC defines it as: "the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data."

Of course the above situation would be radically transformed depending on what the user decides to do next. Game theoretically they have at least seven choices:

1. Do nothing
2. Use information disclosed about the respondent for some secondary purpose

3. Inform the respondent
4. Inform the data provider
5. Inform the information commissioner
6. Publicise the breach in some way
7. Some combination of the above in parallel or in sequence.

Each of these will lead to complex games with different combinations of players and each has different legal and material consequences. Specifically, some of them could lead to the user becoming a data controller. Full analysis of this would require a significant piece of work using Mackey's (2009); Mackey and Elliot (2010) game-theoretic framework and was beyond the scope of this project. However, it seems plausible that a user who accesses OG data and uses response knowledge to identify a respondent in that data could avoid any significant costs through careful strategy selection.

To summarise: (i) with EUL data the expected number of users with response knowledge will be small and the users are constrained from reidentification by the licence (and possibly sanctions), but (ii) with the OGD licence the number of potential users is greater by several orders of magnitude and therefore the number of expected potential users with response knowledge is also much greater. The OGD licence does not constrain the user from reidentification and it appears that they would not be constrained by the DPA until after the identification had occurred.

Given that response knowledge attacks are intrinsically higher risk than any orthodox attack scenario, these became the basis for our penetration test.

4. Methodological Approach to the Penetration Test

Prior to the research taking place, the approach was scrutinised closely by the ONS Ethics Committee. The Committee included a legal representative, senior staff with knowledge of social surveys and others with considerable experience of ethical issues in research. The Committee considered the experience of the researchers, in particular the fact that they had carried out research with ONS in the past, and had previously handled sensitive data appropriately. There is also a long-standing data-handling agreement between ONS and the University of Manchester regarding the latters's secure lab for disclosive data. We had also received written confirmation from the ONS Information Asset Owner for this work to go ahead. ONS Legal Services provided assurances on the legality of carrying out this research under the Statistics and Registration Service Act (SRSA), Section 38: "Use of information by the Board".

ONS staff extracted a set of fifty respondents from each of the surveys at random but excluded respondents who had not consented to further research. Respondents were selected from outside the researchers' areas of residence, in order that no respondent would be known to any of the researchers. These two files were passed to the attack simulation team at Manchester University under strict conditions, in particular that the files must be kept secure, must not be passed to anyone outside the immediate research team, and must be destroyed at the end of the work. The files were transferred securely by courier, the data encrypted and password protected. There was an additional proviso that, for the purposes of this research only, the researchers could attempt to identify individual respondents in the EUL files. The two files were:

The LFS ID file consisting of a list of 50 names, addresses and phone numbers.

The LCF ID file consisting of a list of addresses only.

The simulation consisted of four phases:

i) A search phase where intensive web-based searching was conducted on each of the 100 identifiers. On average, half a day of researchers' time per identifier was spent on this.
ii) Commercial data was purchased from a lifestyle database company corresponding to each of the 100 addresses in the identification files.
iii) The resultant information was matched against the microdata and reidentification attempted.
iv) The matches were verified by ONS.

4.1. The Search Process

In this section we describe the search method and recording system that we used. An initial pilot search was undertaken and the search toolset refined in order to maximise the data returns relative to the search effort involved.

For the cases in the LFS ID file it was necessary to verify whether the named respondent was resident at the property and in the case of the LCF ID file to identify all residents at the property at the time of the survey. For both identification files, attempts were made to gather data on all cases, including the hierarchical structure of the household. At least basic details were successfully gathered on most cases, such as approximate age and/or names of (at least some of) the other household members. On average, in the first round of searches each of the 100 cases had about three hours search time allocated to it. In the second round of searching, ten cases from each file were identified as worthy of extended search time, because the initial search indicated that they were highly visible and therefore more information about them was likely to be found. These cases received an additional 3–5 hours of search time.

4.1.1. Design Process

A search database was developed in order to systemise the search and to maximise the return relative to effort. An Excel workbook was created for each individual case.

There are a number of inherent biases associated with the use of different search engines. These biases include ranking metrics (Vaughan and Thelwall 2004; Vaughn and Zhang 2007; Bar-Ilan et al. 2009), business links associated with the searched sites and 'pay to promote' or sponsored search return services versus organic search results (Ma et al. 2010; Agarwal et al. 2011; Tarantino 2013). Some business and subscription-based websites will use this feature, thereby skewing the search results. Other sites, or individual users of sites, may opt to have their associated data blocked, restricted or removed from search engine rankings, thereby skewing data results further. Using multiple search engines does however increase the information returned. This was achieved by using meta-search tools, a selection of search engines and direct searches of specific sites such as *Facebook*, *LinkedIn*, *192* and ancestry search sites. These are detailed in the appendix.

Several test runs were carried out on pilot data to: (i) minimise the search tools used, (ii) maximise the likelihood of positive returns, and (iii) check accuracy. The most important initial search source was *192* as this supported the cross-referencing of names with addresses and provided an age bracket, all essential basic information to assist further searching.

4.1.2. People and Business Search Tool – 192

For this crucial search tool the following information was retrieved: name, address, estimated age, estimated length of residence, data gathered from the electoral roll, recent nearby house sale prices, business details and director report information.

A recent addition to the data that one can obtain from *192* is a 'background report' on each household resident. This report is available at a relatively low cost (£29.94) and draws information into a single report from a range of public sources, that is, edited Electoral Register, company house information, D&B company listings, 118 Data Resources, Local Data Company, Land Registry, Callcredit plc, HALO Mortality file, The Insolvency Service, and The Registry Trust. The report sets out an individual's and their co-residents' (if any) name, address, length of residence, age in five-year bands, mortality, their solvency status, disqualified director status, home ownership status and whether they have any county court judgements against them. Although these sources of information are publically accessible in the UK, pulling together this information is time consuming and requires some understanding of where and how to access public data sources and knowledge of how to cross-reference sources to verify the information obtained. Thus, the *192* background report makes data collection of this type of information significantly easier. We did not use the reports in our searches but we did purchase several reports to assess the information they offered.

Other data sources were used to cross-reference the data obtained from *192* or to fill in the gaps where it was missing. *Ancestry UK* is a good source of information for gathering more accurate information on date of birth, marital status and family details. It should be noted that it was difficult to search for women or children using this site due to the ambiguity over marital status and married names. Again, if the name is common there will be too many results to search effectively. It is much more accurate with older people who have unusual names and who were less likely to migrate.

The land registry was the most reliable source for both verifying residency and home ownership. It is cheap to use (£3 per search) and has approximately eighty percent coverage. In terms of public sources, *Zoopla* was particularly useful for basic property information, as was *Google Maps*. With street view one can see the property and sometimes vehicles and make inferences about affluence and the presence of children.

The cost for using 192 is £89.94 for 100 credits allowing full access. Your credit is depleted every time you search and even searching with an address and full postcode very rarely returns a correct match within 192. Typically you need to search multiple times (approx. 4+) and you are likely to need to trawl through potentially very long lists of similar names and addresses. Credit is depleted very quickly and so 100 credits do not last very long. Individual background reports can be purchased at a cost of £29.94 each. These reports bring together information from numerous public sources such as Companies House and the Individual insolvency register. UK Ancestry was used to corroborate findings with 192 at a cost of £18.95 per month and was purchased for a period of three months.

4.1.3. Social Media

People are becoming more aware about their privacy settings when using social networking sites such as *Facebook* and *LinkedIn*. Recent studies verify that behaviour in social networking site use is changing and individuals are being encouraged to protect their privacy from different sources (see, for example Moreno et al. (2012) and Whipple et al. (2012)). This clearly inhibits intruders gathering information in this way.

With *Facebook* however, even when security settings are high, it may still be possible to gather data on location, address, employer, likes, films, pages, groups, to view photos, see posts by others on a page and so on. There were only a few cases in the study with *Facebook* accounts, but those were a good source of contextual information for those cases.

For *LinkedIn*, if security settings are activated, little information can be seen unless you are a group member. Searching requires registration and a tracing tool is used which lets a person know when you have searched for them. When the security settings are not activated a range of information is potentially available, such as workplace location, job type/title and education. We found a small number of cases where the privacy settings were not activated.

4.1.4. Difficulties Associated With the Search Strategy

It should be noted that searching in the manner described above is not without its difficulties and/or limitations. With our key search tool *192* we encountered problems:

Stability of the information: Searching within *192* does not always return the exact address/name, and multiple attempts were sometimes needed. We found that searching outside *192*, via *Google* for example, was more reliable but then you have to pay to view any information beyond address and name. It is worth noting that there is a cost for each search on *192*.

Inaccuracy of information: We found conflicting name and age profile data for a small number of our cases, where several results were returned for similar addresses with the same person and co-residents named but with contrasting age profiles. These inaccuracies made it difficult to be confident about the data as there was no way to cross-check for accuracy. Similarly, the data on directors was not always reliable when cross-referenced with business directory sources. Common names, names associated with famous people or ambiguous names, such as those that represent objects, were difficult to search as they produced a large number of returns that were difficult to cross-check for accuracy (given the time contstrains).

In terms of specific groups, it is more difficult to search for middle-aged women, especially if their marital status is unclear or unknown. Often once a ‘husband’ is identified it is easier to confirm the woman’s identity, including date of birth and so on, by tracing her through her husband’s marriage records. This is often the only way to track a woman’s birth record. Even if there are children, this group of women can be difficult to identify without knowing their maiden name to check the birth records against.

Beyond the search limitations there are two other research issues that are worth noting. The first of these is time – or more specifically, the search time used in this study. It is entirely possible that an intruder would be willing to spend more time searching than we did and it would seem more likely that the search frame would be much smaller than our 100 cases.

A second point worth making in relation to our search strategy relates to the issue of who is online, and thus who is more likely to be captured. Studies have shown that people from different age cohorts use the Internet in different ways and develop online trust in dissimilar ways (Obal and Kunz 2013). Fewer older people have online social media accounts, and those that do tend to either use them infrequently or have very little information associated with them.

4.2. *The Commercial Data Purchase*

As well as searching public data sources online for information for the 100 cases in the identification files, we also purchased commercial data from the UK lifestyle data company, CACI. The rationale for this was twofold: (i) to consider what additional information might be available through this route, and (ii) to develop first-hand knowledge and experience of the process of obtaining commercial data.

In terms of the second rationale there were two constraints, the first of which was cost. The minimum purchase from CACI is £1,200, which will get you up to 1,000 names and addresses. It typically costs thousands of pounds to access any variables associated with the names and addresses purchased and the number of variables given is usually limited to between 5 and 20. This makes it unlikely that an intruder searching for a single case is likely to consider purchasing this type of data.

The second constraint relates to the purchasing process itself, which is time consuming and requires several screening stages. The first stage requires that you explain why you want to purchase data. At the next stage, you are allocated an account manager who discusses at length why you want the data and negotiates what variables you can have access to and for what time period. The final stage requires you to sign a contract outlining the cost involved, the variables and cases requested and the intended use. Accessing the data took several weeks and many emails and phone calls. The length of this process and the hoops that the user is required to go through on top of the cost are likely to deter all but the most determined intruder.

We informed CACI that we wanted to purchase data for a study looking at what information they held and what disclosure risk (if any) it might pose. We were able to negotiate for the minimum order value the names and addresses for Sample 1 and 2 and those of any other residents at the properties. We were also given all the variables held on each resident at the 100 households. The available variables can be summarised in the three categories: demographic, lifestyle, and socioeconomic.

The key variables present in the dataset and considered useful for matching purposes were: age (five-year bands), sex of each adult, number of children in household, number of bedrooms, social grade, course occupational grouping, years at property (four bands), tenure (three bands), house type, and number of cars (0, 1, 2+).

4.3. *The Matching Process*

We initially explored the idea of using automated matching techniques. However, it was assumed that the intruder we were simulating was not a technical expert and therefore using a probability based approach was not realistic.

Even more importantly, it became evident that a non-automated approach was actually going to be more productive. An example of why this is the case can be found with Case 4 from the LFS sample. This case threw up an enormous amount of data, some of it contradictory. In particular, the crucial piece of information about Case 4's age seemed unreliable on *192.com*. This left us with several possible people who corresponded to Case 4 in the dataset. However, since we knew that Case 4 was in the sample and the information on Case 4's marital partner seemed solid, it was possible to look for household age-relationship combinations that matched Case 4's spouse's age (which appeared reliable). Fortunately most of the combinations were a priori unlikely (large age differences) and only one threw up an actual match. This was then checked for against other information about both Case 4 and his spouse.

In other cases where we found multiple matches on the available information, we were able to examine the dataset to look for an additional variable that differentiated those matches and then go out to look for information in a more targeted way for the case on that variable. This ability to go to and fro between the data, the matching information and the world made the approach much more like a piece of detective work and less like an orthodox statistical matching process. Because of the ethical and time constraints we were under on the project, we could not go as far with this as we might have done otherwise.

Once the attack simulation team had reduced the number of possible matches down as far as they could, they then made an assessment of the certainty of the best matches. This was essentially a subjective expert judgement but was based on a three factors: (i) the closeness of the next best match; (ii) prior knowledge of data divergence issues (time lag between data collection and the survey date, coding mismatches, etc.); and (iii) confidence in the data we had collected (was the information contradictory, was there doubt over whether we had found the correct person, etc.). The matcher in fact assigned a score on a 100-point scale to represent their confidence, which roughly translates into a subjective estimate of the probability of a match being correct. For the purposes of presentation here this scale is collapsed into High (70–100), Medium (50–69), Low (15–49), and Very low (0–14). It should be stressed that no algorithm was used here – the confidence level that was assigned was essentially based on expert judgement by the researcher, taking account of the above factors. A possible extension of this work would be to investigate whether this expert judgement is easily convertible into an algorithm. However, here we are simply concerned with whether matches could be achieved with minimum technical mediation.

The matches were then verified by the ONS team.

5. Matching Results: LFS

5.1. Stage 1: Openly Available Information Only

At Stage 1 we considered only openly available information that we had collected from the Internet.

In total, matches against nine records were attempted, since for the other 41 cases insufficient good-quality information was obtained to make a match attempt viable. Of the nine match attempts, eight produced a correct match, although in two cases it was one of a multiple match.

Summary information is given below. The critical point here is that if the matcher's confidence was high then the matches were successful. These are invariably cases where large amounts of good-quality information were obtained.

5.2. *Stage 2: Adding in the Commercial Dataset*

Before the Stage 1 matches were verified, the process was repeated, this time adding in the commercial data. This increased the number of correct matches to 14. Adding the commercial data did not have a completely monotonic effect on the matches; two of the matches that were correct using only the openly available information were not even attempted with the commercial data because the commercial data provided contradictory information, reducing the certainty. This nonmonotonicity may seem counterintuitive but is related to more a general phenomenon observed, for example by Elliot (2009). Essentially, increasing information has a diminishing return on the power of a set of key information (because information about people is correlated) but the impact on data divergence is linear. So at some point the noise created by the divergence exceeds the information gain from the increasing key size. Where that point is varies depending on the level of divergence, the level of correlation between the key variables and the power of the key variables. On the other hand, fuzzy and probabilistic matching techniques can reverse the process, trading lower precision for higher recall.

The headline results are that by using the openly available information, six of the 50 records were correctly matched one-to-one (12%). Using the commercial data as well pushed this up to 14 (28%). These headline figures however disguise a more significant fact: that the slope of "matchability" is quite steep; the precision rate for high-confidence matches was 100%. The reason for this steepness is primarily because of the amount and quality of the information obtained. Another factor was household size.

6. Matching Results: LCF

The process for the LCF matches was slightly different. The file was not hierarchical and so lacked that defining feature. On the other hand, it did have a low-level geographical indicator on it: the Output Area Classifier (OAC). Obtaining an OAC from a postcode is quite easy once you know how to do it, but it is certainly not obvious and would not necessarily be something that was available to a non-expert. However, a data journalist or similar should be able to obtain the information. We therefore ran two different scenarios: with and without the OAC codes. As it turned out, the OAC was an incredibly useful differentiator key.

The second feature for this dataset was that ONS only provided the simulated attack team with addresses (no names) and this reduced the certainty of the information, which was reflected in the confidence levels that we recorded.

A third difference in the process here was that where there were multiple matches of equal certainty these were recorded as single joint match. For the purpose of comparison, these are turned into effective confidences by simply dividing the total by the number of matches.

At Stage 1a (without OAC codes) there were a possible 20 matches against eight addresses. The mean effective confidence was 16%, which meant that we were predicting

Table 1. Matching attempts using openly available information against the LFS dataset

	1-to-1			1-to-*n*		No match attempted
Confidence level	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	5	0	1.00	2	0	
Medium confidence	1	0	1.00	0	0	
Low confidence	0	0	–	1	0	
Very low confidence			–			41
Overall	6	0	1.00	3	0	41

that we would obtain 3.3 matches. In fact we obtained two. This information is shown in Table 3. This file looks reasonably safe against this simple attack.

However, when the OAC code is added the situation looks very different, as we can see in Table 4. A total of 42 matches were made against 27 addresses. 16.55 matches were predicted to be correct and in fact 18 were. As with the LFS matches, the high-confidence matches were more likely to be correct.

7. General Discussion

The headline finding of this study is that an intruder could, even with partial response knowledge (an address) and using only publically available and/or purchasable information, obtain some correct high-confidence matches without the use of sophisticated matching software. The overall precision rate for high-confidence matches over Tables 1 to 4 is 91%.

This is an important finding. The data here are fairly standard social survey data and contain a mixture of mundane and sensitive information. A correct match against either file would yield income and health information about the target as well as information about other family members, including children. So beyond the obvious legal requirements, the data custodian has a clear duty of care to respondents.

There are seven caveats that must be placed on the details of our headline finding. First, the datasets were older than the public and commercial data (15 months older in the case of the LFS and 27 months in the case of the LCF). This will have increased the data

Table 2. Matching attempts combining openly available data and the CACI commercial dataset

	1-to-1			1-to-*n*		No match attempted
Confidence level	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	10	0	1.00	1	0	
Medium confidence	1	1	0.50	0	0	
Low confidence	1	1	0.50	1	2	
Very low confidence			–			42
Overall	12	2	0.86	2	2	42

Table 3. Matches against the LCF without OAC codes

	1-to-1			1-to-*n*		No match attempted
Confidence level	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	0	0	–	0	1	
Medium confidence	1	1	0.50	0	1	
Low confidence	0	2	0.00	1	1	
Very low confidence			–			42
Overall	1	3	0.25	1	2	42

divergence. In many cases this would have caused a "no match attempted" outcome and therefore will not have affected rates (against confidence), but it will have affected the number of matches attempted and this was factored into the overall confidence level.

Second, the person doing the matching was a statistical disclosure control expert. So, although he restricted himself to unsophisticated manual matching techniques, he was not able to switch off his understanding of data processes, and in particular concepts such as rareness and uniqueness. This would have made confidence estimates more accurate than might be expected for an intruder without that expertise.

Third, the study team was restricted to carrying out legal and ethical actions. We could not, for example, call the named householder. We drew a strict ethical line around our search behaviour so we did not, for example, create fake accounts, attempt to befriend anyone or pose as another person, all of which could potentially yield further information. We also decided that we would not carry out site visits as this would be potentially intrusive. A malicious intruder would not be restricted in the same way. Relatedly, a wealth of technology sources is available to knowledgeable users that could increase the likelihood of an intruder gaining access to an online account such as *Facebook* to access the data. This study did not use any such technology (for a useful background to the issue of using socialbots to hack social networking accounts, please see Boshmaf et al. 2013).

Fourth, the data gatherer was restricted in time by the need to gather information against a representative number of cases. An intruder who simply wanted to identify a single individual could focus a lot more resources on that case.

Table 4. Matches against the LCF dataset with open data using the OAC code

	1-to-1			1-to-*n*		No match attempted
Confidence level	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	5	2	0.71	3	1	
Medium confidence	8	0	1.00	1	1	
Low confidence	1	3	0.25		2	
Very low confidence						23
Overall	14	5	0.74	4	4	23

Fifth, every dataset will be different in terms of its properties (variables, sample size, data structure, etc.) and those properties will interact with the likelihood of correct identifications.

Sixth, the study is a snapshot, albeit a compelling one. The availability of data in the public domain is changing constantly, with the general trend being upwards. This will tend to increase the risk associated with this type of attack. The importance of this issue is increased by the fact that any move from EUL to OGD *for any given dataset* is a one-shot decision. Once the data are released, then the decision is effectively irreversible.

Seventhly, although we were considering response knowledge we were not able to mimic the entirety of what an individual might know about a respondent, only what is available more or less publically. It is likely that an intruder with response knowledge would also have other personal knowledge about the respondent. A potentially interesting extension to the current study would involve re-contacting respondents and asking them to nominate a friend, neighbour, or colleague and then asking the nominee to complete the survey as if they were the respondent.

Some of these factors mitigate the risks indicated by these findings, while others exacerbate them. This makes drawing general conclusions from the **specifics** of the results reported here hazardous.

Nevertheless, the general shape of the results is indicative that moving the survey datasets from end user licence to open data, without any change in their content, would significantly increase the risk of a statistical disclosure on each such dataset and make the likelihood of a disclosure event far greater. The results of the study presented here do suggest that the level of detail on geographical variables and the level of information about household structure are two issues that would need to be attended to in a data-release decision. The restrictions placed on researchers under the EUL in these two surveys are therefore necessary in order to deter an attack, and to provide ONS with some sanctions in case of an attempted or claimed disclosure. As we lay out in Section 3, the response knowledge scenario makes sense with open data but not with EUL licensing.

The financial resources required for the entire study were modest. The main cost was the commercial dataset, which cost us £1200; in addition we spent under £100 on ad hoc services on *192* and the land registry. This works out at an average of £13 per record. In fact, for the LCF study we ended up not using the commercial data as it did not add any value, so the costs there were considerably cheaper.

To summarise, the study presented here provides an illustration of the importance of carrying out well-formulated penetration tests before decisions are made about data releases, particularly irreversible ones such as the release of a file of individual records as open data.

Appendix

List of search sites used to build identifying information

Search engines	Business finder	People finder
http://www.lycos.com/	http://companycheck.co.uk/index	http://www.192.com/
http://www.clusty.com/	http://www.kompass.com/	http://www.infobel.com/en/UK/
http://www.mamma.com/	http://www.hoovers.com/	http://www.linkedin.com/
http://www.metacrawler.com/	http://www.yalwa.co.uk/	
http://uk.search.yahoo.com/	http://www.infobel.com/en/uk/Business.aspx	**Ancestry**
http://www.bing.com/	http://www.yell.com/	http://home.ancestry.co.uk/
http://www.google.com/	http://www.business.com/	
http://www.hotbot.com/	http://www.lexisnexis.co.uk/en-uk/home.page	**Facebook**
http://www.excite.com/		https://www.facebook.com/
http://www.ask.com	**Estate Agents**	
	http://www.zoopla.co.uk/	**Other**
	http://www.rightmove.co.uk/	http://www.iannounce.co.uk/

8. References

Agarwal, A., K. Hosanagar, and M.D. Smith. 2011. "Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets." *Journal of Marketing Research* 48: 1057–1073. Doi: http://dx.doi.org/10.1509/jmr.08.0468.

Backstrom, L., C. Dwork, and J. Kleinberg. 2007. "Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography." In Proceedings of the 16th international conference on World Wide Web, 8–12 May 2007, Banff, AB, Canada. 181–190. Available at: http://dl.acm.org/citation.cfm?id=1242598 (accessed 9 November 2015).

Bar-Ilan, J., K. Keenoy, M. Levene, and E. Yaari. 2009. "Presentation Bias Is Significant in Determining User Preference for Search Results-A User Study." *Journal of the American Society for Information Science and Technology* 60: 135–149. Doi: http://dx.doi.org/10.1002/asi.20941.

Boshmaf, Y., I. Muslukhov, K. Beznosov, and M. Ripeanu. 2013. "Design and Analysis of a Social Botnet." *Computer Networks* 57: 556–578. Doi: http://dx.doi.org/10.1016/j.comnet.2012.06.006.

El Emam, K., E. Jonker, L. Arbuckle, and B. Malin. 2011. "A Systematic Review of Re-Identification Attacks on Health Data." *PloS one* 6(12) : e28071. Doi: http://dx.doi.org/10.1371/journal.pone.0126772.

Elliot, M.J. 2009. "Using Targeted Perturbation of Microdata to Protect Against Intelligent Linkage." In Proceedings of UNECE Work Session on Statistical Confidentiality, 17–19 December 2007, Manchester. Available at: http://www.unece.org/index.php?id=14503#/ (accessed 14 December 2014).

Elliot, M.J. and A. Dale. 1998. “Disclosure Risk for Microdata Report to the European Union ESP/204 62 361–372.” Available at: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:19b497 (accessed 9 November 2015).

Elliot, M.J. and A. Dale. 1999. “Scenarios of Attack: the Data Intruder’s Perspective on Statistical Disclosure Risk.” *Netherlands Official Statistics* 14: 6–10. Available at: http://bit.ly/1ScX0cS (accessed 9 November 2015).

Elliot, M.J. and E. Mackey. 2014. “The Social Data Environment.” In *Digital Enlightenment Yearbook*, edited by K. O’Hara, S.L. David, D. de Roure, and C. M-H. Nguyen. 253–263. Doi: http://dx.doi.org/10.3233/978-1-61499-450-3-253.

Gymrek, M., A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich. 2013. “Identifying Personal Genomes by Surname Inference.” *Science* 339: 321–324. http://dx.doi.org/10.1126/science.1229566.

Ma, Z.M., G. Pant, and O.R.L. Sheng. 2010. “Examining Organic and Sponsored Search Results: A Vendor Reliability Perspective.” *Journal of Computer Information Systems* 50: 30–38. Available at: http://bit.ly/1MSpcni (accessed 9 November 2015).

Mackey, E. 2009. *A Framework for Understanding Statistical Disclosure Control Processes: A Case Study Using the UK’s Neighbourhood Statistics*. PhD Thesis, University of Manchester. Available at: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.502255 (accessed 9 November 2015).

Mackey, E. and M.J. Elliot. 2010. “The Application of Game Theory to Disclosure Events.” *Proceedings of UNECE worksession on Statistical Confidentiality, Bilboa, December 2009*. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp. 40.e.pdf (accessed 09/11/2015).

Mackey, E. and M.J. Elliot. 2013. “Understanding the Data Environment.” *XRDS* 20: 37–39. http://dx.doi.org/10.1145/2508973.

Malin, B. and L. Sweeney. 2004. “How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-Identification to Evaluate and Design Anonymity Protection Systems.” *Journal of Biomedical Informatics* 37: 179–192. http://dx.doi.org/10.1016/j.jbi.2004.04.005.

Moreno, M.A., A. Grant, L. Kacvinsky, P. Moreno, and M. Fleming. 2012. “Older Adolescents’ Views Regarding Participation in Facebook Research.” *Journal of Adolescent Health* 51: 439–444. http://dx.doi.org/10.1016/j.jadohealth.2012.02.001.

Müller, W., U. Blien, and H. Wirth. 1995. “Identification Risks of Micro Data. Evidence from Experimental Studies.” *Sociological Methods and Research* 24: 131–157. http://dx.doi.org/10.1177/0049124195024002001.

Narayanan, A. and V. Shmatikov. 2008. “Robust De-Anonymization of Large Sparse Datasets.” In Proceedings of the 2008 IEEE Symposium on Security and Privacy, 18–21 May 2008, Berkeley/Oakland, CA, USA. 111–125. Doi: http://dx.doi.org/10.1109/SP.2008.33.

Narayanan, A. and V. Shmatikov. 2009. “De-Anonymizing Social Networks.” In Proceedings of the 2009 IEEE Symposium on Security and Privacy, 17–20 May 2009, Berkeley/Oakland, CA, USA. 173–187. Doi: http://dx.doi.org/10.1109/Sp.2009.22.

Obal, M. and W. Kunz. 2013. “Trust Development in E-Services: A Cohort Analysis of Millennials and Baby Boomers.” *Journal of Service Management* 24: 45–63. Doi: http://dx.doi.org/10.1108/09564231311304189.

Paass, G. 1988. “Disclosure Risk and Disclosure Avoidance for Microdata.” *Journal of Business and Economic Statistics* 6: 487–500. Doi: http://dx.doi.org/10.1080/07350015.1988.10509697.

Tarantino, E. 2013. “A Simple Model of Vertical Search Engines Foreclosure.” *Telecommunications Policy* 37: 1–12. Doi: http://dx.doi.org/10.1016/j.telpol.2012.06.002.

Vaughan, L. and M. Thelwall. 2004. “Search Engine Coverage Bias: Evidence and Possible Causes.” *Information Processing & Management* 40: 693–707. Doi: http://dx.doi.org/10.1016/S0306-4573(03)00063-3.

Vaughan, L.W. and Y.J. Zhang. 2007. “Equal Representation by Search Engines? A Comparison of Websites Across Countries and Domains.” *Journal of Computer-Mediated Communication* 12: 888–909. Doi: http://dx.doi.org/10.1111/j.1083-6101.2007.00355.x.

Whipple, E.C., K.L. Allgood, and E.M. Larue. 2012. “Third-Year Medical Students’ Knowledge of Privacy and Security Issues Concerning Mobile Devices.” *Medical Teacher* 34: e532–e548. Doi: http://dx.doi.org/10.3109/0142159X.2012.670319.

Received December 2014
Revised November 2015
Accepted November 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 349–373, http://dx.doi.org/10.1515/JOS-2016-0020

Interviewer Effects on a Network-Size Filter Question

Michael Josten[1] *and Mark Trappmann*[2]

There is evidence that survey interviewers may be tempted to manipulate answers to filter questions in a way that minimizes the number of follow-up questions. This becomes relevant when ego-centered network data are collected. The reported network size has a huge impact on interview duration if multiple questions on each alter are triggered. We analyze interviewer effects on a network-size question in the mixed-mode survey "Panel Study 'Labour Market and Social Security'" (PASS), where interviewers could skip up to 15 follow-up questions by generating small networks. Applying multilevel models, we find almost no interviewer effects in CATI mode, where interviewers are paid by the hour and frequently supervised. In CAPI, however, where interviewers are paid by case and no close supervision is possible, we find strong interviewer effects on network size. As the area-specific network size is known from telephone mode, where allocation to interviewers is random, interviewer and area effects can be separated. Furthermore, a difference-in-difference analysis reveals the negative effect of introducing the follow-up questions in Wave 3 on CAPI network size. Attempting to explain interviewer effects we neither find significant main effects of experience within a wave, nor significantly different slopes between interviewers.

Key words: Partial falsification; network generator; filter questions; interviewer cheating.

1. Introduction and Research Question

Within the total survey error framework (Groves et al. 2004; Biemer 2010) survey interviewers play a central role as the agents who implement the survey design. Their tasks may comprise selecting the right target person, achieving contact with the target person and eliciting cooperation, explaining the task to the survey respondent, asking the questions, probing and coding answers. Consequently, interviewers can influence almost every error source in a survey.

This article focuses on interviewer effects on measurement. Interviewers may influence survey measurement in a variety of ways: there is ample evidence that the mere presence of an interviewer causes respondents to give more socially desirable answers than in self-administered surveys (Tourangeau and Yan 2007). Furthermore, respondents may be influenced by observable interviewer characteristics like his or her age (Freeman and Butler 1976), gender (Groves and Fultz 1985; Huddy et al. 1997) or ethnic affiliation (Schuman and Converse 1971) – where direction and strength of the effect have often

[1] Opinion Market Research & Consulting GmbH, Rollnerstr. 8, 90408 Nuremberg, Germany. Email: michael.josten@opinion.de
[2] Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany and University of Bamberg, Germany. Email: mark.trappmann@iab.de

Acknowledgments: We would like to thank Stephanie Eckman, three anonymous reviewers, and the associate editor, for their valuable comments on and suggestions for previous versions of this article.

been shown to depend on the interaction between respondent and interviewer characteristics. Another source of interviewer variance is nonstandard behavior during the interview like probing (Freeman and Butler 1976; Mangione et al. 1992; van der Zouwen et al. 2004). All explanations mentioned so far refer to unintended interviewer effects on measurement. A different explanation focuses on deliberate interviewer misbehavior in the sense of partial interview falsifications, emanating from an incentive to cheat.

Following the seminal article by Crespi (1945), there have been regular publications on interviewer cheating (cf. Blasius and Friedrich 2012 for a brief overview), focusing on interviewers' motivation for cheating (Crespi 1945), on methods to detect (Biemer and Stokes 1989) and prevent (AAPOR 2003) cheating and on consequences for estimates from surveys (Schnell 1991; Schraepler and Wagner 2005). Cheating is usually considered to be a problem in CAPI rather than CATI surveys (Guterbrock 2008). While interviewers who fabricate complete interviews run a high risk of detection, other, more subtle techniques are harder to detect (Schnell 2012). These include the selection of the wrong target persons, the fabrication of parts of the interview, but also "[. . .] deliberately miscoding the answer to a question in order to avoid follow-up questions" (Guterbrock 2008, 267). Depending on the payment scheme and the tightness of supervision, interviewers might be tempted to increase their efficiency by editing answers to filter questions in a way that reduces interview duration.

With respect to panel surveys, interviewers face a high risk of detection when fabricating complete interviews. Inconsistencies in answers across waves are easily detected. In addition, in most panel surveys, respondents are routinely contacted by the survey agency in between waves for tracking purposes and a complete fabrication would immediately become apparent in case of interviewer changes. In the German Socio-Economic Panel Survey (GSOEP), such falsifications were thus almost exclusively found in initial wave interviews of new refreshment samples (cf. Schraepler and Wagner 2005). This makes resorting to partial falsifications, like taking shortcuts at filter questions, even more attractive in panel surveys.

Underreporting in filter and screener questions has received increased attention in recent years (Kreuter et al. 2011; Tourangeau et al. 2012). Several projects have been launched in order to investigate the processes leading to such underreporting (Tourangeau et al. 2015; Eckman et al. 2014). Most of the results published so far refer to respondent effects. The investigation of interviewer effects on measurement in filter questions has been identified as an important topic for future research in this literature (Kreuter et al. 2011) as only few studies have been devoted to this.

We identified three studies dealing with interviewer effects on filter questions in general. Schnell and Kreuter (2000) argue that differences in victimization rates between different German victimization surveys are in part due to interviewers taking shortcuts at filter questions (Schnell and Kreuter 2000, 114f.). In a later study, Matschinger et al. (2005) investigated a mental health screening question that triggered a series of follow-up questions if endorsed. They found huge interviewer effects, and a latent class regression clearly revealed classes of dishonest interviewers who exhibited learning effects across fieldwork. Finally, Kosyakova et al. (2015) found interviewers affected the endorsement rate of filter questions in the German PASS panel. Controlling for a large set of interviewer and respondent attributes, about three percent of the variance in responses to filter questions can be shown to be interviewer variance. For CAPI interviewers the endorsement rate and thus the number of follow-up questions decreased with growing interviewer experience.

One specific area where interviewer effects on filter questions can be assumed to be of particular importance is the collection of ego-centered social network data in surveys. The number of persons reported in response to network generator questions (i.e., questions that are designed to elicit the name of alters in the network of a respondent) can have a huge impact on subsequent interview duration if – as is often the case – multiple follow-up questions on each alter or even questions on the relationship between each pair of alters are asked. Given that this question is also less objective than other common filter questions, for example on current employment status, it may stand out to interviewers as the one question where a lot of time can be saved and detection probability is low. Consequently there is some evidence of strong interviewer effects on the number of persons named in response to network generator questions (van Tilburg 1998; Marsden 2003; Brüderl et al. 2013).

Interviewer effects on filter questions are usually reported for PAPI and CAPI surveys where interviewers are not under constant supervision by their survey organization. We do not know of any CATI study that finds strong effects. (Kreuter et al. (2011) even found no effect of experimentally varied payment schemes (payment by case vs. payment by hour) on answers to filter questions in a CATI survey. They argue that the routine monitoring is what keeps CATI interviewers from cheating). A serious drawback of previous research is that interviewers are usually strongly confounded with primary sampling units (PSUs) and thus interviewer and area effects are difficult to separate. In this article CATI and CAPI interviews within identical PSUs are utilized to disentangle this effect.

Using a mixed-mode survey, we will evaluate four research questions in this article: can we find any interviewer effects on a network-size filter question? Does the size of these effects differ by the mode-specific combination of payment scheme and supervision practice? Can these effects be explained by area differences in network size? And do interviewers learn to cheat as they gain experience with the survey?

2. Previous Research

Social network size measures have been discussed as being prone to interviewer effects for some time. This is usually investigated by analyzing to what extent observed network sizes differ across interviewers and what proportion of these variances can be explained by interviewer characteristics or by differences in respondent composition (van Tilburg 1998; Marsden 2003; Brüderl et al. 2013; Paik and Sanchagrin 2013). Typically the intraclass correlation (ICC) is used as a measure for the size of the interviewer effect. The ICC can be interpreted as the (covariate-adjusted) proportion of variance that is due to differences between interviewers. For a formal definition in the context of random-effects models, see the methods section.

Controlling for a large number of sociodemographic respondent and interviewer characteristics, both van Tilburg (1998) and Marsden (2003) found a strong interviewer effect on the number of social contacts elicited, measured by an ICC of $\rho_{int} = 0.15$ and $\rho_{int} = 0.13$, respectively, after controlling for respondent-level predictors of network size. Compared to the average size of interviewer effects for any kind of questions in personal interview surveys reported by Groves (1989, 319), these effects are quite large. Nevertheless, recent studies discovered even larger interviewer effects far exceeding this for network size: Paik and Sanchagrin (2013) found an intraclass correlation of up to $\rho_{int} = 0.22$, while Brüderl et al. (2013) found by far the largest effect of $\rho_{int} = 0.40$ after controlling for respondent-level predictors. However, Marsden (2003) and Brüderl et al. (2013) could not identify interviewer characteristics significantly affecting network sizes. Van Tilburg (1998) found experienced interviewers produced reduced network sizes, whereas interviewers who were better educated and had more experience in the particular survey generated larger networks.

One common approach is to classify interviewers according to the (pattern of) network sizes they generate. For example, Brüderl et al. (2013) identified three types of interviewers by analyzing which interviewers affect the intraclass correlation most, thus distinguishing between diligent, normal, and fraudulent interviewers. This approach does not take into account the temporal pattern of responses. An alternative approach applied by Matschinger et al. (2005) for a mental health screening question makes use of this temporal order of the interviews. They used latent class regression techniques to create classes of interviewers with similar patterns of responses to filter questions across their sequence of interviews. Thereby, they were able to identify classes of interviewers who "learn" to cheat while becoming more experienced with the administration of the questionnaire.

The network delineation instrument that previous studies used in their analysis was an interviewer-administered name generator where respondents were instructed to report the names of persons they are regularly in touch with in specific situations. The delineating procedure thus constituted a complex task for both respondents and interviewers. Respondents had to interpret these questions correctly and select appropriate persons from their network, while interviewers had to check answers and apply appropriate probing strategies in case of difficulties of comprehension or lacking plausibility. Due to the questions' complexity, the explanation for the large interviewer variances provided by van Tilburg (1998) and Marsden (2003) resides in different probing strategies among the interviewers. Thus they advise training interviewers more carefully. In contrast, Brüderl et al. (2013) and Paik and Sanchagrin (2013) explain the differences in network size by deliberate interviewer misbehavior in order to shorten an interview, an explanation that has been largely neglected by previous literature. All of these studies investigate interviewer effects on network size in a face-to-face setting where there was a nonrandom allocation of interviewers to respondents. In these designs, there is not sufficient interpenetration to separate interviewer effects from area effects (Groves 1989, 270f.). Thus regional variance and interviewer variance are confounded (O'Muircheartaigh and Campanelli 1998; Schnell and Kreuter 2005). Consequently, the unexplained interviewer variance that these studies interpret as evidence for interviewer effects might instead be an unexplained area effect and reflect local differences in network size instead. To tackle this, van Tilburg (1998) and Brüderl et al. (2013) included area dummies in their models. However, this strategy leaves only those areas which were worked by more than one interviewer for the identification of interviewer effects. Marsden (2003) argued against area effects by showing that the intra-interviewer correlation for a less complex "global" network size measure, almost identical to the one that is analyzed in this article, amounts to only 0.04–0.05. Paik and Sanchagrin (2013) compared outlier interviewers to the remaining interviewers in the same PSU to prove that there is no significant difference in network size estimates between geographical regions that can account for the existence of outlier interviewers.

In contrast to previous approaches, we are able to exploit the mixed-mode design implemented in PASS as well as the longitudinal character of the data. The mixed-mode design brings with it a direct comparability of a CATI sample, in which interviewer assignment is independent of region, and a CAPI sample, in which interviewer and region are confounded. Different incentive structures and supervision practices between modes can thus be exploited to differentiate between explanations referring to the complexity of the filter question and explanations referring to interviewer cheating.

3. Data and Hypotheses

This article uses data from the third wave of the German panel survey "Labour Market and Social Security" (PASS) (Trappmann et al. 2010; Trappman et al. 2013). PASS is an annual panel survey that focuses on labor market, poverty, and social policy research. While the target population is all households in Germany, low-income households are oversampled. In each household the head of the household answers a household questionnaire. Subsequently, every person aged 15 or older is interviewed with a person questionnaire. PASS is implemented with a sequential mixed-mode design within 300 primary sampling units. Most of the households are interviewed in CATI mode, while households and persons that cannot be contacted by phone or that prefer a personal interview are interviewed in CAPI. In Wave 3, which is analyzed here, 5,663 persons (42.1%) have been interviewed in CAPI and 7,776 (57.9%) in CATI mode (Bethmann and Gebhardt 2011).

The PASS survey offers the rare opportunity of a mixed-mode survey where answers to the same filter question are available from CATI as well as CAPI interviews in the same primary sampling units. This entails a direct comparability of the interviewer influences in the context of different supervision practice and incentive structure as well as the possibility to consider interviewer and area effects separately.

The network-size filter question that is analyzed in this article was asked in the person questionnaire in Wave 3. Respondents were first asked whether they have any close friends or family members outside their household with whom they have a strong relationship. If this first filter question was endorsed, it was followed by the inquiry concerning the number of such contacts outside the household. For each of the three closest relationships, respondents were asked five follow-up questions about the alter's gender, education, employment status, frequency of contact and the kinship relation between respondent and alter. These follow-up questions were limited to the three closest friends in an attempt to constrict the incentive to cheat. (The questionnaire can be found at http://doku.iab.de/fdz/pass/Questionaires_English_W3.zip). While the same two initial questions on personal network size had been asked in the first two waves of the panel study, no follow-up questions had been asked in these waves. Therefore differences in response behavior between Waves 1 and 2 on the one hand and Wave 3 on the other hand can be exploited to analyze the effects of adding additional burden to certain answers to this question.

Depending on the answer to the initial two questions, in Wave 3 there can be between zero and 15 follow-up questions. The easiest way for interviewers to reduce their workload would be to enter “no”, “refused” or “don’t know” without even asking the initial question, thus skipping the whole set of follow-up questions. An alternative would be to ask and record the first question truthfully and then enter a number smaller than three for the number of close friends (irrespective of whether the second question was actually asked or what answer was given). At the same time, there is hardly any chance that respondents who had never been asked this set of questions before would notice these shortcuts.

In PASS Wave 3 overall, 129 CATI interviewers conducted between two and 238 interviews (with a mean of 60) and 243 CAPI interviewers conducted one to 97 interviews (with a mean of 23). All interviewers work for the same survey organization. Nevertheless, CATI and CAPI interviewers work in completely different environments. On the one hand, modes differ in their payment scheme. While CAPI interviewers are paid the same amount for each successful interview, irrespective of its duration, CATI interviewers are paid by the hour. Consequently, interview duration should not matter for CATI interviewers, while in CAPI a shorter interview implies a higher hourly wage.

On the other hand, modes differ in the tightness of supervision: CAPI interviewers can work fairly autonomously as they organize the workload assigned to them themselves and do not face a recording of their interviews. As a means to detect falsifications, a random sample of respondents and nonrespondents is contacted by the survey organization either by postcard or by phone after the interview. These tests focus mainly on verifying the assigned disposition codes and the time and duration of the interview. It seems very unlikely that cheating with a filter question can be inferred from these tests. In contrast to this, CATI interviewers are frequently monitored by a supervisor in the call centers of the survey organization, who would likely detect any deviations from the protocol leading to a dismissal from the study in case of repeated misbehavior (cf. Büngeler et al. 2010).

All in all – unless they have a preference for creating high-quality data – CAPI interviewers face strong incentives to cheat when collecting data on network size. At the same time it seems very unlikely that there are deliberate falsifications by CATI interviewers. They do not gain anything from cheating, as a shorter interview means that they are paid less or have to contact additional respondents within their shift. Interviewer effects in CATI should thus be limited to unintentional effects. Hence, we derive Hypothesis 1a and 1b with respect to Wave 3 of PASS:

> Hypothesis 1a: The interviewer effect on network size is stronger in CAPI than in CATI.
> Hypothesis 1b: Network sizes in CAPI interviews are smaller compared to CATI interviews.

As the same network size questions had been asked in Waves 1 and 2 of the panel survey without triggering any follow-up questions, an incentive for CAPI interviewers to cheat has been introduced for the first time in Wave 3. This should reduce network size in CAPI in Wave 3 compared to network size in CAPI in Wave 2. This leads to Hypothesis 1c:

> Hypothesis 1c: The introduction of up to 15 follow-up questions in Wave 3 leads to a decrease in CAPI network size compared to Wave 2.

There is some evidence that experienced interviewers produce stronger systematic effects on filter questions than unexperienced interviewers (Hughes et al. 2002). Groves et al. (2004, 273) argue that this might be due to reward systems for CAPI interviewers focusing on productivity instead of measurement quality. Thus we derive our fourth hypothesis:

> Hypothesis 2a: The network size in CAPI mode in Wave 3 decreases with the experience of an interviewer *across* studies.

Furthermore, it is highly plausible to assume that interviewers learn how to cheat effectively during the fieldwork of a specific survey. It might take some time for interviewers to realize which questions have a huge impact on interview duration and can be manipulated easily at the same time (Matschinger et al. 2005). In line with this, we expect the cheating behavior not to be constant across time as interviewers have to exhibit a learning effect first. Thus, our fifth hypothesis is:

> Hypothesis 2b: The network size in CAPI mode in Wave 3 decreases with the experience of an interviewer *within* the study.

4. Methods

To test these hypotheses, we will for the most part use multilevel regression models (Swamy 1971; Goldstein 1986; Longford 1993) to take into account the clustering of respondents within interviewers and to determine the proportion of the total variance that can be attributed to the interviewers as a measure of interviewer influence. In the absence of a gold-standard measurement for network size, interviewer effects can be identified by estimating intra-interviewer correlations ρ_{int} of the survey measure. Within the frame of a random-effects ANOVA model, this is given by $\rho_{int} = \sigma_b^2/\sigma_b^2 + \sigma_w^2$ where σ_b^2 is the between-interviewer variance and σ_w^2 the within-interviewer variance. (We follow the notation in Rabe-Hesketh and Skrondal 2012.) We will refer to the intra-interviewer correlation either as ρ_{int} or as ICC (for intraclass correlation). In the presented form, however, ρ_{int} confounds several effects on the interviewer level: respondents assigned to different interviewers may differ in true network size from the beginning. This may be moderated further by the interviewers' differing ability to contact and convince respondents with different network sizes to participate (West et al. 2013). One approach to fix this is to include variables that explain individual network size in a random-intercept model. The intra-interviewer correlation is then only estimated based on the unexplained variance of the model. Drawing on a rich literature on network size, we decided to include gender, age, education, employment status, health, membership in organizations, and the existence of a partner and the number of children outside the person in question's own household as individual predictors of network size (van Tilburg 1998; Marsden 2003; McPherson et al. 2006; Brashears 2011; Brüderl et al. 2013; Paik and Sanchagrin 2013).

There is an incentive to create networks of sizes smaller than three. For each network person reported less than that, the workload for the interviewer decreases by five questions. However, once the size of three has been reached, it does not matter whether three or more persons are named. While this data structure might be analyzed using double-hurdle models (Cragg 1971) (where step one models whether a network size smaller than three is reported and step two models how much smaller it is) or ordinal regression models (O'Connell 2006) that use three or more as maximum category, we decided to choose a more straightforward analysis strategy here. In the subsequent analysis, we will only distinguish between networks of size three or more, where the maximum number of follow-up questions was asked, and networks of size smaller than three, where some number of questions were skipped. We will refer to this variable as the "network-size dummy variable" throughout this article.

With this dummy constituting the dependent variable in our model, we suppose a logistic random-intercept model (Snijders and Bosker 2000, 207 ff.; Rabe-Hesketh and Skrondal 2012, 520 ff.). As a starting point we use an empty model

$$logit\left\{ \Pr\left(y_{ij} = 1 \middle| x_{ij}, z_j, \zeta_j\right)\right\} = \beta_1 + \zeta_j \tag{1}$$

where Y is the dependent variable that takes on a value of 1 if a respondent i's network is size three or larger. It only includes a mean intercept β_I and a random intercept $\zeta \sim N(0,\sigma_b)$ that represents the deviation of interviewer j's intercept from β_I. Note that there is no respondent-specific error term in (1). To be able to compute $\rho_{int,}$ a latent-response formulation can be used where an error term ε_i is assumed to have a standard logistic distribution with mean zero and within-respondent variance $\sigma_w = \pi^2/3 = 3.29$ (Snijders and Bosker 2000, 223 ff.; Rabe-Hesketh and Skrondal 2012, 510 ff.). In a second step, we will extend our model by explanatory variables to a full random-intercept model

$$\begin{aligned} logit\left\{ \Pr\left(y_{ij} = 1 \middle| x_{ij}, z_j, \zeta_j\right)\right\} =& \beta_1 + \beta_2 x_{ij2} + \ldots + \beta_k x_{ijk} \\ & + \beta_{k+1} z_{jl} + \ldots + \beta_{k+l} z_{jl} + \zeta_j \end{aligned} \tag{2}$$

where X_{ij} are predictors for network size on the respondent level and Z_j are predictors on the level of the interviewer.

The random-intercept model in (2) assumes that the slope of the dependent variable by interviewer experience is constant across interviewers and that each additional interview has the same effect on log odds. It is highly plausible, however, that interviewers differ in their motivation to work as an interviewer and thus in their receptiveness to incentives to reduce workload at the cost of data quality. Accordingly, we can introduce a random slope for interview sequence (here denoted as x_{ijk}) to model its effect heterogeneity across interviewers, resulting in a specific coefficient for interview sequence for each interviewer:

$$\begin{aligned} logit\left\{ \Pr\left(y_{ij} = 1 \middle| x_{ij}, z_j, \zeta_{j1}, \zeta_{j2}\right)\right\} =& \beta_1 + \beta_2 x_{ij2} + \ldots + \beta_k x_{ijk} \\ & + \beta_{k+1} z_{jl} + \ldots + \beta_{k+l} z_{jl} + \zeta_{j1} + \zeta_{j2} x_{ijk} \end{aligned} \tag{3}$$

We estimated all random-intercept models using the xtlogit-command for random-effects estimation in Stata 12.1 and estimated the random-slope model using the xtmelogit command. Significance tests for random intercepts and random slopes were performed using likelihood-ratio tests compared to models without random intercept or slope. Stata approximates log likelihoods using adaptive Gaussian quadrature (AGQ) for the xtlogit and AGQ with the multicoefficient extension from Pinheiro and Bates (1995) and the multilevel extension from Pinheiro and Chao (2006) for xtmelogit (cf. StataCorp 2011).

To test Hypothesis 1c, we will make use of the data's longitudinal nature. Within a difference-in-difference approach (Lechner 2011), we compare differences in CAPI network sizes between Waves 2 and 3 to differences in CATI network sizes between Waves 2 and 3 to assess whether the introduction of follow-up questions for each network person brought about a negative treatment effect on network size. We assume that a treatment is only introduced in CAPI, as CATI interviewers face no changed incentives through the change of the question's character.

5. Results

5.1. Descriptive Results

Table 1 shows average network sizes and proportions of networks smaller than three by mode for Waves 2 and 3. Differences in Wave 3 outcomes between modes are obvious: while only 14.1% of the CATI respondents exhibit small networks, 40.2% of the CAPI respondents do so.

The large mode differences suggest that CAPI interviewers were indeed tempted by the incentive to cheat. However, without more rigorous models these differences might be due to self-selection of isolated respondents to the CAPI part of the study. As all respondents whose phone number could not be identified from the sampling frame or from a telephone directory search were approached in CAPI mode, this alternative hypothesis gains credibility: isolated persons should be more likely to have an unlisted number or no phone number at all. When comparing results for Wave 3 to results for Wave 2 in Table 1, one gains the impression that CAPI networks were indeed smaller even before the incentive to cheat was introduced. In Wave 2, 28.7 percent of CAPI networks, but only 13.6 percent of CATI networks were of a size smaller than three. However, there is a striking growth in small networks in CAPI that cannot be found in CATI.

Further insights can be gained by investigating the proportion of networks of size three or larger by interviewer. For this purpose, all interviewers with less than ten interviews are excluded. Among the remaining 116 CATI interviewers, the standard deviation in the proportion of networks of size three or more is 0.076, while among the remaining 165 CAPI interviewers it is 0.247, that is, CAPI interviewers differ more from each other in the network sizes they produce.

Table 1. Network size by mode and wave

		Proportion of networks smaller than three (in %)	Average network size (std dev)	*n*
CATI	Wave 2	13.6	8.82 (8.89)	7,877
	Wave 3	14.1	8.25 (8.09)	7,771
CAPI	Wave 2	28.7	6.61 (8.28)	4,567
	Wave 3	40.2	4.98 (6.49)	5,633

Figure 1 shows a quantile-quantile plot of the network-size dummy variable by interviewers for CATI and CAPI. It plots quantiles of the distribution of the network-size dummy variable by interviewers in one mode against quantiles of the same distribution in the other mode.

This figure shows that differences between modes are most pronounced when the lower quantiles of the distribution are compared. For ease of interpretation, selected quantiles (10, 25, 50, 75, 90) are highlighted in the plot. For example, the ten-percent quantile is at 0.211 for CAPI, signifying that the ten percent of CAPI interviewers who produce networks of size three or more least frequently do so in at most 21.1 percent of their interviews. In contrast, the ten-percent quantile among CATI interviewers is at 0.778.

This difference of 0.567 declines via 0.218 for the median to 0.085 for the last decile. These descriptive results seem to point to a subgroup of CAPI interviewers who produce very small proportions of networks of size three or more, while CATI interviewers are much more homogenous in their results.

Figure 2 displays by mode how the proportion of networks of size three or more changes within the sequence of interviews conducted by an interviewer. A locally weighted regression using the Stata command "lowess" (Cleveland 1979) was used to smooth the curve. In order not to give the few interviewers with more than 50 interviews too much weight, the figure is cut off at the 50th interview.

As could be expected considering there is no incentive to shorten an interview, in CATI mode there is no visible trend from the first to the last interview. In CAPI, however, we find no decreasing trend as we might have expected. On the contrary, the proportion of respondents with networks of size three or more shows a small decline across the first few interviews and then increases from about the tenth interview.

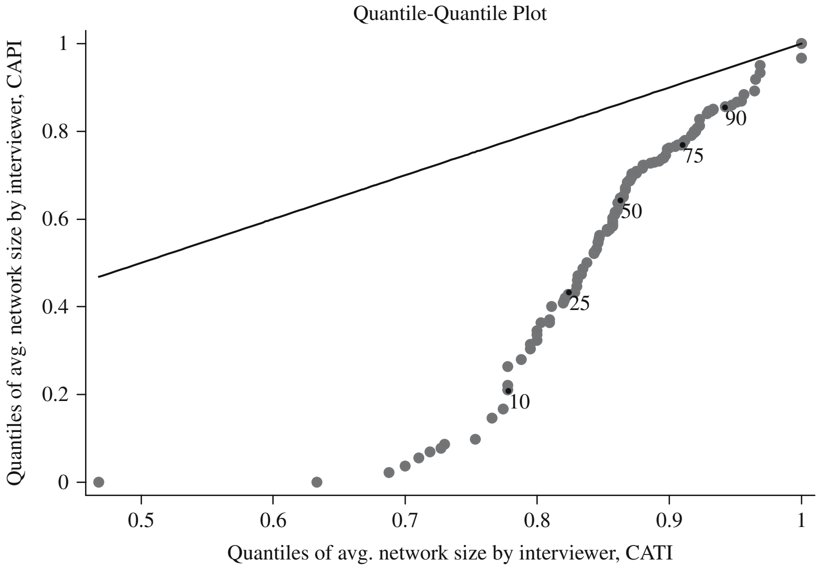


Fig. 1. Quantile-quantile plot of the network-size dummy variable by interviewer for CATI and CAPI (only interviewers with ten or more interviews)

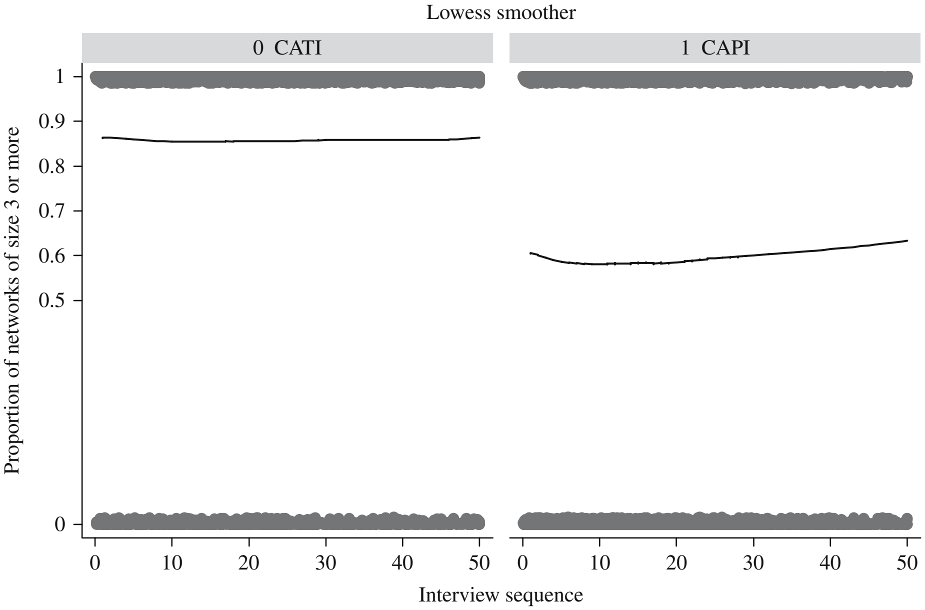


Fig. 2. Proportions of networks of size three or more by interview sequence in CATI and CAPI

5.2. *Results from a Difference-in-Difference Approach*

In this section the longitudinal character of the study is exploited. As the network-size question was not used as a filter question in Wave 2, there was no incentive to cheat on this question for any interviewer in Wave 2. The introduction of up to 15 follow-up questions in Wave 3 brought about an incentive to cheat for CAPI interviewers only. Thus, a difference-in-difference approach (DiD in the following), comparing the differences between Wave 2 and Wave 3 outcomes for CATI and CAPI interviews, can be used to investigate the effect of introducing follow-up questions.

While the number of extra questions is proportional to the network size for up to three network contacts, it remains constant for all network sizes exceeding three. In this section we will therefore use a truncated network-size variable that is equal to the network size for networks sized three or smaller and three for networks larger than three. This variable can be interpreted as the number of loops over the five-name interpreter questions in Wave 3. Of course this is hypothetical in Wave 2, where there were no actual loops.

In contrast to Table 1, where all cases were included that were interviewed in either Wave 2 or 3, Table 2 focuses only on individual respondents who participated in both of these waves. Altogether, there are 9,113 such respondents with valid network size in both waves. Column (3) contains the difference between (2) and (1). A negative number corresponds to a decrease in the workload for the respective group.

We will distinguish between five different groups of respondents, depending on the interview mode they were interviewed in. Those interviewed in CATI in Waves 2 and 3 can serve as a reference group because cheating incentives have remained unchanged. In this large group ($n = 5{,}330$), the mean number of loops has remained almost constant (2.751 to 2.745). The second group was switched from CAPI in Wave 2 to CATI in Wave 3. Although

Table 2. Network-size differences between Wave 2 and 3 by group

Group	(1) Mean number of loops Wave 2	(2) Mean number of loops Wave 3	(3) Difference (2)-(1)	(4) Diff-in-Diff (compared to Group 1)	(5) *n*
1 CATI -> CATI	2.751	2.745	−0.006	–	5,330
2 CAPI -> CATI	2.264	2.622	0.358	0.365 ($p = 0.015$)	53
3 CATI -> CAPI	2.693	2.391	−0.301	−0.295 ($p < 0.001$)	322
4 CAPI -> CAPI (different ivwer)	2.519	2.359	−0.160	−0.154 ($p = 0.066$)	412
5 CAPI -> CAPI (same ivwer)	2.455	2.190	−0.265	−0.259 ($p < 0.001$)	2,996
Total					9,113

there should be no change in incentives in this group, it is the only one where the number of loops increased from 2.264 to 2.622. Note, however, that this group is rather small ($n = 53$). In addition, in the PASS panel design switches from one mode to another are mainly due to the interviewer's inability to conduct another interview with the respondent in the original mode, which might for example be due to the respondent having moved. Thus, whatever caused those respondents to switch modes might have influenced their network sizes as well.

In all other groups we expect negative changes in the number of loops due to the introduction of an incentive to cheat. For respondents who were switched from CATI in Wave 2 to CAPI in Wave 3 and respondents who were interviewed in CAPI mode in both waves but by different interviewers, CAPI interviewers have not worked these cases in the previous wave and thus cannot have any knowledge of the previous-wave network size. As expected, the number of loops decreases from 2.693 in Wave 2 to 2.391 in Wave 3 for respondents who switched from CATI to CAPI. Again, one should be aware that this is a selective group.

Another 412 respondents (Group 4) were interviewed in CAPI mode in Waves 2 and 3, but the interviewer was switched. Again an incentive to cheat has been introduced and no prior knowledge of network size is available to the interviewers. Network size decreases as expected from 2.519 to 2.359.

The last group consists of 2,996 cases that remained in CAPI in both waves with the same interviewer. This group is special in that an incentive to cheat has been introduced, but interviewers might still be aware of the respondent's network size from the previous year, which might make them more careful as they anticipate the survey organization's ability to check for consistency across waves. Nevertheless, the decline in the number of loops (2.455 to 2.190) is very pronounced in this group.

In a next step, the differences in all other groups will be compared to the difference for the reference group of CATI stayers (Group 1). This DiD is negative for all three groups in which a cheating incentive was introduced and positive only for CAPI to CATI changers where there never was an incentive to cheat. All differences are significant at a ten-percent level (two-sided t-tests taking clustering within interviewers into account), and only the relatively small group of CAPI stayers with different interviewers fails to reach the five-percent significance level.

Only comparing those respondents for whom interviewers never had an incentive to cheat (Groups 1 and 2) and those for whom an incentive to cheat was introduced (Groups 3 to 5), the DiD of -0.254 indicates a highly significant treatment effect which supports hypothesis 1c. The more detailed analysis in Table 2, however, shows that something else is going on in the data. CAPI networks were already smaller in Wave 2, which might be the result of a selection effect (poorly connected persons are less likely to have listed phone numbers). However, the results for the small group of respondents who switched from CAPI to CATI suggests that there might also be a mode effect on networks in the absence of incentives to cheat. We can only speculate at this point whether this might be due to differences in standardization and probing related to differences in fieldwork monitoring. As stated above, an alternative explanation for the increased network size of these mode-switchers is that the mode switch is due to changes in their life circumstances that also affected network size. Thus we should be careful about causal interpretations, as Group (2) is likely to be selective in that it contains a relatively large proportion of respondents whose life circumstances have changed.

Table 3. Empty model for CATI and CAPI

	CATI (Model 1)	CAPI (Model 2)
	β (s.e.)	β (s.e.)
Fixed part		
Person-level predictors	Not included	Not included
Interviewer-level predictors	Not included	Not included
β_1	1.900 (0.051)***	0.479 (0.091)***
Random part		
σ_b	0.394 (0.051)***	1.201 (0.081)***
ρ_{int}	0.045 (0.011)***	0.305 (0.028)***
n	7,402	5,348
Log Likelihood	−2901.773	−3179.452

Dependent variable: network-size dummy (1 = three and more contacts, 0 = less than three contacts)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.3. Results from Multilevel Models

To further investigate the mode differences found in the descriptive analysis, empty random-intercept models without any covariates were estimated for both CATI (Model 1) and CAPI (Model 2) as an initial step. The results can be found in Table 3.

In CATI there is a significant ICC of $\rho_{int} = 0.045$. This is well within the range usually found for interviewer effects (Groves 1989, 319). In contrast, in CAPI the ICC amounts to $\rho_{int} = 0.305$ which constitutes a huge interviewer effect. Interviewer effects of this size are extremely rare. Note, however, that these are ICCs from the empty model. They might at least partly reflect differences in interviewer workload. Therefore, in a next step we controlled for individual-level covariates that explain individual network size. Simultaneously we controlled for observed interviewer-level covariates. In a first step, only interviewer demographics (gender, age, and education) are included. Results can be found in Table 4.

Age in years is centered around its mean and respondents' education is coded in three categories for "low", "medium", and "high" educational background, according to the German educational system. For employment status, we created four dummy variables that comprise unemployed people, those who are still in education or military service, and those not in the labor force, whereas employed persons served as the reference category. The health measure in PASS are the so-called SF-12v2 indicators, which constitute an internationally accepted inventory of health measures (Nübling et al. 2007). Of these, the two superordinate scales "physical health" and "mental health" were generated. To account for the social contacts a person's leisure activities entail, we include a dummy variable that indicates the respondent's active engagement in any kind of club or organization. Furthermore, we generated one dummy variable indicating whether the respondent has a partner outside the household. The network-size question only refers to persons outside the respondent's household and we expect that the partner will usually be among the named persons if this applies to her or him.

The composition with respect to observed person-level variables and observable interviewer characteristics only plays a very minor role in explaining intra-interviewer correlations in network size in both modes. The ICC for CATI falls from 0.045 to 0.040,

Table 4. Models for CATI and CAPI including respondent- and interviewer-level predictors

	CATI (Model 3)	CAPI (Model 4)
	β (s.e.)	β (s.e.)
Fixed Part		
Person-level predictors		
Gender (ref. = female)	−0.383 (0.073)***	−0.291 (0.068)***
Age centered	−0.005 (0.003)	−0.007 (0.003)**
Still in school (ref. = low education level)	−0.051 (0.203)	0.289 (0.201)
Medium education level	0.317 (0.086)***	0.304 (0.083)***
High education level	0.498 (0.095)***	0.704 (0.095)***
Partner outside of household	0.542 (0.115)***	0.498 (0.108)***
Physical health	0.007 (0.004)*	0.008 (0.004)*
Mental health	0.032 (0.003)***	0.027 (0.003)***
Unemployed (ref. = employed)	−0.369 (0.090)***	−0.229 (0.087)**
Student/military service	−0.250 (0.165)	0.190 (0.163)
Retired/homemaker/parental leave	0.045 (0.120)	0.074 (0.107)
Active engagement in any organization	0.430 (0.076)***	0.589 (0.078)***
Interviewer-level predictors		
Gender (ref. = female)	0.044 (0.109)	0.207 (0.193)
Age centered	0.001 (0.005)	−0.029** (0.010)
Medium education level (ref. = low education level)	0.188 (0.335)	−0.238 (0.252)
High education level	−0.004 (0.327)	0.000 (0.274)
β_1	−0.278 (0.413)	−1.204 (0.371)**
Random Part		
σ_b	0.370 (0.051)***	1.201 (0.082)***
ρ_{int}	0.040 (0.011)***	0.305 (0.029)***
n	7,402	5,348
Log Likelihood	−2745.015	−2989.010

Dependent variable: network-size dummy (1 = three and more contacts, 0 = less than three contacts)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

while in CAPI it even remains constant at 0.305. Thus, differences in network size between interviewers can be attributed neither to their demographic characteristics nor to characteristics of the respondents they interviewed.

In spite of the huge difference in interviewer effects between the two modes, the variables explaining the network size are strikingly similar in both modes. Coefficients in Models 3 and 4 point in the same direction and are similar in size with only two exceptions, which both comprise very small numbers of cases and are not significant in any of the models. Models for both of the modes consistently show larger networks for women, medium or highly educated respondents, respondents with a partner outside their household, physically and mentally healthy respondents and smaller networks for unemployed than for employed. Age is significant only in the CAPI model, where younger respondents tend to have larger networks. However, the CATI coefficient at -0.005 is very similar to the CAPI coefficient at -0.007. Furthermore, the age of the interviewer has a negative effect on network size only in CAPI.

Turning to Hypothesis 1a, Models 3 and 4 provide evidence confirming that the interviewer effect is much larger in CAPI than in CATI after controlling for composition effects and observable interviewer characteristics. While separate models for CATI and CAPI are well suited to showing the different magnitude of the intra-interviewer correlation in CATI and CAPI, providing support for Hypothesis 1a, these models do not show that network size is downward biased in CAPI after controlling for respondent composition (Hypothesis 1b). This can be done by employing the same respondent- and interviewer-level variables as in Model 4 but including all cases from both modes and adding a CAPI indicator to measure the effect of CAPI interviews compared to CATI interviews. These results can be found in Model 5. Table 5 shows only the coefficient of the additional CAPI variable and the variance components. As expected, the CAPI effect is strongly negative and highly significant, which implies smaller networks in CAPI mode and thus supports Hypothesis 1b.

So far, we have identified an interviewer effect in CAPI that is much larger than the interviewer effect in CATI and leads to significantly smaller networks after controlling for observable person- and interviewer-level predictors. However, the larger interviewer effects in CAPI might be explained by unobserved differences in the assignment to interviewers. In centralized CATI studios, interviewer assignment to respondents is close to random within a given shift. In contrast, in CAPI interviewer workload is assigned much more selectively, as all cases interviewed by one interviewer usually reside in the same region in one or two sampling points. On average, interviewers have performed 77 percent of their workload in their main sampling point. Thus differences in true network size between regions constitute a serious potential alternative explanation for the larger ICC in CAPI mode. In addition, the strong confounding of interviewers and areas makes the use of cross-classified random-effects models (Rasbash and Goldstein 1994) for the separation of interviewer and sampling point effects unadvisable (Vassallo et al. 2016).

In contrast to previous studies, the PASS data offer a rare opportunity to control for regional differences in network size as an alternative explanation for interviewer variances. Model 6 again uses CAPI cases only. The average network size from all CATI interviews in the same sampling point is included as an additional control for regional differences in expansiveness of networks. Again, Table 5 only shows the coefficients of

Table 5. *Models with additional explanatory variables*

	CATI + CAPI (Model 5)	CAPI (Model 6)	CAPI (Model 7)	CAPI (Model 8)
	β (s.e.)	β (s.e.)	β (s.e.)	β (s.e.)
Fixed Part				
Person-level predictors	included	included	included	included
Interviewer-level predictors	included	included	included	included
CAPI	−1.057 (0.192)***			
Avg. CATI netw. size in PSU		0.059 (0.038)	0.064 (0.038)	0.066 (0.039)
Sequence			0.004 (0.004)	0.003 (0.003)
Previous wave			−0.340 (0.265)	−0.346 (0.266)
β_1	−0.368 (0.273)	−1.722 (0.498)***	−1.558 (0.538)**	−1.553 (0.545)
Random Part				
σ_b	0.938 (0.054)***	1.202 (0.082)***	1.195 (0.081)***	1.210 (0.102)***
ρ_{int}	0.211 (0.019)***	0.305 (0.029)***	0.303 (0.029)***	
$\sigma_{sequence}$				0.012 (0.005)
Corr(ζ_j, $\sigma_{sequence}$)				−0.139 (0.346)
n	12,750	5,348	5,348	5,348
Log Likelihood	−5780.530	−2987.884	−2985.913	−2985.615

Dependent variable: network-size dummy (1 = three and more contacts, 0 = less than three contacts)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

interest. The coefficient for average CATI network size in the sampling point is in the expected direction.

It is, however, not significant on a five-percent level. More importantly, inclusion of this variable has no influence on the ICC, which is still 0.305 after controlling for regional differences.[3] It should be noted, however, that the average network size per sampling unit constitutes an imperfect measure for regional differences, as a proper disentangling of interviewer and area effects would require an interpenetrated sample design (Campanelli and O'Muircheartaigh 1999). Nevertheless, results are still in favor of Hypothesis 1a after controlling for regional expansiveness of networks.

We now turn to Hypothesis 2a and 2b. Here we test whether interviewers exhibit a learning behavior over time – within the study as well as across studies. While we are able to create an exact measure for experience within the study, the number of previous interviews within the study, the survey organization did not provide an equally good measure for overall experience as an interviewer. The only available indicator is whether an interviewer already conducted interviews in previous PASS waves.[4] This measure is probably correlated only weakly with general interviewer experience. Those who carried out previous PASS interviews have been interviewing for at least one or two years. With primary sampling units being constant, interviewers who did their first PASS interview in Wave 3 will often be newly hired interviewers. Of all interviewers in PASS Wave 3, 19 percent had no experience within the study.

Model 7 in Table 5 includes both indicators for experience in addition to all variables from Model 6. Both indicators have no significant effect. The coefficient for previous wave PASS experience is negative, as expected. As the indicator is quite crude, it remains unclear whether more experience as an interviewer leads to a reduction in network size. The coefficient for the number of previous interviews within the wave (sequence) surprisingly is positive, that is, CAPI interviewers tend to produce larger networks in later interviews. Thus, there is definitely no general support for Hypothesis 2b, which presumed learning effects across time within a wave. However, interviewers might differ with respect to their reaction to incentives to cheat and in their learning behavior. It thus makes sense to relax the assumption that the slope with respect to sequence is identical for all interviewers. In the multilevel framework this can naturally be done by replacing the fixed slope for sequence by a random slope, which was done in Model 8. This random slope fails to reach the five-percent significance level when – as is advised by Rabe-Hesketh and Skrondal (2012) – we use a likelihood-ratio test between Models 8 and 7. The other coefficients are scarcely affected by the introduction of the random slope.

Models 7 and 8 thus suggest that there is neither a general trend towards smaller network sizes within the temporal sequence of interviews of one interviewer, nor a significant variation between interviewers with respect to such learning effects. This seems

[3] We also estimated a CATI model (not displayed) in which we replaced interviewers by PSUs as level 2 units. The ICC after controlling for observable person- and interviewer-level variables is only 0.013 and not statistically significant. We interpret this as further evidence that regional differences cannot be considered the reason for interviewer effects.

[4] We should recall that in previous waves the network size question did not serve as a filter and thus direct learning is not the mechanism identified by this measure.

surprising and is in contrast to earlier findings by Matschinger et al. (2005) and Kosyakova et al. (2015). One reason might be that the potential to abbreviate the interview by reporting small networks is very obvious and interviewers learn it so quickly that it does not show up in a linear trend across the whole fieldwork. We thus re-estimated Model 7 ten times using only the first five to 14 interviews for each interviewer. The slope for experience within a wave indeed changes to a negative sign in nine out of these ten models. However, it only becomes significant when taking the first nine or the first twelve interviews of each interviewer. We conclude from this that there might be some learning effects during the initial interviews of each interviewer, but that the evidence for this is in no way convincing.

Models 1 to 8 exclusively make use of Wave 3 data. While these models cannot exactly be replicated using Waves 1 and 2, as some constructs like physical or mental health have only been collected in Wave 3, estimating similar models for previous waves can provide valuable hints. If the increased interviewer effect in CAPI was exclusively driven by the incentive to cheat, then Wave 1 and 2 interviewer effects for network size in CAPI mode should be close to the size that is observed in CATI, and thus it is sufficient to focus on the ICC. Re-estimating a model similar to our Model 6 for Waves 1 and 2 (excluding health variables) and controlling for possible sample-point differences results in ICCs which are substantially higher for CAPI (0.158 in Wave 1 and 0.203 in Wave 2) than those we found for CATI, even before an incentive to abbreviate interviews was introduced. At the same time they are much smaller than the ICC of 0.303 found in Model 6 for Wave 3.

Like the difference-in-difference approach, this suggests that two mechanisms are at work at the same time. The increase in ICC from Waves 1 and 2 to Wave 3 indicates that cheating might play a role, while the difference between interviewer effects in CATI and interviewer effects in CAPI in Waves 1 and 2 indicates that other differences between CATI and CAPI interviews play a role as well. We will return to this in the discussion section.

6. Summary and Discussion

We have found large interviewer effects for a network-size filter question in a large-scale German panel survey. While there is a significant interviewer effect in CATI, it is far exceeded by the interviewer effect in CAPI. This latter effect remains identical in size when observed respondent- and interviewer-level variables are included and even when average network size in a region (measured independently) is controlled for. In contrast to earlier research on interviewer effects on filter questions (Matschinger et al. 2005; Kosyakova et al. 2015), there is no evidence for a general learning effect of the same interviewer across interviews within a wave of data collection or for interviewer-specific differences in these learning effects.

How can we interpret these findings? Past studies have attributed large interviewer effects for network generators to differences in probing behavior (van Tilburg 1998; Marsden 2003). These studies investigated complex name generators that involved a lot of probing. The network-size question under consideration here is less complex. Nevertheless, interviewer effects are even larger. Furthermore, large interviewer effects can only be found in CAPI, where interviewers have an incentive to cheat, while interviewer effects in CATI are comparably small. In addition, the findings from our DiD approach suggest that introducing an incentive for CAPI interviewers to produce smaller networks actually results in reduced network size. This leads us to the conclusion that purposeful manipulation of answers on the side of the interviewers – and not only differences in probing – is one likely explanation.

However, this study has some limitations. Interviewer effects in CAPI were much larger than in CATI in Waves 1 and 2, where no additional follow-up questions were triggered and thus manipulation on the side of the CAPI interviewers has no obvious payoff. Furthermore, a difference-in-difference analysis indicates that respondents who were switched from CAPI in Wave 2 to CATI in Wave 3 show significantly increasing network sizes, although in both cases there were no incentives to produce smaller or larger networks. This indicates that other mechanisms are at work as well. Likely candidates are probing or other deviations from standardized interviewing, which are more likely to happen in the less well supervised CAPI field.

Investigating this in detail would require more detailed information on CAPI fieldwork. The IAB, as the institution responsible for PASS, has set up a project for this purpose and recorded CAPI interviews as part of this research. This might help us to gain a better understanding of the origin of large interviewer effects in CAPI even in the absence of incentives to shorten an interview. Other possible explanations include that interviewer characteristics not controlled for in the models could be more relevant in CAPI than in CATI interviews, or that unobserved respondent characteristics lead to smaller networks in CAPI.

We used a rather simple analysis strategy that is adequate in our view as it reflects interviewer incentives. We only distinguished between networks of size three and larger and networks smaller than three. We performed several sensitivity checks that all replicate the main findings, but result in somewhat smaller ICCs. Including network size as a metric variable in a random-effects linear regression results in an ICC of 0.16, while truncating the network variable at three and using a random-effects tobit model results in an ICC of 0.24. Thus, our simple operationalization seems to grasp what creates the largest unexplained differences between interviewers.

Our findings are limited to one study in Germany. The recent publication by Brüderl et al. (2013) finds interviewer effects of a similar magnitude for a second study in Germany. This study used the same fieldwork agency and thus probably even shared interviewers with PASS. Given that the studies by Matschinger et al. (2005) and Kosyakova et al. (2015) also use German data, one might suspect that this problem is particularly prevalent in Germany. Given the results of the recent study by Paik and Sanchagrin based on the US General Social Survey it is obvious, however, that the problem is not limited to Germany. Our evidence suggests that more rigorous supervision techniques should be used in CAPI surveys to counteract incentives to shorten the interview by cheating on questions and thereby impairing question validity. Suitable techniques include re-interviews (Biemer and Stokes 1989) or statistical approaches to detect potential falsifiers (Biemer and Stokes 1989; Bredl et al. 2012). In addition, recordings of at least a considerable proportion of CAPI interviews by each interviewer could be made mandatory.

Alternatively, payment schemes might be changed so that interviewers are paid better for longer interviews. It is not easy, though, to find an increase rate that makes most interviewers indifferent to the length of the interview. If the premium for long interviews is too high, adverse incentives to the ones observed in this study might arise and interviewers might try to artificially stretch interviews. The good news of this article seems to be that in spite of falsifications by a large proportion of interviewers that lead to a pronounced downward bias of average network size, regression coefficients in models explaining network size seem to be only marginally affected. The size and direction of effects of respondent characteristics on network size is very similar between CAPI interviews and CATI interviews that offer no incentive to cheat, and both are in line with the previous literature.

The question whether this still holds for longitudinal analyses is beyond the scope of this article. The data analyzed here are from the third wave of the PASS panel. The same network-size question had been used in two prior waves without any follow-up questions. The consequence is a marked decrease in average network size from Wave 2 to 3, followed by another increase from Wave 3 to 4 where incentives to cheat diminished again as no follow-up questions were asked. It is obvious that this biases estimates of average network size across time. This is not a new finding – several authors have argued that findings about declining social capital (Putnam 1995; McPherson et al. 2006) might be artifacts (Fischer 2009; Paik and Sanchagrin 2013).

A related question is how coefficients of models for longitudinal data that use either network size as a predictor or as an outcome are affected. This topic is a matter for future research.

7. References

American Association for Public Opinion Research (AAPOR). 2003. *Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection, and Repair of its Effects*. Available at: https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf (accessed April 6th, 2016).

Bethmann, A. and D. Gebhardt. 2011. User Guide "Panel Study Labor Market and Social Security" (PASS) * Wave 3. *FDZ Datenreport, 04/2011 (en)*, Nuremberg. Available at: http://doku.iab.de/fdz/reporte/2011/DR_04-11_EN.pdf (accessed April 6th, 2016).

Biemer, P.P. and S.L. Stokes. 1989. "The Optimal Design of Quality Control Samples to Detect Interviewer Cheating." *Journal of Official Statistics* 5: 23–39.

Biemer, P.P. 2010. "Overview of Design Issues: Total Survey Error." In *Handbook of Survey Research*, 2nd ed., edited by P.V. Marsden and J.D. Wright, 27–58. Bingley: Emerald.

Blasius, J. and J. Friedrichs. 2013. "Faked Interviews." In *Methods, Theories, and Empirical Applications in the Social Sciences*, edited by S. Salzborn, E. Davidov, and J. Reinecke, 49–56. Wiesbaden: VS Verlag für Sozialwissenschaften.

Brashears, M.E. 2011. "Small Networks and High Isolation? A Reexamination of American Discussion Networks." *Social Networks* 33: 331–341. Doi: http://dx.doi.org/10.1016/j.socnet.2011.10.003.

Bredl, S., P. Winker, and K. Koetschau. 2012. "A Statistical Approach to Detect Interviewer Falsification of Survey Data." *Survey Methodology* 38: 1–10.

Brüderl, J., B. Huyer-May, and C. Schmiedeberg. 2013. "Interviewer Behavior and the Quality of Social Network Data." In *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention*, edited by P. Winkler, R. Porst, and N. Menold, 147–160. Frankfurt: Peter Lang.

Büngeler, K., M. Gensicke, J. Hartmann, R. Jäckle, and N. Tschersich. 2010. *IAB-Haushaltspanel im Niedrigeinkommensbereich Welle 3 (2008/09): Methoden- und Feldbericht. FDZ Methodenreport, 10/2010 (de)*, Nuremberg. Available at: http://doku.iab.de/fdz/reporte/2010/MR_10-10.pdf (accessed April 6th, 2016).

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

Cleveland, W.S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplot." *Journal of the American Statistical Association* 74: 829–836.

Cragg, J.G. 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica* 39: 829–844. Doi: http://dx.doi.org/10.2307/1909582.

Crespi, L.P. 1945. "The Cheater Problem in Polling." *Public Opinion Quarterly* 9: 431–445.

Eckman, S., F. Kreuter, A. Jäckle, A. Kirchner, S. Presser, and R. Tourangeau. 2014. "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys." *Public Opinion Quarterly* 78: 721–733. Doi: http://dx.doi.org/10.1093/poq/nfu030.

Fischer, C.S. 2009. "The 2004 GSS Finding of Shrunken Social Networks: An Artifact?" *American Sociological Review* 74: 657–669. Doi: http://dx.doi.org/10.1177/000312240907400408.

Freeman, J. and E.W. Butler. 1976. "Some Sources of Interviewer Variance in Surveys." *Public Opinion Quarterly* 40: 79–91. Doi: http://dx.doi.org/10.1086/268-269.

Goldstein, H. 1986. "Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares." *Biometrika* 73: 43–56. Doi: http://dx.doi.org/10.1093/biomet/73.1.43.

Groves, R.M. and N.H. Fultz. 1985. "Gender Effects among Telephone Interviewers in a Survey of Economic Attitudes." *Sociological Methods Research* 14: 31–52. Doi: http://dx.doi.org/10.1177/0049124185014001002.

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

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2004. *Survey Methodology*. Hoboken, NJ: Wiley.

Guterbrock, T.M. 2008. "Falsifications." In *Handbook of Survey Research*, edited by P.J. Lavrakas, 267–270. Los Angeles: Sage.

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

Hughes, A., J. Chromy, K. Giacoletti, and D. Odom. 2002. "Impact of Interviewer Experience on Respondent Reports of Substance Use." In *Redesigning an Ongoing National Household Survey*, edited by J. Gfroerer, J. Eyerman, and J. Chromy, 161–184. Washington: Substance Abuse and Mental Health Services Administration.

Kosyakova, Y., J. Skopek, and S. Eckman. 2015. "Do Interviewers Juggle Filter Questions? Evidence from a Multilevel Approach". *International Journal of Public Opinion Research* 27: 417–431. Doi: http://dx.doi.org/10.1093/ijpor/edu027.

Kreuter, F., S. McCulloch, S. Presser, and R. Tourangeau. 2011. "The Effects of Asking Filter Questions in Interleafed Versus Grouped Format." *Sociological Methods & Research* 40: 88–104. Doi: http://dx.doi.org/10.1177/0049124110392342.

Lechner, M. 2011. "The Estimation of Causal Effects by Difference-In-Difference Methods." *Foundations and Trends in Econometrics* 4: 165–224.

Longford, N.T. 1993. *Random Coefficient Models*. Oxford: Oxford University Press.

Mangione, T.W., F.J. Fowler, and T.A. Louis. 1992. "Question Characteristics and Interviewer Effects." *Journal of Official Statistics* 8: 293–307.

Marsden, P.V. 2003. "Interviewer Effects in Measuring Network Size Using a Single Name-Generator." *Social Networks* 25: 1–16. Doi: http://dx.doi.org/10.1016/S0378-8733(02)00009-6.

Matschinger, H., S. Bernert, and M.C. Angermeyer. 2005. "An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview." *Journal of Official Statistics* 21: 657–674.

McPherson, M., L. Smith-Lovin, and M.E. Brashears. 2006. "Social Isolation in America: Changes in Core Discussion Networks over Two Decades." *American Sociological Review* 71: 353–375. Doi: http://dx.doi.org/10.1177/000312240607100301.

Nübling, M., H.H. Andersen, A. Mühlbacher, J. Schupp, and G.G. Wagner. 2007. "Computation of Standard Values for Physical and Mental Health Scale Scores Using the SOEP Version of SF12v2." *Schmollers Jahrbuch: Journal of Applied Social Science Studies* 127: 171–182. Available at: https://www.researchgate.net/publication/23645941_Computation_of_Standard_Values_for_Physical_and_Mental_Health_Scale_Scores_Using_the_SOEP_Version_of_SF12v2 (accessed April 6th, 2016).

O'Connell, A.A. 2006. *Logistic Regression Models for Ordinal Response Variables*. Thousand Oaks, CA: Sage.

O'Muircheartaigh, C. and P. Campanelli. 1998. "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 161: 63–77. Doi: http://dx.doi.org/10.1111/1467-985X.00090.

Paik, A. and K. Sanchagrin. 2013. "Social Isolation in America: An Artifact." *American Sociological Review* 78: 339–360. Doi: http://dx.doi.org/10.1177/0003122413482919.

Pinheiro, J.C. and D.M. Bates. 1995. "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." *Journal of Computational and Graphical Statistics* 4: 12–35. Doi: http://dx.doi.org/10.1080/10618600.1995.10474663.

Pinheiro, J.C. and E.C. Chao. 2006. "Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models." *Journal of Computational and Graphical Statistics* 15: 58–81. Doi: http://dx.doi.org/10.1198/106186006X96962.

Putnam, R.D. 1995. "Bowling Alone: America's Declining Social Capital." *Journal of Democracy* 6: 65–78. Doi: http://dx.doi.org/10.1353/jod.1995.0002.

Rabe-Hesketh, S. and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata. Volume II: Categorical Responses, Counts, and Survival*. College Station, TX: Stata Press.

Rasbash, J. and H. Goldstein. 1994. “Efficient Analysis of Mixed Hierarchical and Cross-Classified Random Structures Using a Multilevel Model.” *Journal of Educational and Behavioral Statistics* 19: 337–350. Doi: http://dx.doi.org/10.3102/10769986019004337.

Schnell, R. 1991. “Der Einfluß gefälschter Interviews auf Survey-Ergebnisse.” *Zeitschrift für Soziologie* 20: 25–35.

Schnell, R. 2012. *Survey-Interviews: Methoden standardisierter Befragungen*. Wiesbaden: VS Verlag.

Schnell, R. and F. Kreuter. 2000. “Untersuchungen zur Ursache unterschiedlicher Ergebnisse sehr ähnlicher Viktimisierungssurveys.” *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 52: 96–117. Doi: http://dx.doi.org/10.1007/s11577-000-0005-y.

Schnell, R. and F. Kreuter. 2005. “Separating Interviewer and Sampling-Point Effects.” *Journal of Official Statistics* 21: 389–410.

Schraepler, J.P. and G.G. Wagner. 2005. “Characteristics and Impact of Faked Interviews in Surveys – An Analysis of Genuine Fakes in the Raw Data of SOEP.” *Allgemeines Statistisches Archiv* 89: 7–20. Doi: http://dx.doi.org/10.1007/s101820500188.

Schuman, H. and J. Converse. 1971. “The Effects of Black and White Interviewers on Black Responses in 1968.” *Public Opinion Quarterly* 35: 44–68. Doi: http://dx.doi.org/10.1086/267866.

Snijders, T.A.B. and R. Bosker. 2000. *Multilevel Analysis*. London: Sage.

StataCorp. 2011. *Stata. Longitudinal-Data/Panel-Data Reference Manual*. Release 12. College Station, TX: StataCorp.

Swamy, P.A.V.B. 1971. *Statistical Inference in a Random Coefficient Model*. New York: Springer.

Tourangeau, R. and T. Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133: 859–883.

Tourangeau, R., F. Kreuter, and S. Eckman. 2012. “Motivated Underreporting in Screening Interviews.” *Public Opinion Quarterly* 76: 453–469. Doi: http://dx.doi.org/10.1093/poq/nfs033.

Tourangeau, R., F. Kreuter, and S. Eckman. 2013. Motivated Misreporting: Shaping Answers to Reduce Survey Burden. In *Survey Measurement: Techniques and Findings from Recent Research*, edited by U. Engel, 24–41, Frankfurt: Campus.

Trappmann, M., S. Gundert, C. Wenzig, and D. Gebhardt. 2010. “PASS: a Household Panel Survey for Research on Unemployment and Poverty.” *Schmollers Jahrbuch. Journal of Applied Social Science Studies* 130: 609–622 Doi: http://dx.doi.org/10.3790/schm.130.4.609.

Trappmann, M., J. Beste, A. Bethmann, and G. Müller. 2013. “The PASS Panel Survey After Six Waves.” *Journal for Labour Market Research* 46: 275–281. Doi: http://dx.doi.org/10.1007/s12651-013-0150-1.

van der Zouwen, J., W. Dijkstra, and J.H. Smit. 2004. “Studying Respondent-Interviewer Interaction: The Relationship Between Interviewing Style, Interviewer Behavior, and Response Behavior.” In *Measurement Errors in Surveys*, edited by P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman, 419–437. New York: Wiley.

van Tilburg, T.G. 1998. "Interviewer Effects in the Measurement of Personal Network Size. A Non-Experimental Study." *Sociological Methods and Research* 26: 300–328. Doi: http://dx.doi.org/10.1177/0049124198026003002.

Vassallo, R., G.B. Durrant, and P.W.F. Smith. 2016. Separating Interviewer and Area Effects Using a Cross-Classified Multilevel Logistic Model: Simulation Findings and Implications for Survey Designs. Submitted manuscript (available from the author on request: g.durrant@southampton.ac.uk).

West, B.T., F. Kreuter, and U. Jaenichen. 2013. "Interviewer Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?" *Journal of Official Statistics* 29: 277–297. Doi: http://dx.doi.org/10.2478/jos-2013-0023.

Received August 28, 2014
Revised June 27, 2015
Accepted September 8, 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 375–404, http://dx.doi.org/10.1515/JOS-2016-0021

The FEWS Index: Fixed Effects with a Window Splice

Frances Krsinich[1]

This article describes the estimation of quality-adjusted price indexes from ‘big data’ such as scanner and online data when there is no available information on product characteristics for explicit quality adjustment using hedonic regression. The longitudinal information can be exploited to implicitly quality-adjust the price indexes. The fixed-effects (or ‘time-product dummy’) index is shown to be equivalent to a fully interacted time-dummy hedonic index based on all price-determining characteristics of the products, despite those characteristics not being observed. In production, this can be combined with a modified approach to splicing that incorporates the price movement across the full estimation window to reflect new products with one period’s lag without requiring revision. Empirical results for this fixed-effects window-splice (FEWS) index are presented for different data sources: three years of New Zealand consumer electronics scanner data from market-research company GfK; six years of United States supermarket scanner data from market-research company IRI; and 15 months of New Zealand consumer electronics daily online data from MIT’s Billion Prices Project.

Key words: Big data; scanner data; online data; hedonic regression; quality adjustment.

1. Introduction

Ensuring that consumer price indexes reflect only price change is important, and is traditionally achieved by pricing a fixed basket of goods over time. However, products such as consumer electronics are rapidly evolving, and have correspondingly shorter life cycles in the market. This makes the traditional approach of pricing a representative fixed basket challenging.

The potential for using so-called ‘big data’ such as scanner data or online data to measure price change with higher accuracy and frequency makes the challenge of rapid product change for price measurement increasingly pertinent. Price indexes in the context of official statistics need to use the data that are available, and ideally their estimation should be automatable.

Price measurement using scanner data has been an active area of research for Statistics New Zealand over the last five years. Collaborative research with Statistics Netherlands

[1] Statistics New Zealand – Prices Unit, PO Box 2922 Wellington 6041, New Zealand. Email: frances.krsinich@stats.govt.nz

Acknowledgments: The author would like to thank Chris Pike, manager of the Prices Unit at Statistics New Zealand, for his much-appreciated support and encouragement of this research; Richard Arnold, Alistair Gray, Jan de Haan and three anonymous referees for helpful discussion and comments; and Alberto Cavallo for sharing online data from the Billion Prices Project.

Disclaimer: The opinions, findings, recommendations, and conclusions expressed in this article are those of the author. They do not necessarily represent those of Statistics NZ, which takes no responsibility for any omissions or errors in the information contained here.

(de Haan and Krsinich 2014) on a new method called the imputation Törnqvist RYGEKS (ITRYGEKS) has established it as benchmark index which appropriately quality adjusts, both for the changing quality mix of products being bought and for the implicit price movements of new products entering, and old products disappearing, from the market.

Supermarket scanner data and online data, however, do not contain sufficient information on characteristics to use methods such as the ITRYGEKS or the multilateral time-dummy (TD) hedonic index. This has led to a revisiting of the fixed-effects index (also known as the time-product dummy (TPD) index) that was used to benchmark the New Zealand housing rentals index (Krsinich 2011b). ‘Fixed effects’ refers to the fitting of product-specific identifiers. This can be seen as the model controlling for the bundles of characteristics corresponding to each product, rather than controlling directly for the characteristics.

This article demonstrates, theoretically and empirically, that the fixed-effects index is equivalent to a fully interacted time-dummy hedonic index based on all characteristics that are constant across time at the barcode, or detailed product-specification, level. Note that this does not advocate the explicit use of fully interacted hedonic models in practice, rather it demonstrates what is implicitly achieved by the fixed-effects model.

It is also shown empirically that this fully interacted time-dummy hedonic index is virtually the same as the more common main-effects time-dummy hedonic index when a comprehensive set of price-determining characteristics are available for inclusion in the main-effects hedonic model.

Because the fixed-effects index requires at least two price observations before a new product is nontrivially incorporated into the estimation, a modified approach to the ‘splicing’ or updating of the price indexes in production is required. This window-splice approach uses the movement across the entire estimation window, rather than just the movement of the most recent period, thereby incorporating a catch-up revision factor reflecting the implicit price movements associated with the introduction of new products in the period after their introduction along with an updating of the predicted fixed effects for all products observed. The combination of a fixed-effects index estimation with the window splicing is referred to as the ‘fixed-effects window-splice’ (FEWS) index.

So, if there is longitudinal price and quantity information at a detailed product-specification level, the FEWS approach can be used to produce nonrevisable quality-adjusted price indexes.

The article is structured as follows:

Section 2 gives the background to the development of the FEWS index – from benchmarking the New Zealand housing rentals index to evaluating this approach in the context of both supermarket scanner data and online data, and then incorporating fixed-effects indexes for mobile phones and televisions in the New Zealand import price index from the December 2013 quarter.

In Section 3, a result from de Haan and Hendriks (2013) for the time-dummy hedonic index is restated in terms of categorical characteristics. This is then extended to include interaction terms, showing that the fixed-effects index is algebraically equivalent to a fully interacted time-dummy hedonic index that explicitly incorporates all characteristics of the products.

In Section 4, a simulation is used to give some intuition for the equivalence of the fully interacted time-dummy hedonic index and the fixed-effects indexes by basing the product identifiers used in the fixed-effects index on the same cut-down set of characteristics included in the time-dummy hedonic index.

Section 5 describes how the window-splice incorporates a catch-up revision factor along with the most recent period's movement to maintain the integrity of the long-term index, in particular by revising for the implicit price movements of new products entering the market in the period after their introduction.

In Section 6, empirical results for the FEWS index are compared to other indexes for a range of data sources:

- New Zealand monthly aggregated consumer electronics scanner data from market-research company GfK, with quantities and comprehensive sets of characteristic information for each product category.
- United States weekly aggregated supermarket scanner data from IRI marketing, with quantities sold but no characteristic information. For manageability, the data is aggregated to a monthly level for the analysis.
- New Zealand daily consumer electronics web-scraped online data from the Billion Prices Project, with no characteristics or quantities (Billion Prices Project, 2016).

Section 7 concludes the article.

2. Background

A fixed-effects approach was used by Statistics New Zealand to retrospectively benchmark the performance of the current matched-sample approach to measuring the price movement of housing rentals. Due to the lack of sufficient characteristics in the housing rental survey data, the longitudinal nature of the data was exploited by fitting a fixed-effects model – that is, the fitting of dwelling-specific intercepts – to implicitly control for all time-invariant characteristics of the surveyed rental dwellings. As pointed out by de Haan (2015a), this is conceptually very similar to the quality-adjustment approach of 'overlap pricing', though it is based on the price movements of all matched products, rather than just one product being replaced.

This approach was controversial at the time, and so Krsinich (2011b) extended a result from Aizcorbe et al. (2003) to show that the implicit price movement being estimated for new rental dwellings by the retrospective fixed-effects index was appropriate.

De Haan and Krsinich (2014) presented and empirically tested the imputation Törnqvist rolling year GEKS (ITRYGEKS) index. This is an extension of the rolling year GEKS (RYGEKS) of Ivancic et al. (2011).

The GEKS index takes the geometric average of all bilateral superlative indexes (e.g., Törnqvist or Fisher indexes) within an estimation window. A rolling window is used with the GEKS index to make it nonrevisable for production, with the latest month's movement spliced onto the previous index number each month. This results in the rolling year GEKS (RYGEKS) index. The window length is usually just over one year, that is, 13 months for a monthly index.

In contrast to the RYGEKS, the ITRYGEKS is based on bilateral time-dummy hedonic indexes rather than superlative indexes. This enables the ITRYGEKS index to reflect the implicit price movements of new and disappearing products on entry and exit, which the RYGEKS does not do.

It was noted as an interesting result in de Haan and Krsinich (2014) that the rolling year time-product dummy (RYTPD) index – that is, the fixed-effects index – tended to sit closer to the benchmark ITRYGEKS than did the RYGEKS, despite utilising only product identifiers rather than explicitly incorporating information on characteristics.

Krsinich (2013) modified the fixed-effects index by splicing on the movement across the entire window rather than just the movement of the most recent period. This further improved the performance of the fixed-effects index, as measured against the benchmark ITRYGEKS.

Also in Krsinich (2013), it was first noted that the time-product dummy index appeared to be equivalent to a fully interacted time-dummy hedonic model and, therefore, might in a way be more comprehensively quality-adjusting than the ITRYGEKS, which is based on main-effects time-dummy hedonic bilateral indexes.

De Haan and Hendriks (2013) explored the use of fixed-effects indexes for producing high-frequency price indexes from online data, deriving expressions for both the time-dummy hedonic and fixed-effects indexes. However, implausible empirical results for women's t-shirts, along with their observation that "measuring quality-adjusted price indexes without information on item characteristics is just not possible", led the authors to conclude that the fixed-effects index is not appropriate for products "where quality change is important".

One hypothesis about this apparent bias in the women's t-shirts index is that, because fashion is highly seasonal, there was probably an almost complete replacement of old products with new products between seasons. The only matched products on which to base the fixed-effects estimation would be products undergoing heavy discounting. See Appendix 2 for the intuition behind this. This suggests that a condition for using the fixed-effects index might be that there is at least some minimum percentage of products that are matched, and not undergoing seasonal discounting, between any two successive periods. Alternatively, the regression assumption of constant coefficients within the estimation window may be violated for these very seasonal products.

In the December 2013 quarter, Statistics NZ introduced fixed-effects indexes into the production of the New Zealand import price index for both mobile phones and televisions. For the major brands of these two products, comprehensive data on import prices and quantities is available longitudinally at a detailed product level. Because the previous quarter of the New Zealand import price index is revisable, the window-splicing aspect of the FEWS index is not required for the incorporation of new products' price movements into the index, though it may be incorporated into production in the future to improve the estimation.

3. Theory

In this section, results from de Haan and Hendriks (2013) are extended to show that the fixed-effects index is equivalent to a fully interacted time-dummy hedonic index where all characteristics that define products are treated as categorical. "Fully interacted" means that the regression equation includes the interaction terms for all possible combinations of main effects.

In the case of scanner data, where barcodes change with any change in a price-determining characteristic, this means the fixed-effects index (with barcodes, or their equivalent, as the product identifiers) is equivalent to a fully interacted time-dummy hedonic index where all price-determining characteristics are explicitly included in the hedonic model.

3.1. The Main-Effects Time-Dummy Hedonic Index

The estimating equation for the (unweighted) time-dummy hedonic index can be stated as:

$$\ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t \tag{1}$$

where p_i^t is the price of product i in period t; D_i^t is a dummy variable with the value 1 for period t and 0 otherwise; z_{ik} is the quantity of characteristic k and ε_i^t is the error term.

The time-dummy hedonic index is then derived from the estimated parameters on time as follows:

$$P_{TD}^{0,t} = \exp(\hat{\delta}^t) \tag{2}$$

Since Equation (1) includes only main effects for the z characteristics, we make this explicit by referring to its corresponding index as the 'main-effects time-dummy' hedonic index $P_{TD(ME)}$.

From Equation (3) in de Haan and Hendriks (2013) we obtain:

$$P_{TD(ME)}^{0t} = \exp(\hat{\delta}^t) = \frac{\prod_{i \in S^t} \left(p_i^t\right)^{\frac{1}{N^t}}}{\prod_{i \in S^0} \left(p_i^0\right)^{\frac{1}{N^0}}} \exp\left[\sum_{k=1}^{K} \hat{\beta}_k \left(\bar{z}_k^0 - \bar{z}_k^t\right)\right] \tag{3}$$

where S^0 and S^t are the set of products available in times 0 and t respectively and N^t is the number of items in time t.

They also extend the formulation to the weighted case. For simplicity, this theory section is restricted to the unweighted case, but the simulation in Section 4 shows the equivalence of the two indexes in the weighted case.

The exponential factor is the quality-adjustment factor, which adjusts the ratio of geometric mean prices for any changes in the average characteristics of products between period t and the base period 0.

Any change in a characteristic in scanner data will correspond to a change in barcode. In online data, a change in characteristic will correspond to a changed product identifier, or model name. So the set of values that characteristics can take is discrete, as a consequence of the set of barcodes or product identifiers being discrete.

This means that when estimating price indexes from scanner or online data, all characteristics can be treated as categorical. Characteristics correspond to a discrete set of product specifications. That is, even numeric characteristics such as 'screen size' for computers or 'number of pixels' for cameras will take a discrete set of values across the set of product specifications.

A convenient feature of treating all characteristics as categorical is that no parametric form is imposed on the hedonic models.

Note also that, rather than building predictive models, these models are conditioning price-determining characteristics out of the parameters estimated for time, from which the price index is derived. The parameters estimated on the characteristics other than time are of no interest in themselves. Nevertheless the parameter estimates may be useful as regression diagnostics.

With all the characteristics being modelled as categorical, (3) can be restated as follows.

$$P^{0t}_{TD(ME)} = \exp(\hat{\delta}^t) = \frac{\prod_{i \in S^t} (p^t_i)^{\frac{1}{N^t}}}{\prod_{i \in S^0} (p^0_i)^{\frac{1}{N^0}}} \exp\left[\sum_{l=1}^{L} \hat{\beta}_l \left(\bar{D}^0_l - \bar{D}^t_l\right)\right] \tag{4}$$

where L is the total number of categories across the k characteristics (minus a base category for each of the k characteristics) and $\bar{D}^t_l$ is the proportion of $i \in S^t$ with the characteristic D_l.

3.2. *The Fully Interacted Time-Dummy Hedonic Index*

Following the same reasoning as in de Haan and Hendriks (2013), an equivalent expression to (4) can be derived for the fully interacted time-dummy hedonic index, that is, the index derived from a hedonic model that includes the main effects and all the interactions between characteristics. For n characteristics this would be the main effects, the two-way interactions, the three-way interactions, and so on up to the n-way interactions.

For simplicity, and without loss of generality, products with just two characteristics are considered here. For example, characteristics A and B, each of which has three categories *1*, *2*, and *3*. The four main effects correspond to the dummy variables for each of *A1*, *A2*, *A3*, *B1*, *B2*, and *B3* (less the two base categories for each of characteristics A and B). The eight second-order interactions are the dummy variables for each of *A1B1*, *A1B2*, . . ., *A3B3* (less one base category).

The estimating equation for the full hedonic model can be written:

$$\ln p^t_i = \delta^0 + \sum_{t=1}^{T} \delta^t D^t_i + \sum_{l=1}^{L} \beta_l D_{il} + \sum_{m=1}^{M} \beta_m D_{im} + \varepsilon^t_i \tag{5}$$

where p^t_i is the price of item i in period t; D_{il} are the dummy variables for the L main effects and D_{im} are the dummy variables for the M second-order interactions, with β_l and β_m the corresponding parameters; δ^0 is the intercept; δ^t are the time-dummy parameters (from which the index is derived); and ε^t_i are the random errors.

Appendix 1 follows the approach of de Haan and Hendriks (2013) to give the full derivation from (5) of the fully interacted time-dummy hedonic index shown in Equation (6).

$$\begin{aligned} P^{0t}_{TD(full)} &= \exp(\hat{\delta}^t) \\ &= \frac{\prod_{i \in S^t} (p^t_i)^{\frac{1}{N^t}}}{\prod_{i \in S^0} (p^0_i)^{\frac{1}{N^0}}} \exp\left[\sum_{l=1}^{L} \hat{\beta}_l \left(\bar{D}^0_l - \bar{D}^t_l\right) + \sum_{m=1}^{M} \hat{\beta}_m \left(\bar{D}^0_m - \bar{D}^t_m\right)\right] \end{aligned} \tag{6}$$

This is the time-dummy hedonic index with quality adjustment for the change in characteristics in terms of not only the main effects, but all the interactions of characteristics. So the quality adjustment is more comprehensive than that of the main-effects time-dummy index of equation (4).

3.3. The Fixed-Effects Index

Similarly to the case of the time-dummy hedonic index, de Haan and Hendriks (2013) show that, with an estimating equation for the (unweighted) fixed-effects index of

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t \tag{7}$$

where D_i is a dummy variable that has the value of 1 if the observation relates to item i and 0 otherwise, the fixed-effects index can be formulated as follows:

$$P_{FE}^{0t} = \exp(\hat{\delta}^t) = \frac{\prod_{i \in S^t} (p_i^t)^{\frac{1}{N^t}}}{\prod_{i \in S^0} (p_i^0)^{\frac{1}{N^0}}} \exp\left[\bar{\hat{\gamma}}^0 - \bar{\hat{\gamma}}^t\right] \tag{8}$$

where $\bar{\hat{\gamma}}^0 = \sum_{i \in S^0} \hat{\gamma}_i / N^0$ and $\bar{\hat{\gamma}}^t = \sum_{i \in S^t} \hat{\gamma}_i / N^t$ are the sample means of the estimated fixed effects.

The exponential terms in (6) and (8) are the same, because the fixed effect for any item i is, by definition, the net effect of the parameters corresponding to the 'bundle' of characteristics belonging to i, where the characteristics are constant across time at the item, or product, level. That is, the fixed effect is the sum of the parameters on the main effects and all the interactions for the characteristics of the product.

To illustrate using the example given in Subsection 3.2, consider an item i with the characteristics *A3* and *B1* in time t.

In the fully interacted time-dummy hedonic index formulation of Equation (6), this item will contribute $\hat{\beta}_{A3} + \hat{\beta}_{B1} + \hat{\beta}_{A3B1}$ to the expression in the brackets which, when exponentiated, is the quality-adjustment factor.

In the fixed-effects index formulation of Equation (8), the item contributes a fixed effect of $\hat{\gamma}_i$ to the corresponding bracketed term.

Since, by definition, the fixed effect is the net effect of the characteristics of the item, $\hat{\gamma}_i = \hat{\beta}_{A3} + \hat{\beta}_{B1} + \hat{\beta}_{A3B1}$, and it follows that Equation (6) = Equation (8).

So, the fixed-effects index is equivalent to the fully interacted time-dummy hedonic index where all price-determining, and therefore product-identifying, characteristics are included in the time-dummy hedonic index.

Price indexes should appropriately reflect the relative expenditures on different products. The following section demonstrates the equivalence of the fixed-effects and fully interacted time-dummy hedonic indexes empirically for the weighted case (using expenditure shares as weights) by using a subset of three characteristics to both include in the fully interacted time-dummy hedonic index and define the product identifiers used in the fixed-effects index.

4. Simulation

Scanner data from market-research company GfK is used for digital cameras for the three years from July 2008 to June 2011.

To run a fully interacted time-dummy hedonic model, just three of the approximately 40 available characteristics for digital cameras are incorporated into the model.

These characteristics are brand (with 22 categories), depth of device in millimetres (295 categories), and photos per second (78 categories).

To estimate the corresponding fixed-effects index, product identifiers are defined based on these three characteristics.

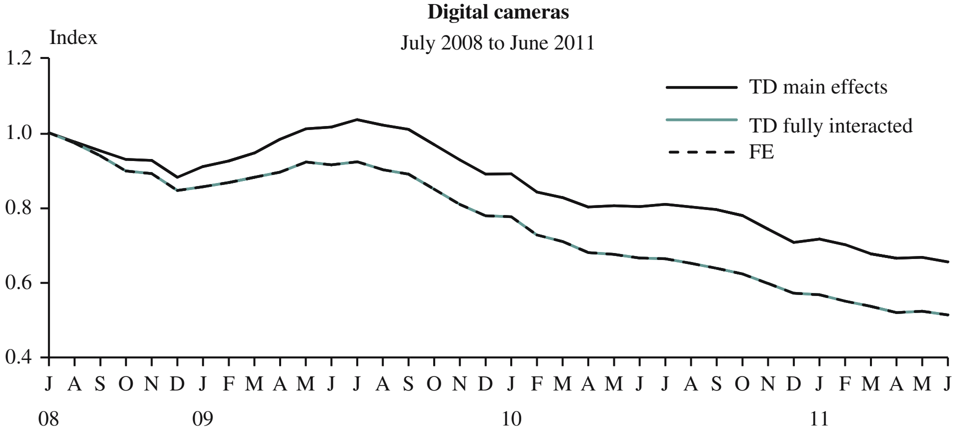
For example, a product of brand *A*, with a depth of device in millimetres of twelve, and photos per second of 100, could be assigned an identifier of 'A_12_100'. A product of brand *B*, depth of device in millimetres of 25, and photos per second of 200, could be called 'B_25_200'. Note that these product identifiers are text strings, and therefore convey no information about the corresponding characteristics to the estimation process.

Figure 1 shows the resulting indexes. The fully interacted time-dummy hedonic index (TD fully interacted) and the fixed-effects index (FE) are precisely the same, despite no characteristics being explicitly incorporated into the fixed-effects estimation.

We also show the standard time-dummy hedonic index, which includes only the main effects for the three included characteristics. The difference between this main-effects time-dummy hedonic index and the fully interacted time-dummy hedonic index shows there is extra price-determining information in the interactions between these three characteristics that is not reflected in the parameters on the main effects. It is likely that particular combinations of characteristics correspond to particular models of camera, and therefore the interactions are indirectly reflecting the price-effects of other characteristics related to those models.

Note that these indexes are estimated on the full set of data – that is, they are not the rolling-window versions which would be used in production to produce nonrevisable indexes, as discussed below in Section 5.

As the set of characteristics included in the main-effects time-dummy hedonic index increases, the difference between the main-effects and fully interacted time-dummy hedonic indexes decreases. This is shown in Subsection 6.1 by the very close correspondence between the time-dummy movement-splice (TDMS) index and the fixed-effects movement-splice (FEMS) indexes for eight consumer electronics products.



Source: GfK

Fig. 1. Comparison of fully interacted time-dummy hedonic and fixed-effects indexes

5. The Window Splice

Consumer price indexes need to be nonrevisable. This is usually achieved by incorporating a 'rolling window' for the set of data used to estimate the index, and splicing on the most recent period's movement to the previously published index number.

Krsinich (2013) proposed a modified approach to the splicing that uses the movement across the entire estimation window rather than just the most recent period's movement. This 'window splice' is a simplified version of a suggestion for improving the splicing of the RYGEKS made by Melser (2011).

Appendix 2 gives the intuition for why the fixed-effects model needs at least two price observations to include a new product nontrivially in the estimation. It also shows how the fixed-effects estimates are improved with more price observations for each product.

The window splice enables a form of implicit revision, which maintains the integrity of the index over the longer term, not only incorporating the implicit price movements of new products being introduced, but also enabling the updating of the fixed-effects estimates as more prices are observed for each product.

Consider a situation where the index over a seven-quarter period is based on three successive five-quarter estimation windows. We use a quarterly, rather than monthly, index to simplify the example.

In practice, when estimating time-dummy hedonic indexes, the index will generally change only slightly for previous periods with each subsequent window's estimations. For the illustration purpose, we here show an extreme situation of quite significant change to the index with each subsequent estimation window.

In the example shown in Figure 2, the index estimated from the first-quarter window of data has a constant increase of ten percent per quarter, the index estimated on window two is steeper, with a 15 percent per quarter increase, and the third window's index reflects a 20 percent increase in each quarter.

The standard approach of splicing on just the most recent periods – the 'movement splice' – would result in the index showing a ten percent increase in each of the first five quarters from Y1Q1 to Y2Q1, then 15 percent from Y2Q1 to Y2Q2, and then 20 percent from Y2Q2 to Y2Q3. This is shown in Figure 3.

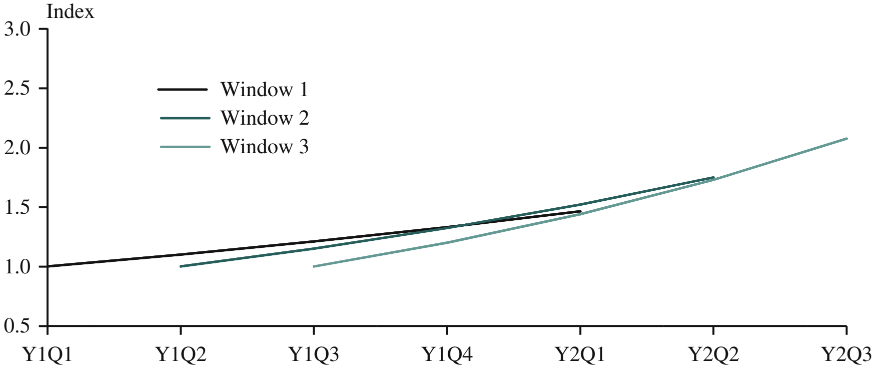


Fig. 2. Indexes estimated from three successive five-quarter windows

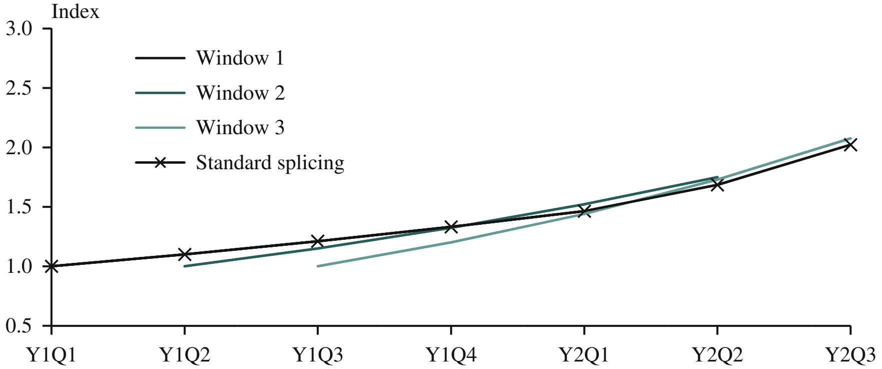


Fig. 3. Splicing on the most recent period's movement

Extending the notation introduced by de Haan (2015b), we can express the movement-spliced index for period $T + t$ (where the estimation window is of length $T + 1$ and $t > 0$) as follows:

$$P_{MS}^{0,T+t} = P_{[0,T]}^{0,T} \times \frac{P_{[1,T+1]}^{1,T+1}}{P_{[1,T+1]}^{1,T}} \times \frac{P_{[2,T+2]}^{2,T+2}}{P_{[2,T+2]}^{2,T+1}} \times \ldots \times \frac{P_{[t,T+t]}^{t,T+t}}{P_{[t,T+t]}^{t,T+t-1}} \qquad (9)$$

where $P_{[c.d]}^{a,b}$ is the index from period a to period b, estimated using data from period c to period d.

So, for the seven-quarter example shown above in Figure 3, the index for the last period Y2Q3 will be:

$$P_{MS}^{0,6} = P_{[0,4]}^{0,4} \times \frac{P_{[1,5]}^{1,5}}{P_{[1,5]}^{1,4}} \times \frac{P_{[2,6]}^{2,6}}{P_{[2,6]}^{2,5}}$$

The problem with this approach is that the revised movement for previous periods is not incorporated into the longer-term index movement. This could result in a bias, likely to accumulate over time, which we can refer to as 'splice drift'.

Another approach to the splicing would be to incorporate the movement across the entire estimation window, so that the longer-term index always reflects the most up-to-date estimation of not only the most recent period's movement, but the entire estimation window.

If we were able to revise the index, then this splicing of the most recent window's index each time would be as shown in Figure 4. The revised index shows a ten percent increase from Y1Q1 to Y1Q2, a 15 percent increase from Y1Q2 to Y1Q3, and 20 percent quarterly increases between Y1Q3 and Y2Q3. So, at all points the index is based on the most recent estimation window available.

We can formulate this 'revisable-index splicing' as follows:

$$P_{RS}^{0,T+t} = P_{[0,T]}^{0,1} \times P_{[1,T+1]}^{1,2} \times \ldots \times P_{[t,T+t]}^{t,T+t} \qquad (10)$$

which, for the last period Y2Q3 in the example shown in Figure 4, will be

$$P_{RS}^{0,6} = P_{[0,4]}^{0,1} \times P_{[1,5]}^{1,2} \times P_{[2,6]}^{2,6}$$

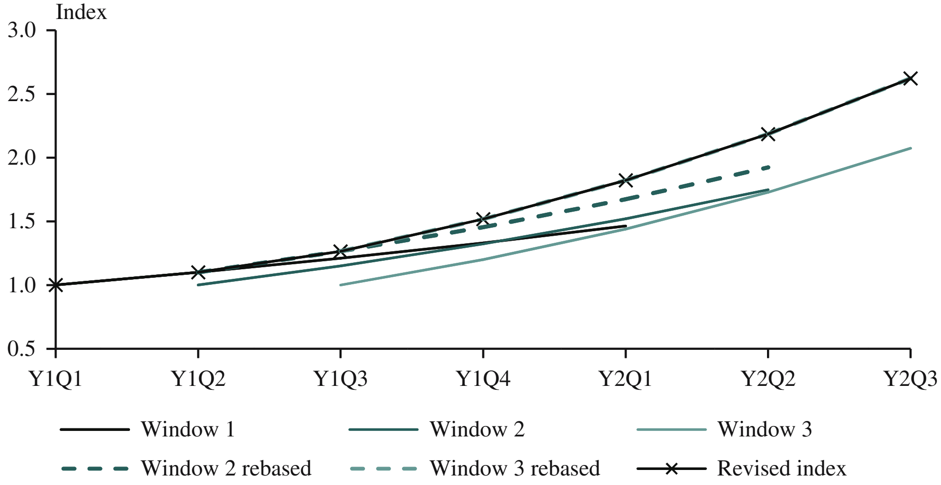


Fig. 4. Revising the index with the most recent window's estimation

And in this quarter, the previous quarter's index can be revised from

$$P_{RS}^{0,5} = P_{[0,4]}^{0,1} \times P_{[1,5]}^{1,5}$$

to

$$P_{RS(rev1)}^{0,5} = P_{[0,4]}^{0,1} \times P_{[1,5]}^{1,2} \times P_{[2,6]}^{2,5}$$

However, consumer price indexes are generally unrevisable.

Instead, a 'catch-up' factor can be incorporated into the most recent period's movement, to maintain the index at the level it would be if revision had been possible. This is the approach taken with the 'window-splice' of the FEWS index, illustrated in Figure 5.

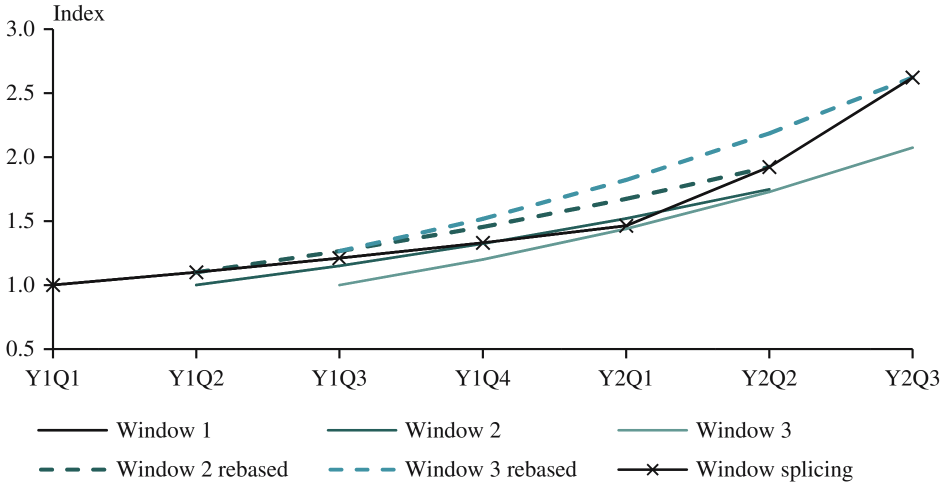


Fig. 5. Incorporating a window splice when the index cannot be revised

As for the revisable-index splicing, the window splicing for period T + t (where t > 0) can be expressed as:

$$P_{WS}^{0,T+t} = P_{[0,T]}^{0,1} \times P_{[1,T+1]}^{1,2} \times \ldots \times P_{[t,T+t]}^{t,T+t} \tag{11}$$

And so the index for the last period Y2Q2 in the example is similarly

$$P_{WS}^{0,6} = P_{[0,4]}^{0,1} \times P_{[1,5]}^{1,2} \times P_{[2,6]}^{2,6}$$

That is, the direct forms of the revisable-spliced and window-spliced indexes are the same.

There is a trade-off with the window-splice approach – between the quality of the most recent period's estimated movement and that of the longer-term index. However, in the case of fixed effects where at least two price observations are required before a new product contributes nontrivially to the estimation, the window splice ensures there is no systematic bias due to the continual omission of the implicit price movements of new products entering the index. It does this by incorporating a revision for them in the period after their introduction. In addition, this window splicing also incorporates implicit revisions for the improvement of the fixed-effects estimates with more price observations.

So the published, and unrevisable, quarter's movement incorporates a revision factor that adjusts the index to the level it would be if revision were possible. Obviously, with such an artificial example as this, where the estimated index changes significantly with each successive window's estimation, the effect on the most recent period's movement is quite substantial. But in practice the change to previous period's movements with the addition of one period of data is usually very small, so there would not be such a sacrifice in the quality of the most recent period's movement to maintain an unbiased longer-term index.

The FEWS index combines this window splice with the fixed-effects index to produce price indexes that are both quality-adjusted and nonrevisable.

To reiterate – the implicit price movement from $t - 1$ to t of a new product at time t will not be reflected at time t until the window splice is incorporated for the index estimated at time $t + 1$. That is, the implicit price movements will be reflected in the appropriate period, but with one period's lag.

De Haan (2015b) compares movement splicing with window splicing. From his Equation (12) we can state the FEWS index for period $T + 1$ as:

$$P_{FEWS}^{0,T+1} = \frac{P_{FE[1,T+1]}^{1,T}}{P_{FE[0,T]}^{1,T}} \times P_{FEMS}^{0,T+1} \tag{12}$$

where $P_{FEMS}^{0,T+1}$ is the fixed-effects movement-splice index for period $T + 1$.

The ratio of $P_{FE[1,T+1]}^{1,T}$ and $P_{FE[0,T]}^{1,T}$ is the implicit revision factor and de Haan's Equation (13) shows it can be expressed as

$$\frac{P_{FE[1,T+1]}^{1,T}}{P_{FE[0,T]}^{1,T}} = \exp\left[\sum_{i \in S^1} s_i^1 \left(\hat{\gamma}_{i[1,T+1]} - \hat{\gamma}_{i[0,T]}\right) - \sum_{i \in S^T} s_i^T \left(\hat{\gamma}_{i[1,T+1]} - \hat{\gamma}_{i[0,T]}\right)\right] \tag{13}$$

where:

$\hat{\gamma}_{i[t,T+t]}$ is the predicted fixed effect for product i estimated on the window from t to $T + t$,

s_i^t is the expenditure share of product i in period t, and S^t is the set of products observed in period t.

6. Empirical Results

Empirical results for the FEWS index are shown, on a range of data sources with different features:

- Scanner data from market-research company GfK. Three years of monthly average prices with a full set of characteristics, and their associated quantities, for eight consumer electronics products in New Zealand.
- Scanner data from market-research company IRI. Six years of weekly average prices for 30 supermarket products in the United States. Information on quantities sold is available in the data, but there is little information on the characteristics of products.
- Online data from the Billion Prices Project at MIT. Fifteen months of daily web-scraped online data for four consumer electronics products in New Zealand. This data has little information on characteristics of products, and no quantity information.

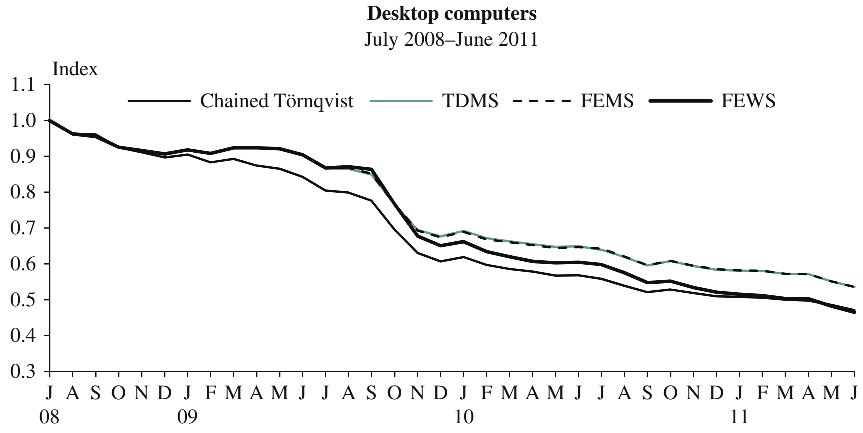
Descriptive statistics on the three data sources are shown in Appendix 3, showing the average number of distinct products per month (for GfK and IRI scanner data) and per week (for the BPP online data), along with the match rates after the first year for each product category, to indicate the levels of product turnover. Table 1 shows the variables available in each of the three data sources, and the price indexes that could be compared for each data source, given these varying data limitations.

Table 1. Price indexes compared for different data sources

		Price indexes compared	
Data source	Available variables	Traditional	Hedonic-based
Scanner data (GfK) *consumer electronics products*	Prices, quantities, characteristics	chained Törnqvist	TDMS, FEMS, FEWS
Scanner data (IRI) *supermarket products*	Prices, quantities	chained Törnqvist	FEMS, FEWS
Online data (PriceStats) *consumer electronics products*	Prices	chained Jevons	unweighted FEMS, unweighted FEWS

6.1. Scanner Data with both Quantities and Characteristics: Eight New Zealand Consumer Electronics Products

The FEWS index is estimated on three years of monthly average prices from New Zealand scanner data for eight consumer electronics products, from market-research company GfK. This data has extensive information on characteristics (around 40 different characteristics for each product) with corresponding total prices and quantities. From



Source: GfK

Fig. 6. Comparison of FEWS to other indexes for desktop computers

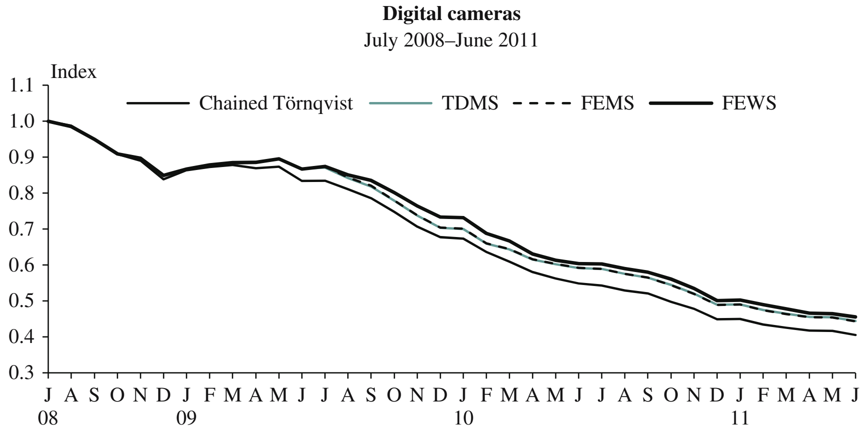
these, average monthly prices and expenditures for each combination of characteristics is derived. There is no outlet or retailer information available in the data, so this is not controlled for.

Because there is a full set of characteristics and expenditure information in the data, the FEWS index can be compared with the time-dummy movement-splice (TDMS) hedonic index, which explicitly incorporates the characteristics into a multilateral hedonic model. The TDMS is more commonly known as a ‘rolling-year time-dummy’ (RYTD) index, but we are renaming it in this article to make the splicing method explicit, for the comparison with the other indexes.

The fixed-effects movement-splice (FEMS) index is shown, in order to see the impact of the window splicing and the (implicit) addition of interaction terms separately.

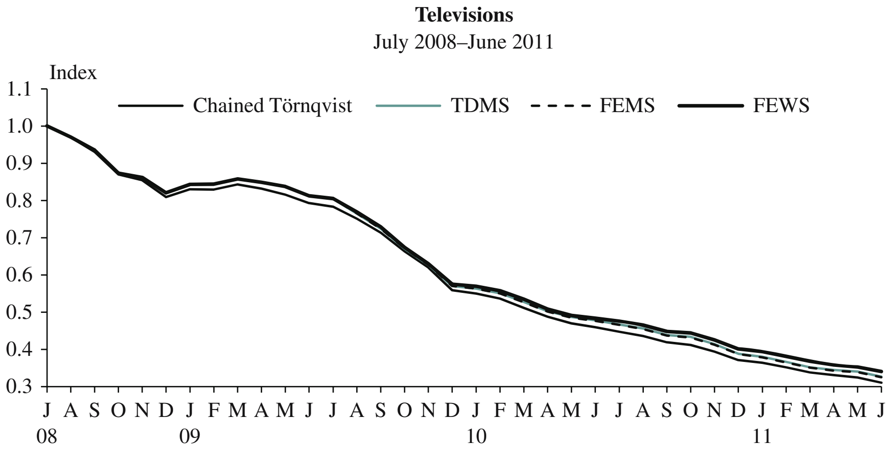
The more traditional chained-Törnqvist index is also included in the comparison:

$$P_T^{t-1,t} = \prod_{i\in M}\left(\frac{p_{it}}{p_{it-1}}\right)^{\sigma_i} \qquad (14)$$



Source: GfK

Fig. 7. Comparison of FEWS to other indexes for digital cameras



Source: GfK

Fig. 8. Comparison of FEWS to other indexes for televisions

where M is the set of products existing in both periods and σ_i is the arithmetic average of the share of expenditure on product i in periods t-1 and t:

$$\sigma_i = \frac{S_i^t + S_i^{t-1}}{2} \tag{15}$$

Note that while earlier versions of this article included the RYGEKS index in the comparison, recent work by Lamboray and Krsinich (2015) has demonstrated that the GEKS index is biased by product turnover, and has suggested a modification called the 'intersection GEKS' (intGEKS). The issue of whether the RYGEKS, as the nonrevisable counterpart of the standard GEKS, is appropriate in situations of high product turnover, such as consumer electronics, is therefore currently an open question, and so it is excluded from these results.

Figures 6 to 8 show the four indexes for desktop computers, digital cameras and televisions, respectively. Table 2 shows the value of the indexes at the end of the three-year study period for all eight consumer electronics product categories analysed.

For all three product categories graphed, the TDMS and FEMS indexes are virtually identical, suggesting that the implicit inclusion of interaction terms in the FEWS index adds little extra price-determining information for these product categories.

Table 2. Value of indexes at June 2011 (base = 1 at July 2008)

	Chained Törnqvist	TDMS	FEMS	FEWS
Camcorders	0.392	0.383	0.382	0.425
Desktop computers	0.471	0.536	0.536	0.465
Digital cameras	0.406	0.445	0.444	0.456
DVD players and recorders	0.588	0.595	0.633	0.631
Laptop computers	0.434	0.460	0.452	0.437
Microwaves	0.825	0.928	0.920	0.919
Portable media players	0.721	0.741	0.723	0.705
Televisions	0.311	0.327	0.325	0.340

Interestingly, the incorporation of the window splice for desktop computers results in a more rapidly declining index from late 2009, which could correspond to the implicit price movements associated with the introduction of new desktop computers being lower than those of existing desktop computers starting around then.

For both digital cameras and televisions, the FEWS index declines slightly less rapidly than the FEMS and TDMS, while the chained Törnqvist declines more rapidly.

Table 2 shows the very close correspondence of the TDMS and the FEMS indexes for most of the products.

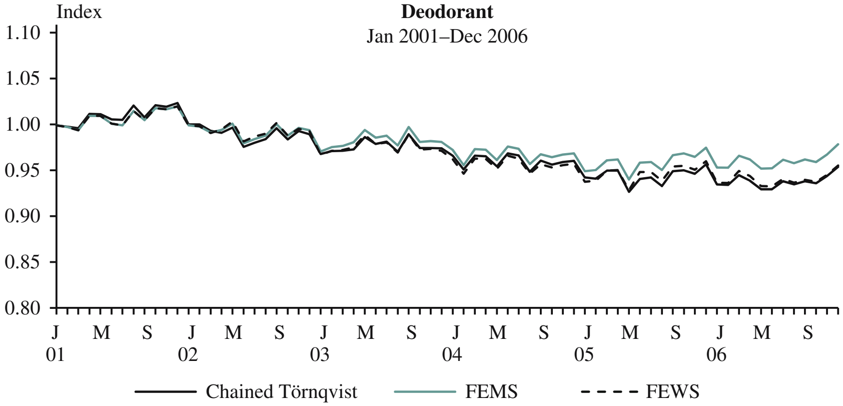
6.2. Scanner Data with Quantities but no Characteristics: 30 United States Supermarket Products

Six years of weekly aggregated supermarket scanner data from IRI marketing (see Bronnenberg et al. 2008) is used to compare the FEWS index with the FEMS and the chained-Törnqvist indexes. The data is aggregated to a monthly level for the sake of manageability.

Unlike the consumer electronics scanner data from GfK, supermarket scanner data tends to contain few, if any, characteristics with which to run either the ITRYGEKS or the time-dummy hedonic (TD) methods, which both rely on hedonic models that explicitly incorporate a full set of price-determining characteristics. However, outlet identifiers are available, and these are incorporated into the product identifiers to reflect and control for the different levels of service that can be experienced across outlets, and which can therefore be considered as part of the overall ‘product’.

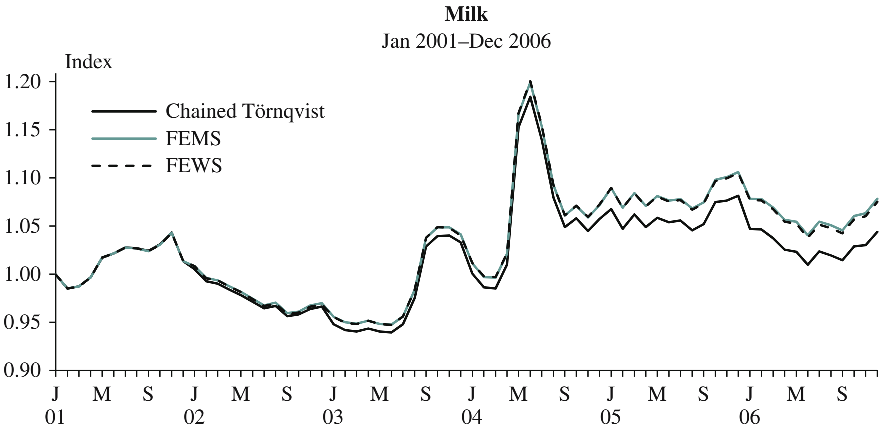
We compare the FEWS index with the FEMS index and the (monthly) chained-Törnqvist index. For most of the product categories, the three indexes track each other relatively closely. In general, there does not appear to be significant chain drift in the chained-Törnqvist index, in contrast to the findings in earlier research such as de Haan and van der Grient (2011) and Ivancic et al. (2011).

For brevity, then, graphs for just four of the 30 supermarket product categories analysed are shown in figures 9 to 12 – deodorant, milk, razors, and facial tissues. These



Source: IRI marketing

Fig. 9. Comparison of FEWS to other indexes for deodorant



Source: IRI marketing

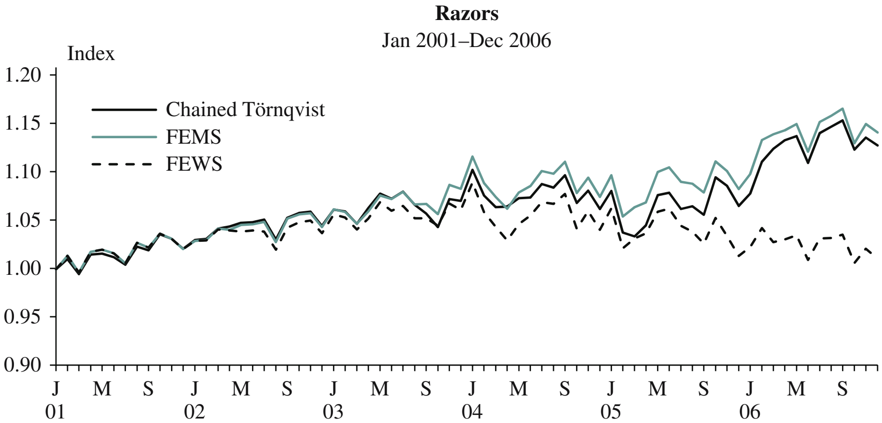
Fig. 10. Comparison of FEWS to other indexes for milk

are product categories for which the indexes do diverge, unlike most of the other product categories.

Table 3 gives the index values for the chained-Törnqvist, FEMS and FEWS indexes at the end of the six-year period analysed for all 30 products. Figures 13 and 14 present the same results graphically.

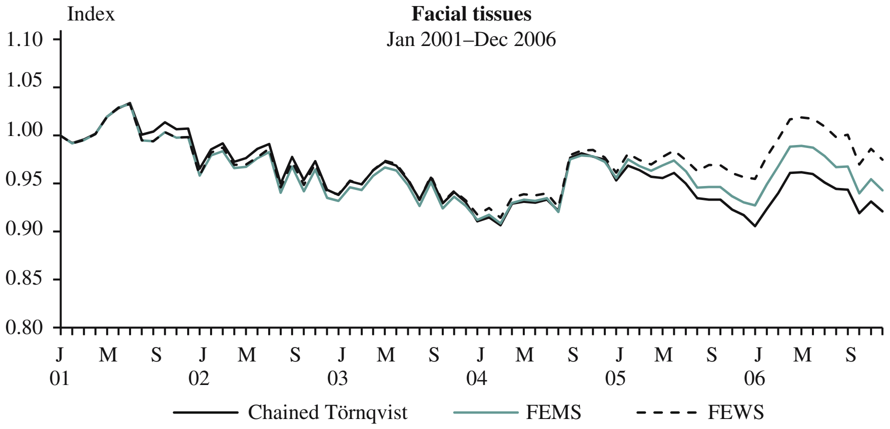
For many of the product categories, the three indexes track each other relatively closely. In general, there does not appear to be significant chain drift in the chained-Törnqvist index, in contrast to the findings in earlier research such as de Haan and van der Grient (2011) and Ivancic et al. (2011).

In Figure 9, for deodorant, the chained-Törnqvist and the FEWS index match closely, while the FEMS index tracks higher. An explanation for this could be that the valuing of characteristics by consumers is changing relatively quickly for this product category, and



Source: IRI marketing

Fig. 11. Comparison of FEWS to other indexes for razors

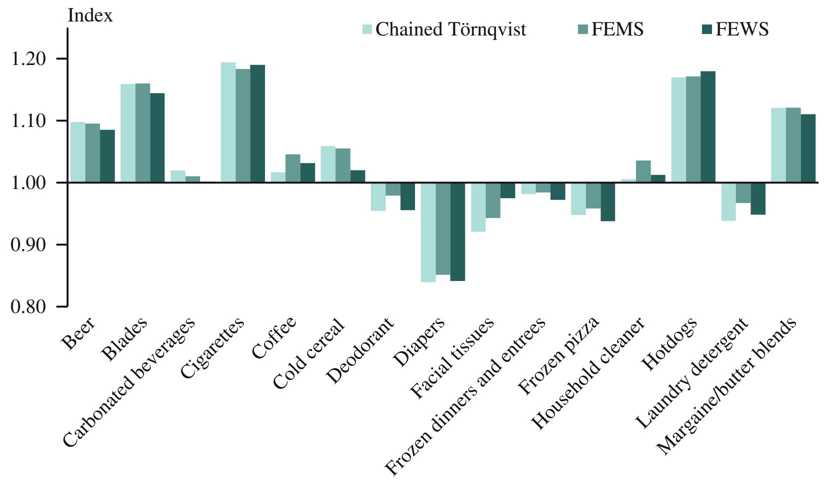


Source: IRI marketing

Fig. 12. Comparison of FEWS to other indexes for facial tissues

Table 3. Value of indexes at December 2006 (base = 1 at January 2001)

	Chained Törnqvist	FEMS	FEWS
Beer	1.098	1.095	1.085
Blades	1.159	1.160	1.144
Carbonated beverages	1.020	1.010	1.002
Cigarettes	1.194	1.184	1.190
Coffee	1.017	1.046	1.032
Cold cereal	1.059	1.056	1.020
Deodorant	0.955	0.979	0.956
Diapers	0.840	0.852	0.841
Facial tissues	0.921	0.943	0.975
Frozen dinners and entrees	0.981	0.984	0.972
Frozen pizza	0.948	0.958	0.938
Household cleaner	1.006	1.036	1.013
Hotdogs	1.170	1.171	1.180
Laundry detergent	0.938	0.967	0.949
Margarine / butter blends	1.121	1.121	1.111
Mayonnaise	1.114	1.109	1.089
Milk	1.044	1.079	1.076
Mustard & ketchup	1.114	1.137	1.126
Paper towels	1.039	1.071	1.087
Peanut butter	1.023	1.047	1.035
Photo supplies	0.819	0.841	0.839
Razors	1.129	1.142	1.011
Shampoo	0.948	0.987	0.942
Soup	1.149	1.172	1.111
Spaghetti / Italian sauce	1.055	1.061	1.048
Sugar substitutes	1.088	1.086	1.090
Toilet tissue	1.106	1.093	1.101
Toothbrushes	0.911	0.931	0.904
Toothpaste	0.964	0.968	0.937
Yoghurt	1.016	1.029	1.026



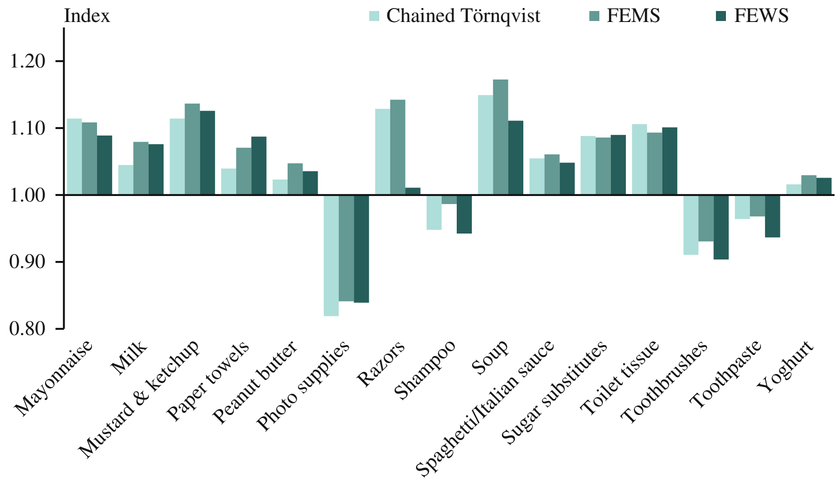
Source: IRI marketing

Fig. 13. Value of indexes at December 2006 (base = 1 at January 2001)

so the implicit revision for the updating of the fixed-effects parameters of the FEWS is making a difference. The chained Törnqvist is based only on current and previous period data, so is 'up-to-date' in this sense. Any difference between the FEWS and the chained Törnqvist would be due to either

- chain drift in the chained-Törnqvist index, or
- the implicit price movements of new products entering and old products disappearing from the market differing from the price movements of matched products.

So, either these two factors are cancelling each other out for this product, or there is no chain drift in the Törnqvist and the implicit price movements from unmatched products are similar to those of matched products.



Source: IRI marketing

Fig. 14. Value of indexes at December 2006 (base = 1 at January 2001)

Figure 10 for milk shows the FEMS and FEWS indexes tracking virtually identically – the indexes at December 2006 are 1.079 and 1.076 respectively – while there appears to be downwards chain drift influencing the chained-Törnqvist index.

Razors, shown in Figure 11, have the most unusual divergence in the three indexes of the 30 product categories analysed. From 2004 onwards the FEWS index trends downwards while the FEMS and chained-Törnqvist indexes continue to rise. This suggests that, during this period, there are downwards implicit price movements associated with the introduction of new products.

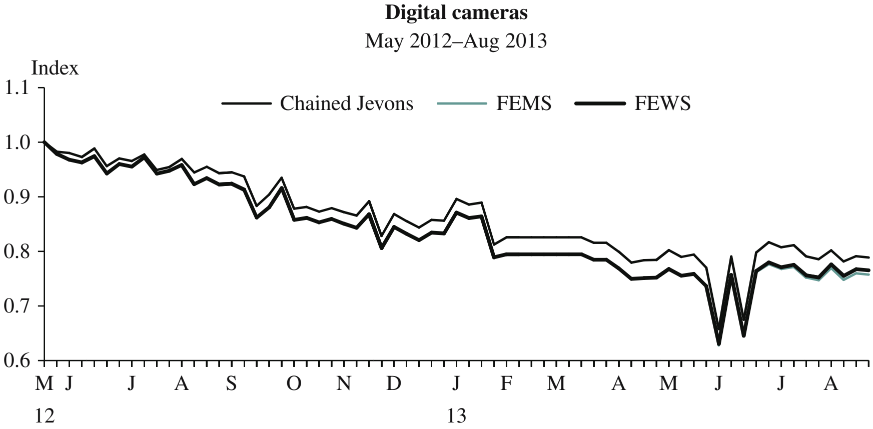
The three indexes for facial tissues, in Figure 12, all gradually drift apart. This is likely to suggest a combination of factors:

- downwards chain drift in the chained Törnqvist
- implicit price movements of new products differing from those of matched products and disappearing products (which the FEMS does reflect) and/or
- updating of fixed-effects parameters reflected by the FEWS but not the FEMS index.

Probably the main conclusion we can draw from the results for all 30 product categories, shown in Table 3 and Figures 13 and 14, is that there is no consistent pattern in the comparison of different index methods across all the products. This echoes the findings of previous research by Krsinich (2011a, 2013, 2014) and de Haan and Krsinich (2014). Further research is required to tease out the underlying factors driving different patterns of results, but this is not the focus of the present article.

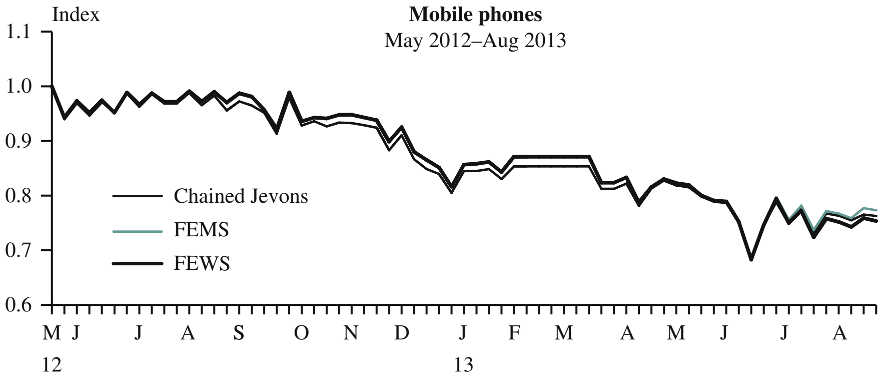
6.3. *Online Data with no Quantities or Characteristics: Four New Zealand Consumer Electronics Products*

Statistics NZ was given access to 15 months of daily web-scraped online data from the Billion Prices Project at MIT for New Zealand consumer electronics products from one major New Zealand retailer.



Source: Billion Prices Project @ MIT

Fig. 15. Comparison of FEWS to other indexes for digital cameras



Mobile phones
May 2012–Aug 2013

Source: Billion Prices Project @ MIT

Fig. 16. Comparison of FEWS to other indexes for mobile phones

Cavallo (2012) discussed possible approaches to quality adjusting this kind of online data, prompting this collaborative research into the fixed-effects index as a solution.

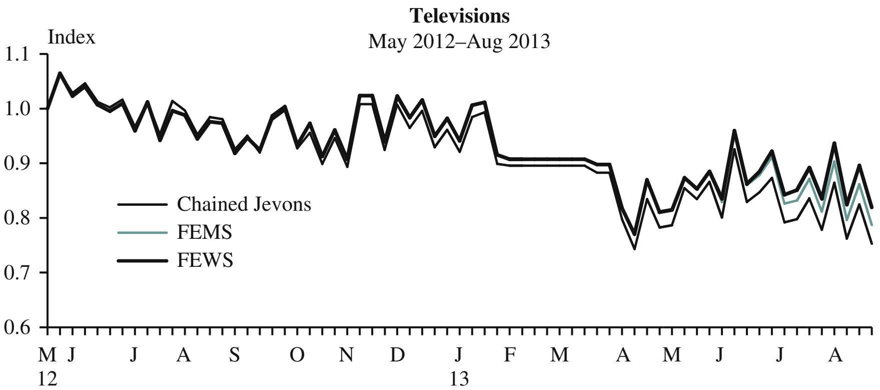
There are very few characteristics available in the online data but, as with scanner data, the products are identified by model name, and therefore any change in characteristics will generally correspond to a different product identifier. This means that the fixed-effects index can be used to implicitly quality adjust for all price-determining characteristics.

Another, less surmountable, limitation of online data is that quantities are not available. The unweighted index method used to estimate elementary-level price indexes by the Billion Prices Project and their commercial counterpart PriceStats is a daily-chained Jevons:

$$P_J^{t-1,t} = \prod_{i \in M} \left(\frac{p_{it}}{p_{it-1}} \right)^{\frac{1}{m}} \tag{16}$$

where m is the number of matched products between periods $t-1$ and t.

An advantage of the chained-Jevons index is that it is transitive and therefore does not suffer from chain drift. However, unlike the FEWS index, the Jevons does not reflect the



Televisions
May 2012–Aug 2013

Source: Billion Prices Project @ MIT

Fig. 17. Comparison of FEWS to other indexes for televisions

Table 4. Value of indexes at the end of August 2013 (base = 1 at mid-May 2012)

	Chained Jevons	FEMS	FEWS
Mobile phones	0.763	0.773	0.753
Digital cameras	0.788	0.757	0.765
Televisions	0.753	0.787	0.819

implicit price movements of new and disappearing products at the time of their introduction and disappearance.

Krsinich (2015) shows that the noninclusion of quantities may not have as much of an impact as might be expected. The effect of having no quantities in online data is simulated by estimating FEWS indexes with and without expenditure share weights on GfK scanner data for consumer electronics.

For the analyses shown here, day-per-week (Saturday) samples of the daily data were used.

Figures 15 to 17 compare the (unweighted) FEWS index to the (unweighted) FEMS and the weekly chained-Jevons indexes. Note that the analysis period is only 15 months long, and the FEMS and the FEWS indexes will be equivalent for most of this period – that is, the first 53 of the full 72 weeks – as this is the length of the initial estimation window.

Note that there was a lapse in data collection between January and March 2013, which results in a flat index for each of the product categories over this period.

Table 4 gives the value of the indexes at the end of the 15 months.

Over the 15-month period, the three indexes are relatively similar. The difference between the chained-Jevons and the FEWS index for digital cameras and televisions reflects the impact of the implicit price movements of new and disappearing products – which are reflected in the FEWS index, but not in the chained-Jevons index.

The high volatility of the indexes probably reflects the lack of smoothing by aggregation to monthly average prices, and the fact that the lack of quantities means that unusual price movements of low-selling products can have more of an impact on the index than they would in a weighted index. The coverage of the online data is also lower than that in the GfK scanner data, as it relates to just one, albeit large, New Zealand retailer.

7. Conclusion

This article has shown that the fixed-effects window-splice (FEWS) index produces nonrevisable quality-adjusted price indexes in the case of ‘big data’, such as scanner data or online data, where there is longitudinal price information at a detailed product-specification level.

There are two key factors behind this:

1. The fixed-effects index leverages off the longitudinal information in the data to implicitly quality adjust in a way that is shown to be equivalent to a fully interacted time-dummy hedonic index.
2. The use of a window splice implicitly revises so that the effects of new products are not systematically omitted as they would be with the more standard ‘movement splice’. This is a form of implicit revision that also enables the estimation of fixed effects for all products to be updated.

The FEWS index may be less appropriate for product areas where product characteristics or consumer preferences change rapidly, for example with seasonal variation or technological development. More research is required to understand and formulate the limits of the method's applicability in this context.

Appendix 1

Derivation of Equation 6

Restating and extending the working of de Haan and Hendriks (2013) so that all characteristics are treated as categorical and all interactions are included in the time-dummy hedonic models along with the main effects, the predicted prices of item (i.e., product) i in the base period 0 and the comparison periods t are as follows.

$$\hat{p}_i^0 = \exp(\hat{\delta}^0)\exp\left[\sum_{l=1}^{L}\hat{\beta}_l D_{il} + \sum_{m=1}^{M}\hat{\beta}_m D_{im}\right] \tag{17}$$

$$\hat{p}_i^t = \exp(\hat{\delta}^0)\exp(\hat{\delta}^t)\exp\left[\sum_{l=1}^{L}\hat{\beta}_l D_{il} + \sum_{m=1}^{M}\hat{\beta}_m D_{im}\right]; \quad (\mathrm{t}=1,\ldots,\mathrm{T}) \tag{18}$$

Taking the geometric mean of the predicted prices for all items belonging to the samples S^0 and $S^1,\ldots,S^T$ gives the following.

$$\prod_{i\in S^0}(\hat{p}_i^0)^{\frac{1}{N^0}} = \exp(\hat{\delta}^0)\exp\left[\left(\sum_{l=1}^{L}\hat{\beta}_l\sum_{i\in S^0}D_{il} + \sum_{m=1}^{M}\hat{\beta}_m\sum_{i\in S^0}D_{im}\right)/N^0\right] \tag{19}$$

$$\prod_{i\in S^t}(\hat{p}_i^t)^{\frac{1}{N^t}} = \exp(\hat{\delta}^0)\exp(\hat{\delta}^t)\exp\left[\left(\sum_{l=1}^{L}\hat{\beta}_l\sum_{i\in S^t}D_{il} + \sum_{m=1}^{M}\hat{\beta}_m\sum_{i\in S^t}D_{im}\right)/N^t\right] \tag{20}$$

Dividing (20) by (19) and rearranging gives the following.

$$\begin{aligned}\exp(\hat{\delta}^t) &= \frac{\prod_{i\in S^t}(\hat{p}_i^t)^{\frac{1}{N^t}}\exp\left[\left(\sum_{l=1}^{L}\hat{\beta}_l\sum_{i\in S^0}D_{il} + \sum_{m=1}^{M}\hat{\beta}_m\sum_{i\in S^0}D_{im}\right)/N^0\right]}{\prod_{i\in S^0}(\hat{p}_i^0)^{\frac{1}{N^0}}\exp\left[\left(\sum_{l=1}^{L}\hat{\beta}_l\sum_{i\in S^t}D_{il} + \sum_{m=1}^{M}\hat{\beta}_m\sum_{i\in S^t}D_{im}\right)/N^t\right]} \\ &= \frac{\prod_{i\in S^t}(\hat{p}_i^t)^{\frac{1}{N^t}}}{\prod_{i\in S^0}(\hat{p}_i^0)^{\frac{1}{N^0}}}\exp\left[\sum_{l=1}^{L}\hat{\beta}_l\left(\bar{D}_l^0 - \bar{D}_l^t\right) + \sum_{m=1}^{M}\hat{\beta}_m\left(\bar{D}_m^0 - \bar{D}_m^t\right)\right]\end{aligned} \tag{21}$$

where $\bar{D}_l^0 = \frac{\sum_{i\in S_0}D_{il}}{N_0}$ and $\bar{D}_l^t = \frac{\sum_{i\in S_t}D_{il}}{N_t}$ are the unweighted sample means of the dummy variable for category l of the main effects and $\bar{D}_m^0 = \frac{\sum_{i\in S_0}D_{im}}{N_0}$ and $\bar{D}_m^t = \frac{\sum_{i\in S_t}D_{im}}{N_t}$ are the unweighted sample means of the dummy variable for category *m* of the second-order interactions.

Because the residual terms sum to zero in each period, (21) can be rewritten.

$$P^{0t}_{TD(full)} = \exp(\hat{\delta}^t) = \frac{\prod_{i \in S^t} \left(p_i^t\right)^{\frac{1}{N^t}}}{\prod_{i \in S^0} \left(p_i^0\right)^{\frac{1}{N^0}}} \exp\left[\sum_{l=1}^{L} \hat{\beta}_l \left(\bar{D}_l^0 - \bar{D}_l^t\right) + \sum_{m=1}^{M} \hat{\beta}_m \left(\bar{D}_m^0 - \bar{D}_m^t\right)\right] \qquad (22)$$

And this is Equation (6) in the main text.

Appendix 2

Intuition

It is difficult to understand how it is possible for quality adjustment to happen without any explicit information on characteristics. The key point is that, in scanner and online data, products are defined at the detailed specification level – that is, at a level at which any change in price-determining characteristics would prompt a new identifier. Rather than estimating each characteristic's effect separately, the fixed-effects model is estimating the effect of each 'bundle' of characteristics corresponding to a particular product.

To get an intuition for how this works, first consider four products, all with the same set of characteristics, which are fixed across time. The standard approach to hedonic modelling fits the log of price against characteristics and time, so the graphs of logged price for each product can be visualised as shown below in Figure 18. *P'0* is the predicted logged price for the base product *P0*, and actual logged prices for each of the four products *P0* to *P3* differ from this predicted *P0* each period by a normally distributed error term.

Now consider the situation where the products each have different characteristics from one another – that is, they are now four distinct products – and these characteristics are fixed across time at the product level – as they would be in the case of scanner data or online data, where barcodes or product names change if there is a change in characteristics.

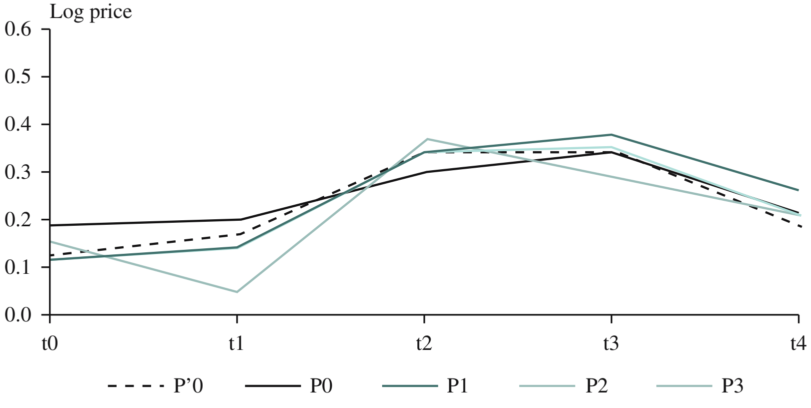


Fig. 18. All products with the same characteristics

Assume that, within the estimation window of the hedonic model, characteristics' parameters are constant. Another, perhaps more useful, way of stating this is that the model will estimate the average value of each characteristic's parameter within the estimation window.

The net effect of each product's 'bundle of characteristics' is constant over time, so the logged price graph is moved up or down accordingly. This constant factor is the 'fixed effect' estimated by the fixed-effects model – that is, the product-specific intercept. This is shown below in Figure 19.

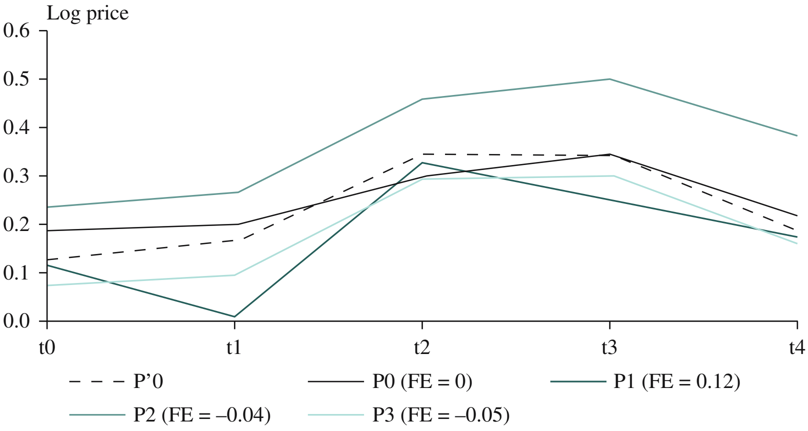


Fig. 19. Each product with a different set of characteristics

So, the fixed effects are estimated such that the sum of squares of the error terms is minimised for each of the products, resulting in the predicted log price for the base product *P'0* (i.e., the dashed line).

Consider, now, the situation of a new product *P4* entering at time *t1*. Figure 20 shows the logged price of this new product *P4* over the full period from *t0* to *t4*, alongside the logged prices of the other products.

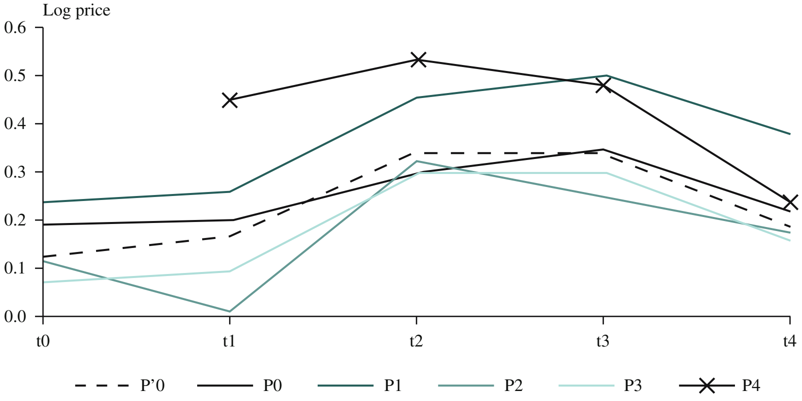


Fig. 20. New product entering at time t1

In the period of introduction, the best estimate of the fixed effect of the new product *P4* – that which minimises the sum of squared error terms – is simply the distance between its log price and the predicted log price of the base item *P0*. This is shown below in Figure 21. Clearly, this is a trivial estimate, and the model (and therefore the index derived from it) is unaffected by the inclusion of the new product in this first period. Consequently, the corresponding implicit price movement associated with introducing this new product is simply the underlying price movement of the matched products.

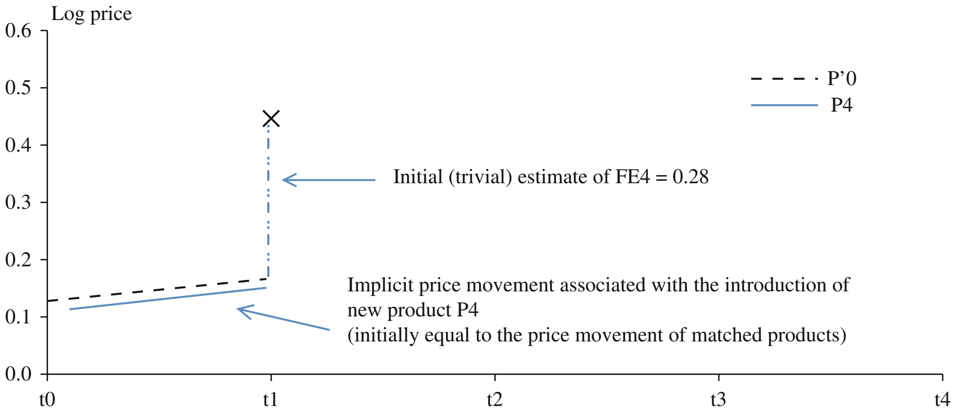


Fig. 21. Period of introduction of new product P5

This is the reason that, as noted by Diewert (2004), matched-model and fixed-effects indexes are equivalent when we consider only bilateral (two-period) model estimations.

The second price observation of *P4* enables non-trivial estimate of the fixed effect of new product *P4* (and consequently of the implicit price movement associated with its introduction), as shown in Figure 22. This can be visualised as shifting the logged price record of *P4* vertically up or down until the sum of squares of the error terms (indicated with bold lines) is minimised.

Note that for the purposes of this example it is assumed that the underlying predicted price movement of the base product *P'0* is unaffected by the new product *P4*. Of course,

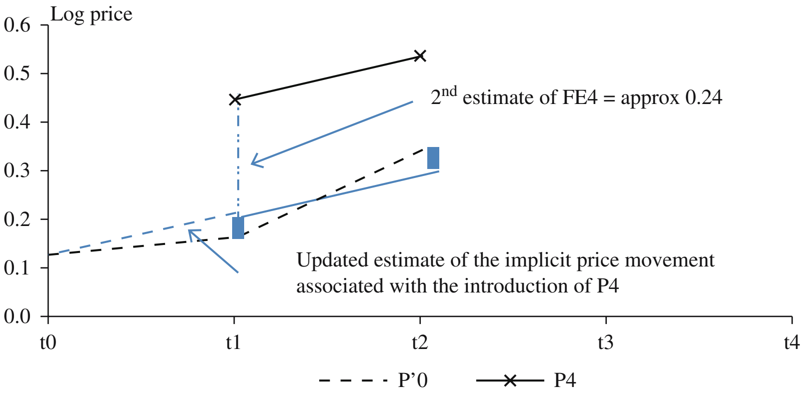


Fig. 22. Second price observation period for new product P5

the predicted price will be slightly affected by the new product and it is the appropriate estimation of the new product's fixed effect, and therefore its influence on the resulting price index, that is the objective.

As more prices are observed for the new product *P4*, the estimate of its fixed effect is updated accordingly. Figures 23 and 24 show how the estimate of the fixed effect gradually reduces each period as a result of fitting the longitudinal logged price record for this new product around the underlying logged price movement predicted from all products.

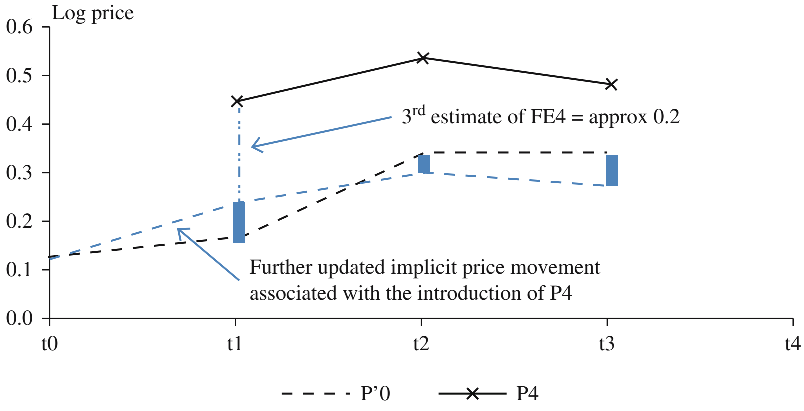


Fig. 23. Third price observation period for new product P5

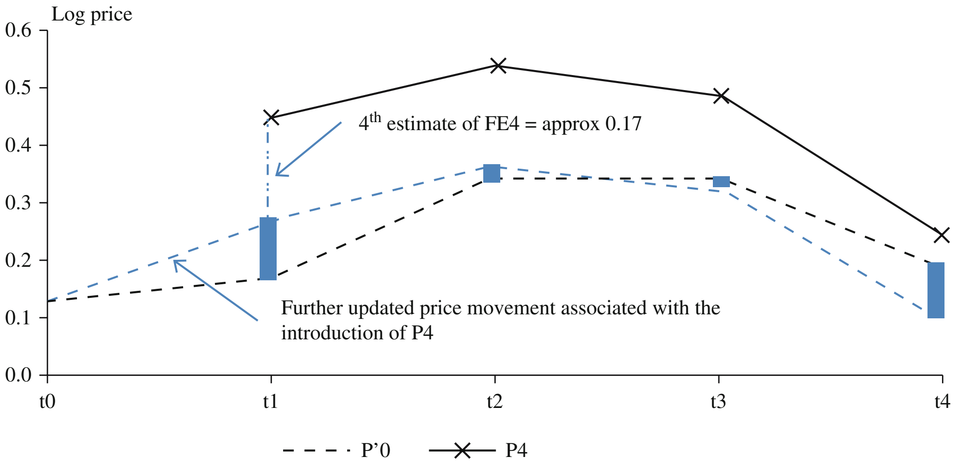


Fig. 24. Fourth price observation period for new product P5

Because the estimate of the fixed effect is trivial in the first period, at least two price observations are required before a new product can contribute meaningfully to the estimation. And with more price observations, the estimate of the fixed effect for the new product converges to its true value – that is, the estimate is being driven more by the shape of the longitudinal price record than by the error terms. The window splice described in Section 5 enables an implicit revision to be incorporated along with the most recent period's estimated movement to account for this updating of the fixed-effects estimates.

Appendix 3

Descriptive Statistics on the Three Data Sources

Tables 5 to 7 show the average number of distinct products per month (for the GfK and IRI scanner data) and per week (for the BPP data, as represented by the Saturday sampling) for each product category.

They also show the percentage match rates after the first year for each product category, to indicate the rates of product turnover. These show that the turnover for consumer electronics can be very rapid – at its most extreme, only three percent and four percent of products are matched after the first year for laptop computers and desktop computers respectively. Even

Table 5. Average number of products per month, and match rate after the first year, for GfK consumer electronics scanner data

	Number of products	Match rate (%)
Camcorders	88	41
Desktop computers	150	4
Digital cameras	289	23
DVD players and recorders	202	26
Laptop computers	432	3
Microwaves	152	33
Portable media players	161	18
Televisions	341	23

Table 6. Average number of products per month, and match rate after the first year for IRI supermarket scanner data

	Number of products	Match rate (%)
Beer	2951	84
Blades	445	87
Carbonated beverages	4138	90
Cigarettes	1650	84
Coffee	2883	87
Cold cereal	1679	88
Deodorant	1039	83
Diapers	401	70
Facial tissues	234	80
Frozen dinners and entrees	1438	89
Frozen pizza	1247	93
Household cleaner	519	82
Hotdogs	929	93
Laundry detergent	785	81
Margarine / spreads / butter blends	258	94
Mayonnaise	358	90
Milk	2794	89
Mustard and ketchup	872	85
Paper towels	419	88
Peanut butter	251	92
Photography supplies	163	82
Razors	83	80
Shampoo	2216	76
Soup	1120	89
Spaghetti / Italian sauce	1418	91
Sugar substitutes	145	90
Toilet tissue	238	78
Toothbrushes	569	78
Toothpaste	731	75
Yoghurt	1751	95

Table 7. Average number of products per week, and match rate after the first year for BPP consumer electronics online data

	Number of products	Match rate (%)
Digital cameras	94	49
Mobile phones	60	16
Televisions	51	18

the consumer electronics product category with the highest match rate in the GfK scanner data – camcorders – has just 41% of products still being observed after the first year.

In contrast, the match rates for the supermarket products are much higher. These range from 70% for diapers to 95% for yoghurt.

8. References

Aizcorbe, A., C. Corrado, and M. Doms. 2003. "When Do Matched-Model and Hedonic Techniques Yield Similar Price Measures?" Working Paper no. 2003-14, Federal Reserve Bank of San Francisco. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=550421 (accessed March 2016).

Bronnenberg, B.J., M. Kruger, and C.F. Mela. 2008. "The IRI Marketing Data Set" *Marketing Science* 27: 745–748. Available at: https://www.researchgate.net/publication/259286152_The_IRI_marketing_data_set (accessed April, 2016).

Cavallo, A. 2012. "Overlapping Quality Adjustment Using Online Data." Presentation to the 2012 Economic Measurement Group, Sydney, November 21–23, Australia. Available at: https://www.business.unsw.edu.au/research-site/centreforappliedeconomicresearch-site/Documents/A.%20Cavallo%20-%20Overlapping%20Quality%20Adjustment%20Using%20Online%20Data.pdf (accessed April 2016).

de Haan, J. 2015a. "The Time-Product Dummy Method and Implicit Quality Adjustment." Unpublished draft, May 2015.

de Haan, J. 2015b. "Rolling Year Time Dummy Indexes and the Choice of Splicing Method." Paper presented at the 14th meeting of the Ottawa Group, Tokyo. 20–22 May, 2015. Available at: http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s3room.pdf (accessed April 2016).

de Haan, J. and H. van der Grient. 2011. "Eliminating Chain Drift in Price Indexes Based on Scanner Data." *Journal of Econometrics* 161: 36–46. Doi: http://dx.doi.org/10.1016/j.jeconom.2010.09.004.

de Haan, J. and R. Hendriks. 2013. "Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes." Paper presented at the 2013 Economic Measurement Group, Sydney, Australia. 28–29 November, 2013. Available at: https://www.business.unsw.edu.au/research-site/centreforappliedeconomicresearch-site/Documents/Jan-de-Haan-Online-Price-Indexes.pdf (accessed April 2016).

de Haan, J. and F. Krsinich. 2014. "Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes." *Journal of Business & Economic Statistics* 32: 341–358. Doi: http://dx.doi/10.1080/07350015.2014.880059.

Diewert, W.E. 2004. "On the Stochastic Approach to Linking the Regions in the ICP." Discussion Paper no. 04-16, Department of Economics, University of British Columbia, Vancouver, Canada. Available at: http://papers.economics.ubc.ca/legacypapers/dp0416.pdf (accessed March 2016).

Ivancic, L., W.E. Diewert, and K.J. Fox. 2011. "Scanner Data, Time Aggregation and the Construction of Price Indexes." *Journal of Econometrics* 161: 24–35. Doi: http://dx.doi.org/10.1016/j.jeconom.2010.09.003.

Krsinich, F. 2011a. "Price Indexes from Scanner Data: A Comparison of Different Methods." Paper presented at the 12th meeting of the Ottawa Group, Wellington, New Zealand. 4–6 May, 2011. Available at: http://www.stats.govt.nz/~/media/Statistics/ottawa-group-2011/Ottawa-2011-Papers/Krsinich-2011-paper-PImethods-comparison.pdf (accessed April 2016).

Krsinich, F. 2011b. "Measuring the Price Movements of Used Cars and Residential Rents in the New Zealand Consumers Price Index." Paper presented at the 12th meeting of the Ottawa Group, Wellington, New Zealand. 4–6 May, 2011. Available at: http://www.stats.govt.nz/~/media/Statistics/ottawa-group-2011/Ottawa-2011-Papers/Krsinich-2011-paper-Rentals-cars.pdf (accessed April 2016).

Krsinich, F. 2013. "Using the Rolling Year Time-Product Dummy Method for Quality Adjustment in the Case of Unobserved Characteristics." Paper presented at the 13th meeting of the Ottawa Group, Copenhagen, Denmark. 1–3 May, 2013. Available at: http://www.dst.dk/da/Sites/ottawa-group/~/media/Kontorer/12-Priser-og-forbrug/Ottawa-Group/Frances%20Krsinich%202%20Ottawa%20Group%202013%20RYTPD%20final.pdf (accessed April 2016).

Krsinich, F. 2014. "Fixed Effects with a Window Splice – Non-Revisable Quality-Adjusted Price Indexes with No Characteristic Information." Paper presented at the meeting of the group of experts on consumer price indices, 26–28 May 2014, Geneva, Switzerland. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/New_Zealand_-_FEWS.pdf (accessed April 2016).

Krsinich, F. 2015. "Price Indexes from Online Data Using the Fixed-Effects Window-Splice Method." Paper presented at the 14th meeting of the Ottawa Group, Tokyo, Japan. 20–22 May, 2015. Available at: http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s2p7_pap.pdf (accessed April 2016).

Lamboray, C. and F. Krsinich. 2015. "A Modification of the GEKS Index When Product Turnover is High." Paper presented at the 14th meeting of the Ottawa Group, Tokyo, Japan. 20–22 May, 2015. Available at: http://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s1p2_pap.pdf (accessed April 2016).

Melser, D. 2011. "Constructing Cost of Living Indexes Using Scanner Data." Unpublished draft, September 2011 – an updated version of "Constructing High Frequency Indexes Using Scanner Data." Paper presented at the 12th meeting of the Ottawa Group, Wellington, New Zealand. 4–6 May, 2011. Available at: http://www.stats.govt.nz/~/media/Statistics/ottawa-group-2011/Ottawa-2011-Papers/Melser-2011-paper-Constructing-indexes.pdf (accessed April 2016).

The Billion Prices Project. 2016. Available at: http://bpp.mit.edu (accessed April 2016).

Received August 2014
Revised October 2015
Accepted November 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 405–431, http://dx.doi.org/10.1515/JOS-2016-0022

"Do the Germans Really Work Six Weeks More than the French?" – Measuring Working Time with the Labour Force Survey in France and Germany

Thomas Körner[1] *and Loup Wolff*[2]

Measuring working time is not only an important objective of the EU Labour Force Survey (LFS), but also a highly demanding task in terms of methodology. Against the background of a recent debate on the comparability of working time estimates in France and Germany, this article presents a comparative assessment of the measurement of working time in the Labour Force Survey obtained in both countries. It focuses on the measurement of the hours actually worked, the key working-time concept for short-term economic analysis and the National Accounts. The contribution systematically analyses the differences in the measurement approaches used in France and Germany in order to identify the methodological effects that hinder comparability. It comes to the conclusion that the LFS overstates the difference in hours actually worked in France and Germany and identifies question comprehension, rounding, editing effects, as well as certain aspects of the sampling design, as crucial factors of a reliable measurement in particular of absences from work during the reference week. We recommend continuing the work started in the European Statistical System towards the development of a model questionnaire in order to improve cross-national harmonisation of key variables such as hours actually worked.

Key words: Nonsampling errors; measurement error; questionnaire design; working hours; international comparability.

1. Measuring Working Time as an Objective of the Labour Force Surveys

Measuring working time is one of the most important of the numerous objectives of Labour Force Surveys (LFS). In many countries, working-time estimates based on the LFSs are vital data for inputting into the National Accounts (e.g., to calculate indicators on the volume of work and labour productivity) as well as for a whole range of socioeconomic analyses. The LFS carried out in the member states of the European Union provide data regarding two main working-time concepts, both of which are defined according to the Resolution of the Eighteenth International Conference of Labour Statisticians (ICLS) endorsed in 2008 (International Labour Organization, ILO 2008) – the hours actually worked and the hours usually worked.

[1] Federal Statistical Office Germany, 65180 Wiesbaden, Germany. Email: thomas.koerner@destatis.de
[2] Ministére de la culture et de la communication, 182 rue Saint Honoré, 75033 Paris cedex 01. Email: loup.wolff@culture.gouv.fr

Acknowledgments: This contribution is a revised, updated and expanded version of two papers presented at the 7th European Workshop on Labour Force Survey Methodology, Madrid, 10–11 May 2012. The authors would like to thank the anonymous reviewers and the associate editor for their useful comments. The views expressed in this article are those of the authors and not necessarily those of their institutions.

In practice, measuring the hours actually worked and those usually worked can only be based on household surveys, since the employed persons are the only respondents able to quantify the number of hours they actually and usually work. It could be suggested that time-use surveys are better than LFSs for this purpose. However, in this respect, the major disadvantage of time-use surveys is that, in comparison to the LFSs, they provide limited labour market information, have a small sample size and long intervals between the survey waves, do not cover the entire year, and do not use the concept of the reference week.

This explains the widespread use and key role of LFSs in measuring working time, despite the fact that LFSs often do not allow extensive questionnaire space or interviewing time for questions on specific topics such as working time. The strengths of LFS data are that they are based on large samples, they are able to provide timely and intra-annual data, that they cover all economic activities, including formal and informal employment, and that they are (at least in theory) in a position to implement the concepts as defined in the ICLS Resolution.

Nevertheless, the difficulties in measuring working time are regularly seen in the results of LFSs. There are often discrepancies between the results of the LFS and the National Accounts. The results regarding hours actually and hours usually worked are often inconsistent and international comparability is limited. Sometimes such (apparent or real) inconsistencies cause controversy, as in the case of the results for France and Germany. In 2012, a study based on Eurostat tabulations noted that the hours actually worked were much lower in France compared to Germany (Coe-Rexecode 2012). At the time, a lively debate ensued: French newspapers such as *Libération*, *Les Echos* and *Le Figaro* were incredulous and investigated the reliability of the study. The criticism was aggravated even further due to the fact that the working-time estimates in the National Accounts show inverted differences. The findings continue to provoke intense debate even today (Coe-Rexecode 2014).

Despite this controversy, the problem of measuring working time in LFSs is addressed surprisingly rarely in the literature. Still, the issues around measurement do feature in the discussion: in a broad overview, Mata-Greenwood (2001, 9–11) mentions recall errors, proxy errors, question comprehension issues and social desirability bias. While these effects may lead to over- as well as underestimations of working time, most contributions assume that, generally speaking, the effects lead to a underestimation of absences in the responses to the hours actually worked (for an overview, see OECD 2004).

As mentioned previously, empirical studies on the accuracy of working-time estimates are rare. Regarding the measurement of the hours actually worked, a number of studies compare data from LFS using a question covering the entire reference week with diary-based measurement approaches, mostly implemented in the context of time-use surveys. In most cases, the LFS-based estimates are 2% to 10% higher than the respective diary information (see e.g., Niemi 1993; Robinson and Bostrom 1994; Robinson et al. 2011). Smaller differences were found by Williams (2004), in whose analysis the LFS estimates were only 1.6% higher than time-use survey data, and by Frazis and Stewart (2004), who found only small differences between LFS and time-use survey data. In addition to a number of methodological limitations relating to the comparison of time-use survey and LFS data, all these studies have the drawback that they either completely exclude those employed persons who were not at work during the reference week from their analyses or at least have difficulties in comparing data based on reference weeks. Nevertheless, people who are employed but not at work in the reference week need to be included in the comparative analysis, as it is often assumed that absences due to holidays or illness have a major impact on the accuracy of working-time measurements. An international aggregate-level comparison of estimates on hours actually worked (including absences) also shows higher values for LFS-based results, compared to administrative data (Fleck 2009, 27).

A similar, but less pronounced, tendency to produce higher working-time estimates has also been confirmed by a number of studies regarding the hours usually worked and the contractual hours of work. These studies are based on comparisons with establishment survey data (Williams 2002) and administrative registers (Villund 2009). For instance, in the case of the hours usually worked, the most common problem is that the hours usually worked are difficult to collect from establishment surveys, as employers ordinarily only have the information about the contractual hours of work and not the hours usually worked.

Unlike most previous studies, this analysis is not primarily based on comparisons with other data sources, be it time-use surveys or administrative data. After an overview of the relevant statistical concepts and their operationalisation in France and Germany (Section 2), our analytical point of departure is a comparison of the LFS results from France and Germany. We discuss whether these results are consistent with other data sources regarding important working-time components such as entitlement to paid leave, sick leave, and the number of public holidays (Section 3). Given that the LFS in Germany and France use quite different measurement approaches, we then attempt to identify possible sources of error that may contribute to the differences observed (Section 4).

2. Working-Time Concepts and their Implementation in the Labour Force Surveys of France and Germany

Conceptually, the measurement of working time in official statistics is based on the *Resolution Concerning the Measurement of Working Time* adopted by the Eighteenth International Conference of Labour Statisticians (ICLS) in 2008 (ILO 2008). This Resolution distinguishes no less than seven concepts of working time: the hours actually worked, the hours usually worked, the hours paid for, the normal hours of work, the contractual hours of work, the overtime hours of work, and the absence from work hours. As acknowledged by the Resolution, the variety of these seven concepts makes it impossible to capture all of them in one single data source: the Resolution suggests using LFSs to measure the hours actually worked and the hours usually worked, and to use establishment surveys as well as administrative registers to measure the hours paid for, the contractual hours of work, and the normal hours of work. This distinction is important, as it highlights the fact that there are few reference data sources to compare the LFS results for the hours actually worked and the hours usually worked.

The concept of hours actually worked is the key concept of the ICLS Resolution and it is closely related to the definition of the other concepts, in particular the hours usually worked. The hours actually worked refer to "the time spent in a job for the performance of activities that contribute to the production of goods and/or services during a specified short or long reference period" (ILO 2008, 43). The reference period for the LFSs, usually the reference week used for the measurement of the employment status, is also applied to working-time measurement.

Following the recommendation to use LFSs to measure the hours actually worked is not straightforward, as the ICLS definition is very complex, reflecting the complexity of the issue in national working-time regulations. Regarding the productive activities within the production boundary of the System of National Accounts (SNA), it contains at least nine distinct components to be included and seven components to be excluded when determining the hours actually worked (the definition of which is problematic in itself). Table 1 provides a simplified overview of the working-time components. In addition to the time spent actually carrying out the tasks that are part of the job, working time also includes the time spent preparing the work, waiting time, most travel time between work locations, time for work-related training, as well as short breaks other than the lunch break. Time spent on call is considered working time depending on the extent to which the worker is prevented from participating in other activities. Conversely, annual leave, sick leave, public holidays, other types of leave, as well as educational activities not required by the job and longer breaks are excluded from the hours actually worked.

For the hours usually worked, the same rules apply; they are defined as "the typical value of hours actually worked [. . .] over a long observation period of a month, quarter, season or year" (ILO 2008, 45).

In contrast to this complex definition, LFSs usually try to determine the number of hours actually worked and the hours usually worked with one single question each, tacitly assuming that the inclusion and exclusion criteria are similar to respondents' everyday-life perceptions (and can consequently be clarified with a short instruction). Nevertheless, the resulting questions, usually of the type "How many hours did you actually work in the week from [. . .] to [. . .]?" are highly demanding cognitively, even if one assumes that the distinctions made in the ICLS Resolution are close to those in the respondents' minds (see Chapter 4).

The LFSs is carried out in all European Union member states, as well as in EFTA countries, in the Former Yugoslav Republic of Macedonia, and in Turkey. It is based on a number of European legal acts, the main one being the Council Regulation No. 577/1998. The legal acts are based on the principle of output harmonisation, that is, they specify the output data to be provided to Eurostat, but give a certain leeway to take national circumstances into account for survey implementation. The regulations in particular specify the required precision (and indirectly the effective sample size), the equal distribution of the sample over the calendar weeks, the variables and items to be provided as well as the transmission deadlines. The actual survey implementation, including the sampling frame, question wording, data collection modes and so on, is largely not regulated by law (although some provisions exist regarding the measurement of labour status). The member states are provided with a set of recommendations for implementation, however (for an overview of the LFS harmonisation strategy, see Körner 2012). Nevertheless, substantial differences exist regarding the operationalisation of the variables in questionnaires (see, e.g., Massarelli 2011) and the data collection modes applied (Blanke and Luiten 2014). The differences recently led to the development and joint international testing of model questionnaire modules for key variables of the core LFS (such as labour status and working time) and ad-hoc modules. Table 2 provides an

Table 1. Components to be considered for the measurement of hours actually worked (adapted from ILO 2008)

Hours actually worked . . .	
. . . include time spent	**. . . exclude time spent**
- carrying out the tasks and duties of the job (in any location and during periods not dedicated to work) - cleaning or maintaining tools, instruments, processes, changing time, decontamination or washing-up time - purchasing or transporting basic materials - waiting for business, customers or patients (as part of working-time arrangements) - on temporary interruptions of a technical, material or economic nature - on on-call duty (depending on the extent to which other activities and movements are restricted) - on training and skills enhancement (required by the job) - travelling between work locations (e.g., business trips) - on short breaks	- on annual leave - on sick leave - on public holidays - on other leave for personal or family reasons or public duties (e.g., parental leave) - commuting between work and home (if no productive activities are performed) - on long breaks (meal breaks and resting time during long trips) - on educational activities not required by the job

overview of important design elements of the LFS in France and in Germany (regarding the situation in other member states, see the summary in Eurostat 2013).

Regarding the number of hours actually worked, the legal basis of the LFS currently does not include anything other than the name of the variable (HWACTUAL) and the specification that the "number of hours actually worked during the reference week in the main job" (Regulation 377/2008) is to be collected and that the persons employed, but not at work in the reference week, are to be coded "00". The recommendations for the implementation of the questionnaire basically summarise the ICLS Resolution (still currently the resolution adopted in 1962), without giving any recommendations on questionnaire implementation.

Consequently, there is considerable variation in the measurement instruments used for data collection, and this is also the case for working time. France and Germany are particularly interesting cases for comparison, as they can be considered extreme cases on the continuum between very detailed measurement involving many questions (France) and a very limited number of questions (Germany).

Like most other member states, the German questionnaire basically uses a single-question approach with one question asking for the hours actually worked in the reference week, while the French LFS questionnaire includes a detailed sequence of questions introducing the concept and the situation in the reference week.

The presentation of the findings from the present study to the working committees of the European Statistical System has already triggered further development work: since April 2013, a Eurostat task force has been trying to improve and further harmonise the measurement of working time in the LFS. The task force is developing a model

Table 2. Important design elements of the French and the German LFS (2012)

	France	Germany
Net quarterly sample size (persons 15–74 years)	106,000	129,500
Quarterly sampling rate	0.25%	0.25%
Sample design	Area sampling	One-stage area sample (average size of sampling districts 9 dwellings)
Sampling frame	Tax register (habitation tax)	Census 1987 for former territory/Central Population Register 1990 for new federal states/Annual updates based on the statistics of building permissions
Rotation scheme	Quarterly rotation (6 consecutive quarters)	Annual rotation only (four consecutive years)
Reference week	Fixed reference weeks (see Chapter 4)	Sliding reference weeks (see Chapter 4)
Response rate	84.7%	98.2%
Mandatory or voluntary response	Mandatory	Mandatory
Rate of proxy interviews	29%	25%
Data collection modes first interview	CAPI	CAPI, self-administered PAP, Telephone interview
Data collection modes follow-up interview	CATI, CAPI	CAPI, self-administered PAP, Telephone interview

questionnaire with the aim of improving the measurement of absences and achieving greater harmonisation. The first version of the model questionnaire has already been tested in several member states. The task force's approach is discussed in relation to our findings in Chapter 5.

3. Assessing the Measurement of Working Time in France and Germany

As mentioned in the introduction, the number of hours actually worked per year is a crucial indicator for the National Accounts. In some countries, the estimation for the National Accounts is derived directly from the LFS, while in others, like France and Germany, it is at least partly independent of the LFS (see Wanger 2013 for Germany and Lefèvre, Rakotomalala, and Toutlemonde 2012 for France). Table 3 shows that comparing working-time estimates is a delicate task, not only internationally, but also as regards comparisons between the LFS and the National Accounts. Starting with the working time of employed persons, it could be concluded that there is hardly any difference between France and Germany, according to the LFS. In contrast, in the National Account estimates, Germany registers a significantly lower number of hours actually worked per year (−6.9%). The difference in the German results is not only due to the actual measurement of working time, but also to the deviations in the share of part-time workers. The LFS may

Table 3. Number of annual hours actually worked per person in France and Germany (2012)

		Germany	France	Relative difference %
Employed persons	LFS (main job only)	1,636	1,632	0.2
	National Accounts	1,393	1,489	−6.9
Employees	LFS (main job only)	1,586	1,554	2.0
	National Accounts	1,297	1,402	−8.1
Full-time employees	LFS (main job only)	1,863	1,684	9.6
	National Accounts	1,646	not available	
Part-time employees	LFS (main job only)	838	985	−17.5
	National Accounts	634	not available	

have difficulties comprehensively capturing marginal employees (Körner and Puch 2011). For this reason, the German National Accounts mainly use other sources, for instance the Employment Statistics Register, to estimate the number of full-time and part-time employees (in 2012, the share of part-time employees in Germany was 38.1% according to the National Accounts (volume of labour accounting) compared to only 27% according to the LFS). German media coverage often focuses on the low number of hours actually worked, as estimated by the National Accounts, suggesting that "Germans work comparatively little" (Groll 2014). In contrast, in France, the debate generally focuses on the LFS estimates for full-time employees, suggesting much longer working times in Germany (+9.6%; Coe-Rexecode 2014).

Table 3 clearly demonstrates that much care must be taken in order to correctly interpret working-time indicators. When the aim is to identify the methodological effects leading to different working-time estimates, it is important to be able to distinguish the effects due to the actual measurement of working time in the survey and effects due to deviating structures for full-time and part-time employees. Against this backdrop, our article focuses on the measurement of hours actually worked in the LFS for the group of full-time employees. This not only makes the reference population more homogeneous, but also focuses on the group with the greatest impact on the average hours actually worked. Moreover, the focus on full-time employees makes it possible to make comparisons with regulations and collective agreements, for example, regarding the entitlements to paid leave and public holidays. It should be noted that the measurement of working time for part-time employees and self-employed is complicated for other reasons, which are however not elaborated on any further in this article (for a recent study, see Vallé et al. 2014).

Table 4 presents the distribution of hours actually worked in hour bands. The results already suggest some possible hypotheses regarding the methodological effects on the measurement of hours actually worked in France and Germany: France registers a higher proportion of full-time employees with 0-hour weeks (due to holidays, sickness or other reasons) or with less than 35 hours worked in the week. The modal value is 35 hours in France, the legal threshold beyond which any additional hours worked fall under the overtime regime. The German modal value is higher, equalling 40 hours per week. The differences of the distributions of hours actually worked in Germany and France, as shown in Table 4, require further investigation: full-time employees necessarily experience a certain number of shorter weeks (including weeks with zero hours) during the

Table 4. Number of hours actually worked by full-time employees in hour bands by reference week (full-time employees; LFS 2012)

Hours actually worked in the reference week	Germany		France	
	in 1,000	%	in 1,000	%
0 hours	2,973	11.5	2,711	14.6
1–19 hours	390	1.5	398	2.2
20–29 hours	531	2.1	1,306	7.0
30–34 hours	1,267	4.9	1,033	5.6
35 hours	1,131	4.4	4,333	23.3
36–39 hours	4,210	16.3	3,532	19.0
40 hours	8,916	34.5	1,722	9.3
41–49 hours	3,864	14.9	1,952	10.5
50 hours or more	2,575	10.0	1,582	8.5
Total	25,854	100.0	18,570	100.0

year for various reasons including paid holidays, bank holidays, sickness, and so on. The measurement of these nontypical weeks is crucial for estimating the average number of hours actually worked per week. The accuracy of these distributions has to be established in order to guarantee the robustness of the aggregates.

3.1. Absences Due to Sickness

Regarding absences due to sickness, both the French and the German LFSs deliver consistent messages: almost 3% of full-time employees were on sick leave for the whole reference week in 2012. The share of employees working a shortened week is also similar: around 1%. These results lead to an estimation of 1.7 weeks of absence due to sick leave per full-time employee in Germany and 1.5 in France in 2012 (Table 5). These results are close to the calculations in the National Accounts in Germany (based on statutory health insurance data), which amount to 1.8 weeks of absence due to sick leave (IAB 2014).

3.2. Absences Due to Holidays

The measurement of sick leave in Germany and France provides coherent figures in both countries. The results contrast much more strongly regarding absences for holiday leave. On average, in any given week, 6.1% of full-time employees are absent from work in Germany on holiday leave. This proportion is much higher in France: 10.3%. Working weeks shortened due to holiday leave are also more frequent in France (5.4%) than in Germany (2.0%).

These differences explain most of the discrepancies observed in Table 4: French full-time employees more frequently report unworked or shortened weeks than Germans do. The sum of 6.3 weeks of absences due to holidays is reported in the French LFS, whilst only 3.6 weeks are declared in the German survey (Table 6).

Is this because French employees have more holiday entitlement than Germans? There is no direct empirical evidence for this. However, the total number of estimated weeks of absence due to holiday leave per full-time employee in Germany seems conspicuously low: only 3.6 weeks, whereas the volume of labour accounts, based on information from collective bargaining agreements, assumes a value of around 6 weeks (see IAB 2014;

Table 5. Absences due to sickness (LFS, 2012)

	Full-time employees	with no work in reference week due to sick leave	with fewer working hours than usual in the reference week due to sick leave
Germany			
Persons (in 1,000s)	25,854	733	203
Persons (in %)	100	2.8	0.8
Average hours actually worked	35.8	0.0	21.0
Estimated sum of sick weeks (in 1,000)	43,225	38,116	5,109
Estimated weeks of absence due to sick leave per full-time employee	1.7	1.5	0.2
France			
Persons (in 1,000)	18,570	491	130
Persons (in %)	100	2.6	0.7
Average hours actually worked	32.4	0.0	20.7
Estimated sum of sick weeks (in 1,000s)	28,761	25,543	3,218
Estimated weeks of absence due to sick leave	1.5	1.4	0.2

Spitznagel 2003; Wanger 2013). According to the data compiled by the European Industrial Relations Observatory (EIRO), the number of days of paid leave is around 30 days in Germany (according to collective agreements) and significantly more than 25 days in France (statutory minimum), based on a five-day working week (Eurofound 2013, 18–19). Against these figures, the results for Germany presented in Table 6 seem implausible and indicate an underreporting of absences due to holiday leave in the German LFS.

Another factor explaining this situation is public holidays. We have analysed the weeks affected by a public holiday for both France and Germany. If French and German respondents respond accurately to the questions regarding their absences during the reference weeks, those whose reference week comprises a national public holiday should in a large majority report a shortened working week.

Table 7 shows that this is the case, but demonstrates also that the effect is much stronger in France than in Germany. In Germany, the working time during weeks affected by a national public holiday is only shortened by 7.6% compared to an average week. In France, the factor is multiplied by almost four: 26.5%.

Table 8 confirms the suspicion of a declarative bias: while 78.1% of the full-time employees declared reduced working hours in reference weeks affected by a nationwide public holiday in France, this percentage is almost halved in Germany (40.1%). German respondents declare fewer weeks with both total and partial absences from work. Similar measurement issues were also confirmed for the Italian LFS, in which the measurement of

Table 6. Absences due to holidays (LFS, 2012)

	Full-time employees	with no work in reference week due to holiday	with fewer working hours than usual in the reference week due to holiday
Germany			
Persons (in 1,000s)	25,854	1,580	516
Persons (in %)	100	6.1	2.0
Average hours actually worked	35.8	0.0	22.2
Estimated sum of holiday weeks (in 1,000)	94,356	82,160	12,196
Estimated weeks of absence due to holidays per full-time employee	3.6	3.2	0.5
France			
Persons (in 1,000)	18,570	1,913	1,009
Persons (in %)	100	10.3	5.4
Average hours actually worked	32.4	0.0	26.2
Estimated sum of holiday weeks (in 1,000s)	117,092	99,486	17,606
Estimated weeks of absence due to holidays	6.3	5.4	0.9

absences due to public holidays was considerably improved after the introduction of an improved questionnaire (Loriga and Spizzichino 2013).

These pieces of evidence strongly suggest the existence of a declarative bias in Germany in comparison with France. German respondents are less likely to mention their days off due to holiday than the French. However, this assumption has not been verified for absences due to sickness.

Table 7. Average hours actually worked by reference week with/without public holiday, full-time employees (LFS, 2012)

	Germany	France
Average	35.8	32.4
Weeks with national public holiday[(1)]	33.1	23.8
Diff (with average)	*−7.6%*	*−26.5%*
Weeks without any public holiday (neither national nor regional)	36.5	33.7
Diff	*1.7%*	*4.1%*

[(1)] Not coinciding with a weekend in 2012;

Germany: Good Friday, Easter Monday, Labour Day, Ascension Day, Whit Monday, National Holiday, Christmas Day.

France: Easter Monday, Labour Day, Armistice Day, Ascension Day, Assumption Day, All Saint's Day, Christmas Day.

Table 8. *Full-time employees actually working less than usual in weeks with national public holidays (LFS, 2012)*

			Germany	France
Weeks with nationwide public holiday				
total		in 1,000s	3,588	2,489
reduced working hours due to any reason		in 1,000s	1,440	1,944
		in %	40.1	78.1
of which	zero hours worked in the reference week	in 1,000s	496	617
		in %	13.8	24.8
	working time shorter than hours usually worked	in 1,000s	944	1,327
		in %	26.3	53.3

There are multiple interpretations to explain this phenomenon: a cultural bias making it easier to declare holiday in France than in Germany? Or perhaps the questionnaires and the sample design of these surveys are to blame (see Section 4)? The effects of this bias are clearly visible in Figure 1: the number of hours actually worked in Germany and France fluctuate throughout the year, with drops during the holiday seasons. However, the peaks are much more pronounced in France than in Germany, due to the greater rate of declaration of absences from work.

4. Understanding the Measurement Effects Leading to the Differences

The previous section revealed a number of clearly implausible results regarding the number of hours actually worked. In contrast to the German figures, in which absences due

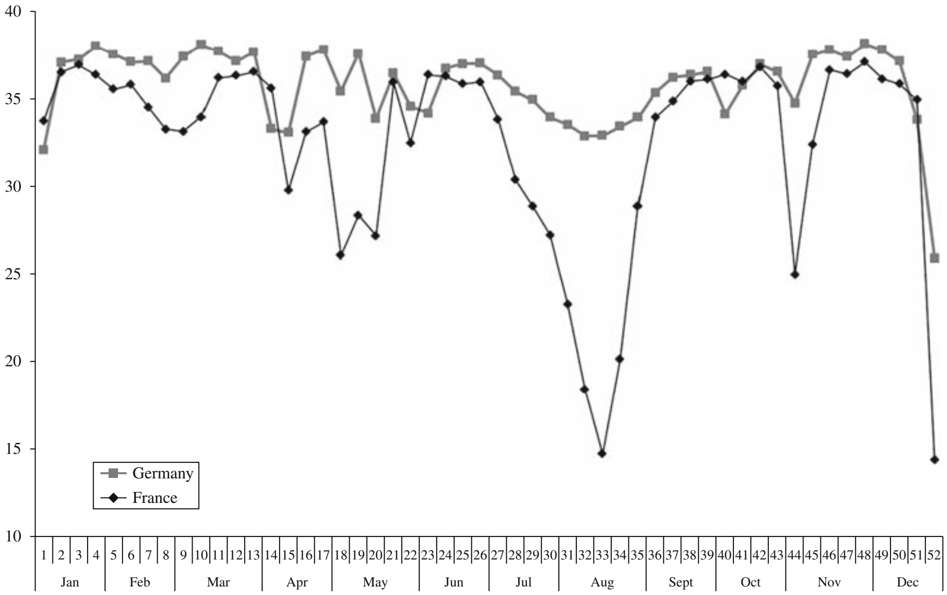


Fig. 1. *Number of hours actually worked by reference week (full-time employees; LFS, 2012)*

to paid leave and public holidays are conspicuously rare, the results for France correspond more or less to expectations based on the legal entitlement to sick leave and the specific calendar situation in the year 2012. Following the comparative assessment of the results on hours actually worked, this chapter tries to identify possible methodological sources that may contribute to the differences in the French and German results.

We have tried to take into account all major sources of error that may apply in the case of the LFSs in Germany and France. Based on the usual error typologies (see, e.g., Groves 1989; Biemer and Lyberg 2003; Groves et al. 2009; Eurostat 2009), possible sources of error include bias due to the sampling design, frame coverage bias, nonresponse bias, measurement bias as well as bias introduced in data processing (e.g., editing and weighting). Given the large sample size of the LFS, we do not focus on sampling errors, considered marginal in magnitude compared to the deviations identified in Section 3. The 95% confidence interval for the average hours actually worked amounts to +/− 0.1 hours in Germany and +/− 0.3 hours in France (annual results; see Eurostat 2014b).

Most of the potential sources of bias are neither easy to identify nor easy to quantify. The analysis of bias requires dedicated experiments that are demanding in terms of their design (and frequently do not meet the expectations) and focus on a reduced number of specific measurement effects. No specific experiments were carried out for the present study. The purpose is instead to give an overview of the major contributors to the astonishing results introduced in Section 3. This analysis is broadly based on the regular LFS data as well as related methodological studies in both countries.

4.1. Nonresponse Bias

Nonresponse bias is a widely researched source of error. Nevertheless, in the case of measuring working time in the LFS in France and in Germany, it is obviously not of major importance. Survey participation is mandatory in both countries, which results in response rates of more than 80% in France and slightly more than 98% in Germany. The level of response alone leaves relatively little room for nonresponse bias. Regarding the effects described in Chapter 3, it also does not seem plausible to assume that nonresponse plays a major role in these effects, that is, it is unlikely that employed persons who were on holiday in the reference week are less likely to participate in the LFS *ceteris paribus*.

4.2. Bias Due to the Sampling Design

The EU-LFS has been specifically designed to measure variables showing seasonal variations. The number of hours actually worked in the reference week is a variable that is subject to particularly strong seasonal and calendar-specific variation. Bank holidays, typical periods when paid leave is taken, as well as the winter season directly translate into a reduction in the number of hours worked. The design of the LFS has been developed to take this variation into account and to produce genuine quarterly and annual results. EU Regulation No. 577/1998 lays down the basic rule that “the reference weeks are uniformly distributed throughout the whole year”, that is, the number of sampling units per week is 1/52 in each calendar week. In the case of hours actually worked, an unequal distribution of the sampling units over the calendar weeks can lead to biased results. For example, smaller sample sizes in typical holiday weeks could contribute to an overestimation of the annual averages of hours actually worked (if not corrected for by the weighting scheme).

France and Germany apply different rules to allocate the reference weeks, so it might be assumed that these differences impact upon the differential results regarding the hours actually worked. While France, like most other member states, uses fixed reference weeks, Germany (like the Netherlands and Slovenia) relies on the principle of sliding reference weeks. When using fixed reference weeks, each sampling unit is allocated to a specific calendar week when the sample is selected. The interview will refer to this specific week, no matter when it takes place. Under the principle of sliding reference weeks, the interview always refers to the week directly preceding the week in which the interview takes place. While the principle of the sliding reference week facilitates fieldwork organisation and may reduce memory bias, it usually leads to a certain variation of weekly sample sizes. In the German case, the weekly sample size varies between 0.8% and 3.0% of the annual sample, while in France the variation is between 1.7% and 2.1%, close to the theoretical share of 1.9% (1/52). It should be noted that the variation in weekly samples in Germany, in addition to the effects of the sliding reference week, is also due to the fact that interviewers receive work packages with addresses to be interviewed 24 times a year (and not 52), that is, the distance between two work packages is two weeks, except for the end of each quarter when it is three. This contributes further to the uneven distribution and leads to a situation in which relatively few respondents are being interviewed at the end of each quarter (for details, see Körner and Puch 2011). Nevertheless, no significant correlation is observed between the weekly sample size and the average number of hours actually worked observed for each week. Even in the unweighted data, the correlation coefficient is no larger than 0.013, so that the effect of the numerical distribution of the sample over the weeks on working-time estimates is negligible (a rough estimation shows an effect of below 0.1 hours).

Still, the principle of the sliding reference week plays a more important role in another respect. It can be assumed that respondents are unlikely to make an appointment for a face-to-face interview in a week directly following an absence, for example, after one week of holiday. If the interview then “slides” to the next calendar week, the reference week will no longer be the one in which the respondent was absent. In other words, respondents who were on leave in the reference week would be systematically underrepresented in the sample. Based on the data available, it is unfortunately very difficult to analyse whether such effects might be empirically relevant. Some information can be gained from looking at the delay between the day the interviewers are provided with their work packages and the time the interviews are actually carried out. On average, 25.5% of the interviews take place in the first possible week after the distribution of a new work package. In weeks outside the usual holiday periods, this share amounts to nearly 30%, while it is only 23% if the interview week falls inside usual school holidays and 22% in weeks directly following the usual school holidays. This could be interpreted as an indication that respondents are more difficult to reach directly following an absence and that the reference week then “slides” to a week in which they are no longer absent. Although the share of interviews within the first week of receiving a new work package is not a sufficient basis on which to quantify potential effects (as nothing is known regarding the actual absences of these respondents), it suggests that the effect on estimates of hours actually worked might be substantial.

4.3. Measurement Bias

The most obvious source of error is measurement bias. Measurement bias includes a vast array of different error types, ranging from effects due to the data collection mode, effects due to proxy interviewing, effects due to the measurement instrument, and the interviewer. The conceptual backbone that helps respondents to understand the mechanisms leading to measurement bias is the cognitive process of answering survey questions. We will first present the data available on effects due to the data collection mode, proxy effects, and the structure and wording of the questionnaire, and then discuss these results in the context of the cognitive processes at play when formulating a survey response regarding the hours actually worked.

There are many reasons why the data collection mode may affect survey measurement. The type of social interaction and of communication, mode-specific questionnaire design options, as well as the use of computer assistance may contribute to differential measurement (for an overview, see Körner 2014). Studying mode effects requires a specific – ideally experimental – design, either with a split-sample, re-interview, or record-linkage design or at least with a reweighting of the data. None of these approaches was implemented for the present study. Despite this limitation, the comparison of LFS data can provide some insight as regards the hours actually worked, as data from an experimental study is able to corroborate the findings in the German case. In the case of France, CAPI is used in the first interview, while CATI is the predominant mode in follow-up interviews, so that the groups can be considered to be roughly structurally equivalent.

Both the German and the French data show a slight but significant difference between CAPI and CATI (see Table 9). The differences are larger for the self-administered paper-and-pencil questionnaire that is used in Germany, but not in France. In the case of Germany, it could be argued that the telephone interviews, as well as the use of the self-administered PAP questionnaire, concern rather specific groups. The results of an experimental study on mode effects largely confirm these results (Körner and Liersch 2014): while there were almost no differences between CATI and CAPI in this experiment, the average hours actually worked determined from the PAP (35.2 hours) was higher than in CAPI or CATI (34.3). However, these differences were still not significant, which might also be due to the limited sample size in this experiment.

Table 9. Hours actually worked per week by data collection mode, 2012 (full-time employees)

	Germany		France	
	hours	%	hours	%
Total	35.8	100	32.4	100
CAPI	35.5	66	32.2	32
CATI[1]	36.0	9	32.5	68
Self-administered PAP	36.7	25	N.A.	

[1] *Germany: Telephone interviews administered by field interviewers and "passive" telephone interviews administered by the state statistical office (both only partly computer assisted).*
France: CAPI and CATI administered by Insee field interviewers. First and last contacts with interviewees have to be CAPI, and interviews in-between (hence Waves 2, 3, 4, and 5) are administered by telephone by the interviewer who has established the contact (unless the household insists on continuing face-to-face).

Table 10 provides further insight into the different measurements underlying the different averages. While once again the distributions of data collected via CAPI and CATI are very close in both countries (no significant differences), PAP respondents are more likely to report a weekly working time of more than 40 hours. Correspondingly, the share of PAP respondents reporting 20 to 39 hours is significantly lower. Given that the share of PAP cases in the German LFS is 25%, the effect on the overall average of the hours actually worked is smaller than 0.3 hours (under the unrealistic assumption that the entire difference is due to a measurement effect).

Effects due to proxy responding are another obvious effect in household surveys, that is, answers provided to questions that refer to another member of the household. It is often assumed that proxy answers are less accurate, mainly because the proxy respondents do not have the necessary information to answer survey questions properly at their disposal. Despite the popularity of using proxy effects to explain the issues experienced with measurements, very few experimental studies have been carried out in an LFS context, and those that do exist rarely focus on working time. One exception is the study by Ole Villund (2009, 5), who concluded in a record-linkage study regarding the contractual hours of work that "proxy interviews seem to cause more overestimation of contractual working hours, especially in jobs with varying contractual working hours and long hours." The data from the LFS only partly confirm this finding: As shown in Table 11, the average hours actually worked by full-time employees are higher in the case of proxy answers in France as in Germany. The difference is 1.4 hours in both countries. Again, these findings should be treated with caution as proxy interviews are not randomly allocated in groups in the LFSs, so some of the difference might be in differential working-time patterns of proxy respondents.

Table 12 clarifies that the differences in the average are due to a greater emphasis on the usual number of hours and not simply an overestimation of working hours in proxy interviews. The responses given by proxy respondents tend towards the modal value (40 hours in Germany and 35 hours in France) more frequently than in direct interviews. Similarly, proxy interviewees more rarely indicate that the reference person did not work at all in the reference week, a finding that again is consistent in both countries. Finally,

Table 10. Full-time employees by number of hours actually worked and data collection mode, 2012 (percentage of all full-time employees)

	Germany			France	
	PAP %	CAPI %	CATI[(1)] %	CAPI %	CATI %
Did not work in reference week	11.4	11.5	11.5	14.9	14.5
1–19 hours	1.3	1.6	1.2	2.4	2.0
20–29 hours	1.4	2.3	2.1	7.2	6.9
30–34 hours	3.4	5.4	5.1	5.8	5.5
35 hours	3.8	4.5	4.9	22.2	23.9
36–39 hours	14.2	17.0	16.9	18.9	19.1
40 hours	34.3	34.7	33.7	9.1	9.4
41–49 hours	18.6	13.7	14.2	11.0	10.3
50 hours or more	11.6	9.3	10.6	8.5	8.5

[(1)] *Germany: Telephone interviews administered by field interviewers and "passive" telephone interviews administered by the state statistical office (both only partly computer assisted).*

Table 11. Hours actually worked per week by proxy interview, 2012 (full-time employees)

	Germany		France	
	hours	%	hours	%
Total	35.8	100	32.4	100
Proxy	36.8	25	33.4	29
Non proxy	35.4	75	32.0	71

particularly long working hours are indicated more rarely in proxy interviews, at least in the case of Germany. These results suggest that the question on hours actually worked is sensitive to proxy effects and can have a substantial impact on the implausible results presented in Section 3. However, it can explain only part of the differences between France and Germany, as the proxy rate in both countries is similar (see Chapter 2).

Mode effects and proxy effects, as well as other types of measurement bias, can originate from any of the cognitive steps in the survey response. The standard cognitive steps of survey response are considered to be: comprehension, retrieval, judgement, and reporting (Tourangeau et al. 2000; for an overview, see Biemer and Lyberg 2003, 123–148). Applied to questions on working time, these steps can be presented as follows: the respondent needs to *understand* the question, in particular the underlying concepts, that is, the items to be included and excluded as well as the reference concepts (reference week, focus on the main job). Having understood which information he or she is expected to provide, the respondent needs to *retrieve* the necessary information, that is, remember the reference week in question, the working hours undertaken, including any absences from work or overtime. In a household survey, respondents will usually have to rely on their memory to perform this task, although it would probably be useful if they checked with their calendar. In the *judgement* step, the respondent needs to assess the information retrieved regarding its completeness and relevance to the question, taking into account the required response format. In the case of working-time questions, this includes some kind

Table 12. Full-time employees by number of hours actually worked and proxy/direct interview, 2012 (percentage of all full-time employees)

	Germany		France	
	Proxy interview %	Direct interview %	Proxy interview %	Direct interview %
Did not work in reference week	8.8	12.5	13.0	15.3
1–19 hours	1.1	1.7	1.5	2.4
20–29 hours	1.8	2.2	6.3	7.3
30–34 hours	5.1	5.0	4.4	6.1
35 hours	4.0	4.6	28.0	21.3
36–39 hours	16.1	16.5	17.8	19.5
40 hours	42.7	31.5	9.4	9.2
41–49 hours	10.9	16.1	9.8	10.8
50 hours or more	9.5	9.9	9.8	8.0

of calculation, either by adding up the daily working time or, more frequently, by deducting absences from a usual week and adding on any overtime. Finally, the respondent needs to provide the *response*, which may deviate from the result obtained in the judgement phase. In working-time questions respondents commonly round off, but the respondent may also over- or understate his or her working time with regard to social norms. It should be noted that the cognitive steps are distinguished analytically but often take place simultaneously.

The French and the German questionnaires have found different solutions to help the respondent provide the correct response. Starting with the comprehension stage, the German questionnaire, following the question ("How many hours did you actually work in the last week?"), restricts the explanation of the concepts to a short instruction saying "The number of hours actually worked may differ from the hours usually worked because of overtime, holidays, extra shifts, public holidays, illness and the like. The number of hours actually worked includes continuing and advanced training, stand-by duty, work done at home provided that it is a normal part of your job, such as for teachers." Note that, despite its length, this instruction already omits some of the elements outlined in Section 2, tacitly assuming that they correspond to a common perception (such as the distinction between coffee breaks and lunch breaks). In the context of an interviewer-administered survey, it is probably doubtful that this instruction would be read out to the respondent in any case. According to the concept of sliding reference weeks, the question refers to the reference week as "last week", which omits the specific date. The question does not contain a specific cue to remind the respondent that the response should refer only to their main job (provided once at the beginning of the module on the characteristics of the main job). This is an important element, as recent cognitive tests have shown that multiple job-holders have a tendency to sum up the working hours in all of their jobs (Vallé et al. 2014).

The French questionnaire has similar instructions, but introduces the main components of the working-time concepts in a series of dedicated questions asking specifically about absences due to paid leave, compensated leave, or sick leave, slack periods of work, training, labour disputes and extra time worked due to overtime hours. The duration of each type of absence and overtime is measured with dedicated questions. It is only after a series of about 20 questions has been answered that the question on hours actually worked appears ("(In total) In the week from Monday [. . .] to Sunday [. . .], how many hours did you actually work in your main job?"). The instruction below the question on hours actually worked therefore serves simply as a reminder: "Do not count hours or days of ordinary holiday, special leave, public holidays, compensated leave, unpaid leave, partial unemployment, education and training, strike, labour dispute". It is easy to imagine that the respondent will be more familiar with the deductions and additions needed to come up with a correct reply regarding the hours actually worked. In contrast to the instruction in the German questionnaire, the French one mentions only deductions, and not any overtime that would have to be included in the answer. Further differences are that the precise dates of the reference week are mentioned in the French questionnaire ("from Monday, . . . , to Sunday, . . .") and that the French question explicitly specifies the reference to the main job. Table 13 provides the original question wording as well as an English translation.

It seems obvious that the questions regarding absences and overtime applied in the French questionnaire not only help to ensure that the complex concept is understood, but

Table 13. Questions on hours actually worked in the LFS questionnaires in France and Germany (2012)[1]

Germany	France
Original version	
Wie viele Stunden haben Sie in der letzten Woche tatsächlich gearbeitet? Die tatsächliche Arbeitszeit kann von der normalerweise geleisteten Arbeitszeit abweichen, zum Beispiel wegen Überstunden, Urlaubstagen, Sonderschichten, Feiertagen, Krankheit o. Ä. Zur tatsächlichen Arbeitszeit gehören auch Weiter- und Fortbildungen, Bereitschaftszeiten, Arbeiten von zu Hause, sofern sie Bestandteil Ihrer Erwerbstätigkeit sind, z. B. bei Lehrkräften.	(Au total) La semaine du lundi . . . au dimanche . . . , combien d'heures avez-vous effectuées dans votre emploi principal? (Ne pas compter les heures ou jours de congés ordinaires, exceptionnels, fériés, ponts, RTT, récupération, congé personnel non rémunéré, chômage partiel, activité de formation, grève, conflit du travail)
English translation	
How many hours did you actually work in the last week? The number of hours actually worked may differ from the hours usually worked because of overtime, holidays, extra shifts, public holidays, illness and the like. The number of hours actually worked includes continuing and advanced training, stand-by duty, work done at home provided that it is a normal part of your job, such as for teachers.	"(In total) In the week from Monday [. . .] to Sunday [. . .], how many hours did you actually work in your main job?" Do not count hours or days of ordinary holiday, special leave, public holidays, compensated leave, unpaid leave, partial unemployment, education and training, strike, labour dispute

[1] *The full LFS questionnaires are available at http://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey_-_methodology#Core_questionnaires*

also facilitate the retrieval of the necessary information. It is however difficult to obtain direct empirical evidence on the cognitive process from quantitative data. Nevertheless, a qualitative pretest of a working-time questionnaire module inspired by the French approach carried out at the Federal Statistical Office Germany indicates that a vast majority of respondents found it helpful for the question on hours actually worked to be introduced after a sequence of questions of absences and extra time. However, only few respondents in this test were actually absent in the reference week (Vallé et al. 2014).

Another important potential source of error at the retrieval stage is memory effect: the longer the time span between the reference week and the interview, the more difficulty respondents will have in comprehensively retrieving the necessary information. For this reason, the legal act mandating the LFSs (Council Regulation No. 577/1998) stipulates that "the interview normally takes place during the week immediately following the reference week". As in the German design the reference week is by definition the week before the interview, memory effects can only be analysed for the French LFS. After the cognitive step of information retrieval, the steps of judgement and reporting follow, and in both cognitive processes the phenomenon of rounding is observed. As shown in Figure 2, there is clear empirical evidence that respondents make use of rounding when preparing and reporting their responses. Rounding is obviously not restricted to the next whole number, but extends to numbers divisible by 5: Conspicuous peaks can be seen for all numbers ending in a "5" or a "0", in both Germany and France. While one might argue that the modal values (35 in France and 40 in Germany) could actually represent the most frequent weekly working time, the peaks for the rest of the distribution clearly indicate the effect of rounding. Even the high percentage of respondents indicating 40 hours in Germany does not seem entirely plausible: the average contractual hours of work laid down in collective agreements in 2012 were 38.3 hours, and almost all economic activity groups had working hours fixed by collective agreements of less than 40 hours (except for the NACE rev. 2 divisions "construction of buildings" and "civil engineering", in which working hours equalled 40 hours; see Statistisches Bundesamt 2014). In Germany, one would expect higher percentages for 38 and 39 hours and a lower percentage for 40 hours. For France, 35 hours is a national legal threshold beyond which the employer has to pay overtime. Since 2008 however, branches and firms have the possibility of renegotiating this threshold at their convenience. Many branches have done so, so it is now hard to give an average duration. According to an estimation by the *Direction de l'animation de la recherche, des études et des statistiques* (DARES), the average collective agreed working hours in France amount to 35.6 hours (Eurofound 2013). Beyond that, the "35-hour" theme has become a political slogan and is omnipresent in public debate. This is why it is no surprise that there is a peak in the responses at 35 hours (Figure 2). It is still difficult to distinguish between the actual hours and the rounding effects.

Although the rounding effect is clear both for France and for Germany, it seems to be slightly more frequent in Germany: 54.2% of the responses given to the question on hours actually worked end in a "5" or a "0", while this is the case for only 47.9% of the responses

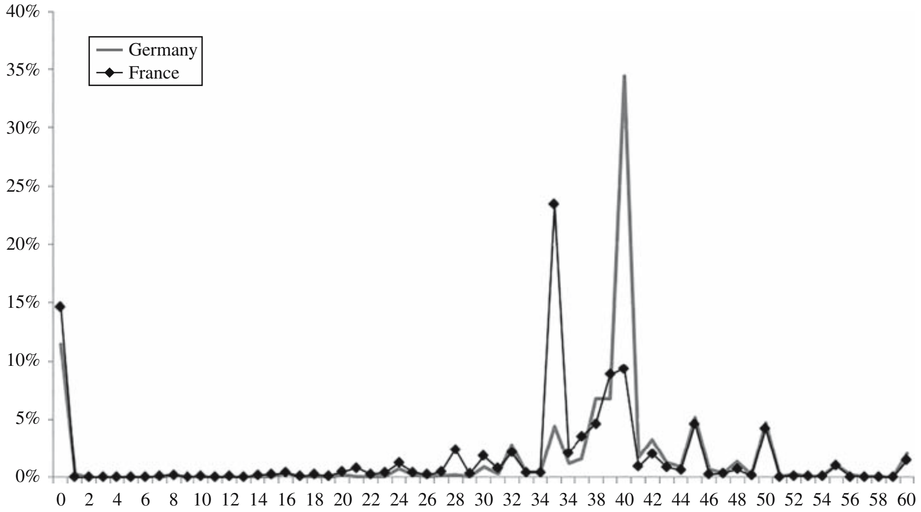


Fig. 2. Distribution of hours actually worked (full-time employees, 2012)

in France. Greater use of rounding in Germany can also be assumed when comparing the questionnaires: While respondents in Germany are required to answer in whole numbers and are explicitly invited to round ("Round up or down to the nearest full hour"), respondents in the French LFS are asked to give their answer to one decimal place.

Rounding off may explain most of the difference between the French and German results. Given the working-time provisions in the law as well as in collective agreements, it seems likely that the peaks for 35 hours in France and 40 hours in Germany suggest that rounding in France usually means rounding down, while in Germany it is more likely to mean rounding up.

4.4. *Editing Effects*

The data provided by the respondents might also be influenced by edit checks applied during data collection. Both Germany and France apply edit checks during the interview itself, using both soft checks (asking to verify the response given) and hard checks (requiring a change in the response given in order to be able to proceed with the interview). In the context of measuring hours actually worked in Germany, two plausibility checks are particularly relevant. Both refer to the measurement of the employment status in the reference week (EU-LFS variable WSTATOR), for which respondents have to indicate whether they worked for at least one hour in the reference week (WSTATOR = 1) or whether they had a job from which they were absent in the reference week (WSTATOR = 2). For respondents coded WSTATOR = 2, in Germany there was a hard check applied to the question on the number of hours actually worked: persons who answered that they worked for at least one hour in the reference week were not allowed to answer "0" to the question on hours actually worked, while those who reported having not been at work in the reference week were requested to report zero hours. Both checks were implemented as hard checks, that is, the respondent could only continue the interview after having corrected his or her response. For instance, the first plausibility check might be problematic: the absence from the job is not the main purpose of the variable WSTATOR, but was initially included to make sure that all employed persons were captured, even if temporarily not at work. Furthermore, in the German translation of "having worked last week", it is not entirely clear whether this includes absences, for example, due to paid leave (Nebel et al. 2013). Therefore it seems likely that the questions on the employment status do not entirely cover the absence of employed persons in the reference week. As a result, respondents might be pushed to omit information on absences in the reference week when asked for the hours actually worked, much later in the questionnaire. It is difficult to find empirical data to substantiate this speculation. Some, albeit limited, insight can be gained from an experimental study on measurement effects due to the data collection mode, carried out by the Federal Statistical Office together with eight State Statistical Offices in 2009 and 2010 (see Körner and Liersch 2014; Statistisches Bundesamt 2010). In this split-sample experiment, the edit checks were applied to the interviewer-administered data collection modes (CAPI and CATI), but not to the self-administered modes. In a tabulation of the experimental data set produced for this article, it was shown that the answers provided to both questions were inconsistent in 5% to 16% of the responses when no edit checks were performed. While the PAP self-administered questionnaires show no clear pattern, the online questionnaire tested among the implausible answers contains almost only cases that stated they did work in the reference week, but reported zero hours when asked for the hours actually worked. The difference in the percentage of full-time employees having worked zero hours in the reference weeks in the online mode was 11% without edit checks compared to 7% with edit checks, and 12% and 9% respectively in the self-administered PAP mode (note that the reference period did not cover the entire year, so a comparison with regular LFS data is not possible). Despite the small sample size and the corresponding low statistical power, this result suggests that effects of this type could have a substantial impact. In France, no hard checks have been implemented for these questions: it is possible for a respondent to simultaneously declare having worked and declare a duration of zero hours. Regarding this specific issue, the French questionnaire is more flexible than the German one. The opposite, however, is not possible in France: a respondent declaring not having worked during the reference week cannot later in the questionnaire give a number of hours worked. To the best of our knowledge, the impact of this filter has not been tested: respondents may give inconsistent answers and give a positive number of hours worked, despite their answer to WSTATOR (for part-time and casual work, for instance).

5. Conclusions

The various issues raised in this contribution show that any comparison of hours actually worked in Germany and France is a complicated affair, since the conditions for ensuring the robust comparability of data are not met.

The main issue is with the measurement of hours not worked during the reference week: given the structures of questionnaires, the effects of rounding off, the sampling design, and the use of data editing, the way this information is collected is heterogeneous and leads to divergent estimates. The study of French and German cases clearly highlights this limitation. The differences in Germany and France regarding the measurement of working time and absences are a concern not only in these two countries, but should be raised for all EU member states. The differences shown here between Germany and France are in fact rather slight when compared with the results for countries like Romania (with only 1.9% of full-time employees totally absent from work in the reference week), Bulgaria (2.9%), Greece (3.5%) or Hungary (4.0%). If taken seriously, these statistics demonstrate that employees in these countries very rarely take holidays or days off from work at all. Conversely, countries like Sweden (15.6%), Finland (15.0%) and France (14.6%) show the highest rates of full-time employees absent during the reference week. In this regard, Germany appears to be amongst the countries where the underreporting of absences is the lowest (see Table 14 in the Appendix).

Overall, as French full-time employees seem to work 3.4 hours less than Germans (respectively 35.8 hours and 32.4 hours actually worked in the reference week in Germany and France), the difference is reduced to 2.0 hours when only considering respondents who worked not less than usual in the reference week (41.9 hours in Germany and 39.9 hours in France). In the same way, the difference regarding the average hours usually worked is much smaller: regarding this concept, which focuses on a typical week (thus largely disregarding any periods of absence), the average working time in Germany (40.7 hours) is only 1.3 hours more than in France (39.4). These uncertainties demonstrate that the international comparability of hours actually worked is limited in the LFS data and that the results require very careful interpretation. However, when making comparisons of the results for the hours usually worked it needs to be borne in mind that the measurement of this variable might be subject to specific measurement effects as well, which cannot be dealt with in the context of this contribution.

This methodological issue can have an impact on the public debate. Unless sufficient care is taken, comparisons based on these data can lead to false conclusions regarding "insufficient" or "excessive" working time either in France or Germany. Given the importance of these issues, we would like to emphasise the need to make the data collected throughout Europe more comparable, notably concerning the measurement of the hours actually worked. We hope that this contribution will help to push this issue further up the European agenda and help countries to find a convergent position.

Within the European Statistical System, the first conclusions have already been drawn: in April 2013, a Task Force on the Measurement of Absences and Working Time was created by Eurostat and the EU member states. The Task Force's main objective is to develop a model questionnaire to help improve and harmonise the measurement of the different working-time concepts in the EU-LFS. The draft model questionnaire addresses many of the issues raised in this contribution: it is based on a similar approach to that used in France, but is considerably simplified in order to ensure its implementation is feasible in all member states. The question on the number of hours actually worked in the reference week is preceded by a series of questions regarding absences from work due to public holidays, holidays and time off in lieu, as well as sick leave and other reasons. Two further questions focus on overtime in the reference week, which also needs to be considered when asking for the number of hours actually worked. A first round of empirical tests showed promising results but revealed a number of different problems, for instance in the case of irregular working hours, shift work, and part-time work. A revised version of the model questionnaire is currently being tested in other member states (Eurostat 2014a). Although this work is considered extremely important to improve cross-national harmonisation, it should be noted that a model questionnaire alone cannot guarantee comparability of the results. As we have shown, differences in the data collection modes, the share of proxy interviews, as well as the allocation of the reference weeks can have substantial impact on the measurement of the hours actually worked as well. It should be investigated further whether further harmonisation is possible regarding these critical elements, or at least whether experimental studies could be established to quantify the magnitude of the effects. With regard to the measurement of working time, it would in particular be desirable to reduce the rate of proxy interviews. Regarding mode effects, the introduction of web interviewing in LFSs (currently fostered by many national statistical institutes) may be a particular challenge. Further experimental research is needed to learn more about the potential impact of web interviewing on working-time estimates.

In the meantime, we would encourage users to make cautious use of the variables related to working time in the LFS, in particular the hours actually worked. Until further methodological improvements have been made, it is difficult to draw robust conclusions from these data, for instance regarding international comparisons. For international comparisons, users should use information regarding the hours usually worked or refer to the hours actually worked for respondents who worked at least one hour in the reference week (current publication at the Eurostat online database).

However, this contribution does demonstrate that analyses of LFS data are possible and can be highly profitable. The results shown here would benefit from being replicated in all member states. The issues might differ from a country to another, as the questionnaires and sample designs for the LFSs are only partially comparable throughout Europe.

Cognitive tests and qualitative approaches to investigate respondents' behaviour, in a comparative framework, together with quantitative experimental studies, would also improve our understanding of these surveys and their divergences.

Appendix

Table 14. Percentage of employed persons having worked less than usual in the reference week, full-time employees (LFS, 2012)

	Hours actually worked		
	0 hours %	more than 0 hours %	Average hours actually worked
Austria	12.2	20.4	35.3
Belgium	10.5	10.7	34.2
Bulgaria	2.9	5.3	39.1
Croatia	5.9	7.5	37.9
Cyprus	6.4	11.4	37.7
Czech Republic	7.6	12.7	37.2
Denmark	13.2	12.1	33.5
Estonia	6.3	14.0	37.5
EU-28	9.8	12.5	35.8
Finland	15.0	21.9	31.9
France	14.6	18.9	32.2
Germany	11.5	10.7	35.8
Greece	3.5	5.9	38.6
Hungary	4.0	10.8	37.9
Ireland	8.9	10.0	34.9
Italy	9.4	7.6	34.5
Latvia	6.0	3.8	37.5
Lithuania	7.3	2.5	36.7
Luxembourg	12.0	7.1	36.2
Malta	5.4	17.6	37.5
Netherlands	11.6	13.9	35.1
Poland	6.2	9.8	37.6
Portugal	9.4	13.1	36.4
Romania	1.9	0.8	40.4
Slovakia	6.4	13.3	37.1
Slovenia	12.3	13.1	34.7
Spain	9.5	10.8	35.7
Sweden	15.6	21.7	32.6
United Kingdom	9.0	18.5	37.4

Source: Eurostat, own calculations.

6. References

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

Blanke, K. and A. Luiten. 2014. *Query on Data Collection for Social Surveys. ESSnet Project "Data Collection for Social Surveys using Multiple Modes"*. Available at: http://www.cros-portal.eu/content/data-collection (accessed 18 December 2014).

Coe-Rexecode. 2012. *La durée effective du travail en France et en Europe*. Document de Travail no. 29. Paris: Coe-Rexecode. Available at: http://www.coe-rexecode.fr/public/content/download/32364/323546/version/2/file/Document-de-travail-Coe-Rexecode-Duree-effective-du-travail-Fance-Europe-2012-29.pdf (accessed 8 April 2016).

Coe-Rexecode. 2014. *La durée effective du travail en France et en Europe*. Résultats de 2013 et mise à jour de l'étude de 2012. Document de Travail no. 49. Paris: Coe-Rexecode. Available at: http://www.coe-rexecode.fr/public/content/download/34852/351622/version/3/file/Doc-trav-49-Duree-du-travail-France-et-Europe-2013-Juin-2014.pdf (accessed 8 April 2016).

Eurofound. 2013. *Developments in Collectively Agreed Working Time 2012*. Dublin: European Foundation for the Improvement of Living and Working Conditions. Available at: http://www.eurofound.europa.eu/docs/eiro/tn1305017s/tn1305017s.pdf (accessed 8 April 2016).

Eurostat. 2009. *ESS Handbook for Quality Reports*. Luxembourg: Office for Official Publications of the European Communities. Available at: http://unstats.un.org/unsd/EconStatKB/Attachment286.aspx?AttachmentType=1 (accessed 8 April 2016).

Eurostat. 2013. *Labour Force Survey in the EU, Candidate and EFTA countries. Main Characteristics of National Surveys, 2012*. Luxembourg: Eurostat. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-TC-13-003 (accessed 13 August 2014).

Eurostat. 2014a. *Interim Report of the Task Force on Measurement of Absences and Working Time*. Working Group on Labour Market Statistics (LAMAS), June 2014. Document Eurostat/F3/LAMAS/07/14. Luxembourg: Eurostat.

Eurostat. 2014b. *Quality Report of the European Union Labour Force Survey 2013*. Luxembourg: Eurostat. Available at: http://ec.europa.eu/eurostat/documents/3888793/6194252/KS-TC-14-010-EN-N.pdf/39a79a33-4442-49b6-b83f-f0a5e81b02ef (accessed 29 June 2015).

Fleck, S. 2009. "International Comparisons of Hours Worked: an Assessment of the Statistics." *Monthly Labor Review* 5: 3–31.

Frazis, H. and J. Stewart. 2004. "What Can Time-Use Data Tell Us About Hours of Work?" *Monthly Labor Review* 127: 3–9.

Groll, T. 2014. "Arbeitszeit: Deutsche arbeiten vergleichsweise wenig." *Zeit Online Beruf* 13 May 2014. Available at: http://www.zeit.de/karriere/2014-05/arbeitszeit-oecd-info-grafik (24 October 2014).

Groves, R. 1989. *Survey Errors and Survey Costs*. Hoboken, NJ: Wiley.

Groves, R. et al. 2009. *Survey Methodology*, 2nd ed. Hoboken, NJ: Wiley.

IAB. 2014. *Durchschnittliche Arbeitszeit und Ihre Komponenten in Deutschland*. Nuremberg: Institut für Arbeitsmarkt- und Berufsforschung (IAB). Available at: http://doku.iab.de/arbeitsmarktdaten/AZ_Komponenten.xlsx (accessed 21 October 2014).

International Labour Organization (ILO). 2008. *Resolution Concerning the Measurement of Working Time*. Adopted by the Eighteenth International Conference of Labour Statisticians, November-December 2008. Geneva: International Labour Organization. Available at: http://www.ilo.org/global/statistics-and-databases/standards-and-guidelines/resolutions-adopted-by-international-conferences-of-labour-statisticians/WCMS_112455/lang–en/index.htm (accessed 11 August 2014).

Körner, T. 2012. "Measuring the Labour Status in Official Statistics: The Labour Force Concept of the International Labour Organisation and its Implementation in the Labour Force Survey." In *Demographic Standards for Surveys and Polls: National and European Dimensions*, edited by J. Hoffmeyer-Zlotnik and U. Warner, 123–138. Cologne: GESIS 2012.

Körner, T. 2014. "Report on the Definition, Identification and Analysis of Mode Effects." Deliverable for Work Package III of the ESSnet on Data Collection for Social Surveys Using Multiple Modes, Wiesbaden: Federal Statistical Office Germany. Available at: http://www.cros-portal.eu/content/data-collection (accessed 18 December 2014).

Körner, T. and A. Liersch. 2014. "Case Study on Mode Effects in the Germany Labour Force Survey." Deliverable for Work Package III of the ESSnet on Data Collection for Social Surveys Using Multiple Modes. Wiesbaden: Statistisches Bundesamt. Available at: http://www.cros-portal.eu/content/data-collection (accessed 18 December 2014).

Körner, T. and K. Puch. 2011. *Coherence of German Labour Market Statistics*. Vol. 19 of *Statistics and Science*. Wiesbaden: Statistisches Bundesamt.

Lefèvre, L. J. Rakotomalala, and F. Toutlemonde. 2012. "Méthodologie des comptes annuels de l'emploi, des heures travaillées et de la durée du travail dans la base 2005 de la comptabilité nationale." Note méthodologique des comptes nationaux, Insee.

Loriga, S. and A. Spizzichino. 2013. "Working hours: Analysis of Italian LFS Results versus Administrative Data and Business Survey." Paper presented at the 8th Workshop on Labour Force Survey Methodology, Gdansk, Poland, 23–24 May 2013. Available at: http://old.stat.gov.pl/lfs2013/papers/F3_Silvia_Loriga,Andrea_Spizzichino_IT.pdf (accessed 24 October 2014).

Massarelli, N. 2011. "Harmonisation Issues for the Measurement of Employment and Unemployment." Paper presented at the 6th Workshop on LFS Methodology, Wiesbaden, Germany, 12–13 May 2011. Available at https://www.destatis.de/EN/AboutUs/Events/LFS/PapersP/G1_HarmonisationIssues_Massarelli.pdf?__blob=publicationFile (accessed 24 October 2014).

Mata Greenwood, A. 2001. "The Hours that We Work: The Data We Need, the Data We Get." *ILO Bulletin of Labour Statistics* 2001-1

Nebel, S. et al. 2013. *Implementation Study of the Model Questionnaire for the Measurement of Employment and Unemployment and the Variables of the ad hoc Module 2015 on Work Organisation and Working Time Arrangements*. Final report, European Commission Grant Agreement No. 10201.2012.001-2012.961, Wiesbaden: Statistisches Bundesamt.

Niemi, I. 1993. "Systematic Error in Behavioural Measurement." *Social Indicators Research* 30: 229–244.

OECD. 2004. *OECD Measures of Total Hours Worked*. The OECD Production Database, March 2004. Available at www.oecd.org/std/productivity-stats/29867131.pdf (accessed 10 August 2014).

Robinson, J. et al. 2011. "The Overestimated Workweek Revisited." *Monthly Labor Review* 134: 43–53.

Robinson, J. and A. Bostrom. 1994. "The Overestimated Workweek? What Time Diary Measures Suggest." *Monthly Labor Review August 1994*: 11–23.

Spitznagel, E. 2003. "Hours and Volume of Work in Germany The IAB Concept of Measurement." Paper presented at the Paris Group Meeting, 4–5 September 2003, London. Available at: http://www.insee.fr/en/insee-statistique-publique/colloques/citygroup/pdf/Germany-Session2.pdf (accessed August 2014).

Statistisches Bundesamt. 2010. *Ergebnisse des Projektes Q-MED/LFS. Quantifizierung von Methodeneffekten unterschiedlicher Erhebungsinstrumente auf die Datenqualität im Labour Force Survey*. Ein Gemeinschaftsprojekt der Statistischen Ämter des Bundes und der Länder zur Qualitätssicherung im Mikrozensus/Labour Force Survey. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt. 2014. *Verdienste und Arbeitskosten. Index der Tarifverdienste und Arbeitszeiten 2. Vierteljahr 2014*. Fachserie 16 Reihe 4.3, Wiesbaden: Statistisches Bundesamt. Available at: https://www.destatis.de/DE/Publikationen/Thematisch/VerdiensteArbeitskosten/Tarifverdienste/Tarifverdienst2160430143224.pdf?__blob=publicationFile (accessed September 2014).

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

Vallé, J. et al. 2014. *Implementation Study of the Model Questionnaire for the Measurement of Absences and Working Time and the ad hoc Module 2016 on Young People on the Labour Market*. Final report, European Commission Grant Agreement no. 07131.2013.001-2013.372, Wiesbaden: Statistisches Bundesamt.

Villund, O. 2009. *Measuring Working Hours in the Norwegian Labour Force Survey: A Pilot Study of Data Quality Using Administrative Registers*. Report 2009/3, Oslo–Kongsvinger: Statistisk sentralbyrå.

Wanger, S. 2013. "Arbeitszeit und Arbeitsvolumen in Deutschland – Methodische Grundlagen und Ergebnisse der Arbeitszeitrechnung." *Wirtschafts- und Sozialstatistisches Archiv* 7: 31–69.

Williams, R. 2002. "Hours Worked: A Comparison of Estimates from the Labour Force and New Earnings Survey." *Labour Market Trends* 110: 429–441.

Williams, R. 2004. "Investigating hours worked measurements." *Labour Market Trends* 112: 71–80.

Legal Texts

Commission Regulation (EC) no 377/2008 of 25 April 2008 Implementing Council Regulation (EC) no 577/98 on the Organisation of a labour Force Sample Survey in the Community As Regards the Codification to Be Used for Data Transmission from 2009 Onwards, the Use of a sub-sample for the Collection of Data on Structural Variables and the Definition of the Reference Quarters. Official Journal of the European Union, 26 April 2008, L 114/57.

Council Regulation (EC) no 577/98 of 9 March 1998 on the Organisation of a labour Force Sample Survey in the Community. Official Journal of the European Union, 14 March 1998, L 77/3.

Received December 2014
Revised July 2015
Accepted September 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 433–459, http://dx.doi.org/10.1515/JOS-2016-0023

Random Walks on Directed Networks: Inference and Respondent-Driven Sampling

Jens Malmros[1], *Naoki Masuda*[2], *and Tom Britton*[3]

Respondent-driven sampling (RDS) is often used to estimate population properties (e.g., sexual risk behavior) in hard-to-reach populations. In RDS, already sampled individuals recruit population members to the sample from their social contacts in an efficient snowball-like sampling procedure. By assuming a Markov model for the recruitment of individuals, asymptotically unbiased estimates of population characteristics can be obtained. Current RDS estimation methodology assumes that the social network is undirected, that is, all edges are reciprocal. However, empirical social networks in general also include a substantial number of nonreciprocal edges. In this article, we develop an estimation method for RDS in populations connected by social networks that include reciprocal and nonreciprocal edges. We derive estimators of the selection probabilities of individuals as a function of the number of outgoing edges of sampled individuals. The proposed estimators are evaluated on artificial and empirical networks and are shown to generally perform better than existing estimators. This is the case in particular when the fraction of directed edges in the network is large.

Key words: Hidden population; social network; nonreciprocal relationship; Markov model.

1. Introduction

Hidden or hard-to-reach populations include several groups of importance to public health research, for example, men who have sex with men (MSM), sex workers (SWs), and injecting drug users (IDUs). A hidden population is typically characterized by i) strong privacy concerns due to illicit or stigmatized behavior, and ii) there is no sampling frame, that is, the size and composition of the population are unknown (Heckathorn 1997). Therefore, it is in general difficult for survey researchers to access hidden populations and draw valid conclusions from sampled data. Several methods have been used to sample from hidden populations, for example, key informant sampling (Deaux and Callaghan 1985), venue-based sampling (Muhib et al. 2001), and snowball sampling (Erickson 1979). However, because of the substantial selection bias inherent in these methods, the

[1] Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden. Email: jensm@math.su.se (corresponding author)
[2] Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan. Email: naoki.masuda@bristol.ac.uk
[3] Department of Engineering Mathematics, University of Bristol, Merchant Venturers Building, Woodland Road, Clifton, Bristol BS8 1UB, United Kingdom. Email: tomb@math.su.se

Acknowledgments: J. M. was supported by Grant No. 2009-5759 from the Swedish Research Council. N. M. was supported by Grants-in-Aid for Scientific Research (No. 23681033) from MEXT, Japan, the Nakajima Foundation, and the Aihara Project, the FIRST program from JSPS, initiated by CSTP, Japan. T. B. is grateful to Riksbankens Jubileumsfond (contract P12-0705:1) for financial support. The authors would like to thank Prof. Fredrik Liljeros and Dr. Xin Lu, Department of Sociology, Stockholm University, for use of the Qruiser dataset.

samples obtained have been considered only for convenience purposes (Magnani et al. 2005). Respondent-driven sampling (RDS) is a more recent sampling methodology for hidden populations (Heckathorn 1997; Salganik and Heckathorn 2004; Volz and Heckathorn 2008). RDS combines an improved link-tracing sampling mechanism, similar to snowball sampling, with a mathematical model that is able to produce asymptotically unbiased estimates of population characteristics given that some assumptions about the sampling procedure are fulfilled. Because of these advantages, RDS has become the primary choice for the study of hidden populations. Some recent examples of RDS studies includes MSM in Panama (Hakre et al. 2014), Dar es Salaam, Tanzania (Bui et al. 2014), SWs in Shiraz, Iran (Kazerooni et al. 2013), Kampala, Uganda (Schwitters et al. 2014), IDUs throughout India (Solomon et al. 2015), methamphetamine users in Cape Town, South Africa (Hobkirk et al. 2015), unauthorized migrant workers in San Diego (Zhang et al. 2014), and low-wage workers in US cities (Bernhardt et al. 2013).

In RDS, the social network of the population is used both in the sampling procedure and for inference. Before we describe RDS in more detail, we will introduce some concepts from social network theory (for a comprehensive reference on social network theory, see Wasserman and Faust 1994). Formally, a *social network* is a (finite) set of actors, for example, individuals, couples, or organizations, that are connected through some type of relation, for example, friendship, kinship, or professional agreements. In graph-theoretical terms, the actors are referred to as *vertices* and their relations as *edges*. The relation between two actors can be reciprocal, that is, the relation is mutual, or it can be nonreciprocal. For example, an individual may choose another individual as a friend. If the other individual in turn chooses the first individual as a friend, the relation is reciprocal, and if that individual does not choose the first one, the relation is nonreciprocal. A reciprocal edge is called an *undirected edge* and a nonreciprocal edge is called a *directed edge*. A network in which the directions of edges are ignored is referred to as an *undirected network*. A network in which the directions of edges are meaningful is referred to as a *directed network*. Note that a directed network may include nonreciprocal and reciprocal edges. In Figure 1, we see three nonreciprocal edges.

One might also consider individual properties of the vertices in the network. The *neighbors* of a vertex are the set of vertices to which it connects by an edge. In Figure 1, the neighbors of vertex *v* includes all vertices except one to the lower left. The *degree* of a vertex refers to the number of neighbors it has in an undirected network. If we ignore the directions of edges in Figure 1, vertex *v* has degree four. If the network is directed, one

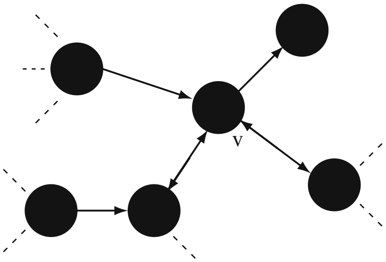


Fig. 1. Sample illustration of a part of a directed social network with six vertices and thirten edges. Vertex v has an undirected degree of two, an incoming degree of one, and an outgoing degree of one.

must consider the directions of edges. A directed edge is either *incoming* to or *outgoing* from a vertex. Because a directed network also may include reciprocal edges, one can identify three types of edges for a vertex w in a directed network, and hence three different degrees: the *undirected degree*, which refers to the number of vertices w connects to by an undirected edge, the *incoming degree*, which refers to the number of vertices w connects to by an incoming edge, and the *outgoing degree*, which refers to the number of vertices w connects to by an outgoing edge. In Figure 1, vertex v has an undirected degree of two, an incoming degree of one, and an outgoing degree of one. The *out-degree* of a vertex is obtained by adding the undirected degree and the outgoing degree. Similarly, the *in-degree* is obtained by adding the undirected degree and the incoming degree. The distribution of vertex degrees in the whole network is called the *degree distribution*. In an undirected network, this distribution is given by the random variable D. In a directed network, the distributions are given by $D^{(\text{un})}$, $D^{(\text{in})}$, and $D^{(\text{out})}$ for the undirected, incoming and outgoing degree, respectively. Formally, these random variables are the degrees of a vertex drawn uniformly at random from the set of all vertices in the network.

An RDS study begins with the selection of a seed group of individuals from the population. The seeds are typically chosen among population members who are well-known to researchers and that supposedly have a large number of contacts. Each seed is provided with a fixed number of coupons, typically between three to five, which are to be distributed among each seed's neighbors. The coupons effectively act as tickets for participation in the study, and each neighbor who has received a coupon is allowed to enter the study upon presenting the coupon at the study site. Those who have received a coupon and joined the study (i.e., respondents) are also provided with coupons to be distributed to their neighbors who have not yet obtained a coupon. This procedure is then repeated until the desired sample size has been reached. The sampling procedure ensures that the identities of participating individuals are not revealed, but because the coupons are numbered, it is possible to obtain the pattern of recruitment throughout the population. Rewards are given to a respondent for his or her participation and for the participation of his or her coupon recipients. This results in social pressure on coupon recipients, which is believed to facilitate effective recruitment. For each respondent, the properties of interest (e.g., HIV status and number of recent sexual encounters) are recorded. Respondents are also asked to provide the number of people they know in the population; this corresponds to the degree in an undirected network and the out-degree in a directed network.

Suppose that we are interested in estimating the proportion of individuals in a population of unknown size N with a specific trait A (e.g., HIV status), denoted p_A. Assume that we have obtained a sample s from an RDS study on this population. In order to estimate p_A from s, we assume that the RDS recruitment process behaves like a random walk on the social network of the population. To this end, it is assumed that (i) respondents recruit peers from their social contacts with equal probability, (ii) each recruitment consists of only one peer, (iii) sampling is done with replacement, (iv) the degree of respondents is reported without error, (v) the social network of the population is undirected, and (vi) the population forms a connected network. Assumption (vi) essentially means that any vertex in the network can be reached from any other vertex in the network, that is, regardless of where the sampling procedure starts, it is possible to sample all members of the population. If the recruitment process has reached equilibrium, we may then estimate p_A by

$$\hat{p}_A^{VH} = \frac{\sum_{i \in s} 1_i(A)/d_i}{\sum_{i \in s} 1/d_i}, \tag{1}$$

where $1_i(A)$ equals one if i has trait A and zero otherwise and d_i is the degree of vertex i (Volz and Heckathorn 2008). In general, when the random walk is in equilibrium and has a known stationary distribution $\{\pi_i;\ i = 1, \ldots, N\}$, we obtain an unequal probability estimator for p_A as

$$\hat{p}_A = \frac{\sum_{i \in s} 1_i(A)/\pi_i}{\sum_{i \in s} 1/\pi_i}. \tag{2}$$

In an undirected network, the stationary distribution is proportional to degree, that is, $\pi_i \propto d_i$ (Doyle and Snell 1984; Lovász 1993). Hence, the estimator in Equation (1) is obtained by using this fact to replace π_i with d_i in Equation (2). Note that the estimators in Eqs. (1) and (2) are the ratio of two Horwitz-Thompson estimators, of the population total and the population size, respectively, from which asymptotically unbiased estimates can be obtained (Särndal et al. 1992, ch. 5.6). This follows because we sample from the random walk model in equilibrium. In practice, Assumptions (i)-(vi) put RDS recruitment in the framework of an irreducible Markov chain for which equilibrium will be approached asymptotically. Although asymptotic equilibrium will not be reached in an RDS study, the recruitment process may come to an approximate equilibrium, and the use of the estimator in Equation (1) can be motivated. Hence, the Markov model obtained from Assumptions (i)-(vi) facilitates the transition from a convenience sample of seeds to a probability sample for which unequal probability estimation procedures can be used.

In most RDS studies, it is not likely that Assumptions (i)-(vi) will hold simultaneously. In this case, the random walk model of the recruitment process will at best be an approximation to the true process and the estimator in Equation (1) may be subject to substantial bias and variance. In recent years, much RDS research has focused on the sensitivity of RDS estimators to violations of Assumptions (i)-(vi). For example, in Gile and Handcock (2010) it was shown that the violation of Assumption (iii) from large sample fractions (>50% of the population) may result in large bias, and in Tomas and Gile (2011) it was shown that bias can be large when Assumption (i) is violated by differential recruitment, that is, the tendency of individuals to preferentially recruit neighbors with certain properties. In Lu et al. (2012), it was found that bias can be substantial if the social network of the population is directed (violation of Assumption (v)), or if recruitment is correlated with study variables (violation of Assumption (i)). Moreover, RDS has been empirically evaluated in, for example, Goel and Salganik (2010), where simulations on empirical networks showed that variance in RDS estimates can be five to ten times larger than in estimates from simple random sampling, and in McCreesh et al. (2012), where it was shown that only 50%–74% of 95% RDS confidence intervals (using bootstrap variance estimates) covered the true population values in an RDS study on a known population of male households in rural Uganda. Several attempts have been made to find new estimators for RDS. In Gile (2011), a successive-sampling estimator that utilizes prior

Table 1. Proportion of directed edges in social networks

Real-life social networks	Prop. dir.	**Online social networks**	Prop. dir.
High-tech managers (Wasserman and Faust 1994)	0.71	Google+ (Oct 2011) (Gong and Xu 2014)	0.62
Dining partners (Moreno 1960)	0.76	Flickr (May 2007) (Gong and Xu 2014)	0.55
Radio amateurs (Killworth and Bernard 1976)	0.59	LiveJournal (Dec 2006) (Mislove et al. 2007)	0.26
Dutch college students (Van De Bunt, Van Duijn, and Snijders 1999)	0.19	Twitter (June 2009) (Kwak et al. 2010)	0.78
Campus hall residents (Freeman, Webster, and Kirke 1998)	0.38	University e-mail (Newman, Forrest, and Balthrop 2002)	0.77
Jazz musicians (Gleiser and Danon 2003)	0.52	Enron e-mail (Boldi and Vigna 2004) (Boldi et al. 2011)	0.85

information on the population size is derived and in Gile and Handcock (2015), an estimator utilizing a superpopulation model for the social network is presented. In Lu et al. (2013), an estimator for RDS on directed social networks utilizing prior information on the in-degrees of groups of population members is presented. Lu (2013) gives an estimator that uses additional information on the composition of sampled individuals' contacts.

Current RDS estimation procedures (except Lu et al. 2013) assume that the social network of the population is undirected (cf. Assumption (v)). However, real social networks are directed in general and often include a considerable number of nonreciprocal edges. Examples of real-life social networks and social networks from online communities, including e-mail social networks, and their fraction of nonreciprocal edges among the total number of edges are shown in Table 1. In real-life social networks, such as those listed in Table 1, network data are often gathered by asking individuals to list, for example, all or some of their friends (Marsden 1990). Then, if an individual i lists j as his or her friend, but j does not list i, there will be a directed edge from i to j. For example, in the network of Dutch college students in Table 1, students were asked to list all their friends among the other residents (Van De Bunt et al. 1999), and in the dining partners' network, individuals were asked to name their two most preferred choices of dining partners. In online social networks, directed edges typically occur because an individual can add another member of the social network to his or her friend list without that member adding him or her, and in the e-mail networks, edges are formed from an individual to another if the latter is present in the former's address book (Newman et al. 2002) or if a message has been sent from the former to the latter (Boldi and Vigna 2004; Boldi et al. 2011).

The presence of directed edges may induce substantial bias and variance in the estimator in Equation (1) and other RDS estimators. For example, in their evaluation of RDS by simulations on an empirical network, Lu et al. (2012) found that the presence of directed edges caused bias as high as 0.06 in estimates from Equation (1); this can be compared with the bias of less than 0.01 induced by violation of the sampling with replacement assumption in the same study. In Lu et al. (2013), simulations on generated networks for which the proportion of directed edges was controlled showed that even a small proportion of directed edges can introduce bias in the estimator in Equation (1) and that the bias can be large (≈ 0.075) when the proportion of directed edges increases. There is also evidence of recruitment taking place by nonreciprocal relations in empirical RDS studies. For example, in an RDS study of IDUs in Sydney, Australia, 29% of participants described their relationship with their recruiter as “Not very close” (Paquette et al. 2011), and in an RDS study of IDUs in Tijuana, Mexico, 62% characterized their relationship to their recruiter as “friend” (Abramovitz et al. 2009). In an RDS study of MSM in Chicago, 13% said that they were “Not at all close” to their recruiter, and 17% characterized the relationship as “other” (instead of friend/acquaintance/partner/relative/coworker) (Phillips et al. 2014), and in an RDS study of an aboriginal community in Labrador, Canada, 80% of those recruited indicated that their recruiter was a “close relative”, “distant relative”, “close friend”, or “friend” (Dombrowski et al. 2013).

The purpose of this article is to develop an estimator for p_A that does not require prior information on population properties for RDS in populations with directed social networks. To estimate p_A without bias from an RDS sample in such cases, we need to accurately calculate Equation (2). Because the RDS estimation method assumes a random walk behavior of the recruitment process, a random walk framework for directed networks is a key component of this expansion. This is no trivial task, because the random walk behaves very differently in undirected and directed networks. In particular, the stationary distribution of the random walk is simply proportional to the degree of the vertex in undirected networks, whereas it is affected by the entire network structure in directed networks (Donato et al. 2004; Langville and Meyer 2006; Masuda and Ohtsuki 2009). We aim to develop such a framework through which we can find estimators for the stationary distribution $\{\pi_i\}$ of the random walk on a directed network to be used in Equation (2) to estimate p_A. We will do this in several steps. Initially, we assume that we observe both the undirected degree, the incoming degree, and the outgoing degree of all vertices that are sampled. We consider the probability of returning to the same vertex after two steps in the random walk and use renewal theory to find an estimator for $\{\pi_i\}$. Then, we consider this estimation procedure in the more realistic situation when we only observe the out-degrees of sampled individuals. First, we derive results for the situation in which the expectations of the degree distributions are known. Then, we drop this assumption and by assuming a model for the social network of the population, we can estimate the unknown expectations. This gives our final estimator. All estimators are then evaluated and compared to existing RDS estimators by means of simulations.

2. Random Walks on Directed Networks

We consider a directed, strongly connected network G with N vertices. The assumption that the network is strongly connected is the equivalent of Assumption (vi) for directed networks, and means that it is possible to go from any vertex v to any other vertex w and then back (Newman 2010). Let $e_{ij} = 1$ if there is a directed edge from i to j and zero otherwise. An undirected edge exists between i and j if and only if $e_{ij} = e_{ji} = 1$. We denote the number of undirected, incoming, and outgoing edges at vertex i by $d_i^{(\mathrm{un})}$, $d_i^{(\mathrm{in})}$, and $d_i^{(\mathrm{out})}$, respectively. The degree distributions are given by the corresponding random variables $D^{(\mathrm{un})}$, $D^{(\mathrm{in})}$, and $D^{(\mathrm{out})}$. For an undirected network, we obtain $d_i^{(\mathrm{in})} = d_i^{(\mathrm{out})} = 0$, and refer to the degree of vertex i as $d_i = d_i^{(\mathrm{un})}$. Otherwise, the degree of vertex i refers to the triplet $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$. Then, $d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}$ and $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ is the in-degree and out-degree of vertex i, respectively. In this notation, vertex v in Figure 1 has $d_v^{(\mathrm{un})} = 2$, $d_v^{(\mathrm{in})} = 1$, and $d_v^{(\mathrm{out})} = 1$. If the network in Figure 1 was undirected, we would obtain $d_v = 4$. It should be noted that, during the random walk, we may observe for example the out-degree $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$, but not the $d_i^{(\mathrm{un})}$ and $d_i^{(\mathrm{out})}$ values separately.

Consider the simple random walk $X = \{X(t); t = 0, 1, \ldots\}$ with state space $S = \{1, \ldots, N\}$ on G such that the walker staying at vertex i moves to any of the $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ neighbors reached by an undirected or outgoing edge with equal probability. We denote the stationary distribution of X by $\{\pi_i; i = 1, \ldots, N\}$, where $\pi_i = \lim_{t\to\infty} P(X(t) = i)$. The stationary distribution exists if the network is aperiodic, that is, the walker will not return periodically to the same vertex repeatedly during the walk. If we sample from the random walk in equilibrium, we refer to $\{\pi_i\}$ as the *selection probabilities* of the vertices in G.

For an arbitrary network, we obtain

$$\pi_i = \sum_{j=1}^{N} \frac{e_{ji}}{\sum_{\ell=1}^{N} e_{j\ell}} \pi_j = \sum_{j=1}^{N} \frac{e_{ji}}{d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}} \pi_j, i = 1, \ldots, N, \tag{3}$$

where the stationary distribution is fully defined by $\sum_{i=1}^{N} \pi_i = 1$. In undirected networks, we obtain $\pi_i = d_i / \sum_{j=1}^{N} d_j$. In contrast, there is no analytical closed-form solution for $\{\pi_i\}$ in directed networks. If a directed network has little assortativity (i.e., degree correlation between adjacent vertices), $\{\pi_i\}$ is often accurately estimated by the normalized in-degree (Fortunato et al. 2008; Ghoshal and Barabási 2011) because

$$\pi_i \approx \sum_{j=1}^{N} \frac{e_{ji}}{d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}} \bar{\pi} \propto \sum_{j=1}^{N} e_{ji} = d_i^{(\mathrm{in})} + d_i^{(\mathrm{un})}, \tag{4}$$

where $\bar{\pi}$ is the average selection probability. Equation (4) depends only on the in-degree of vertices, that is, it provides a local description of the global solution to Equation (3). However, the estimate given by (4) is often inaccurate in general directed networks (Donato et al. 2004; Masuda and Ohtsuki 2009). Moreover, since it is much easier for individuals to assess their out-degree, that is, how many people they know, than their in-degree, that is, by how many people they are known, it is common to observe only the out-degree. In this case, Equation (4) cannot be used with an RDS sample.

3. Estimating Selection Probabilities

We now derive estimators of the selection probabilities for the random walk on directed networks. We first derive an estimation scheme when the full degree $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$ is observed for all the vertices i visited by the random walk. Then, we restrict this estimation to the situation in which only the out-degree $d_i^{\mathrm{un}} + d_i^{\mathrm{out}}$ of the visited vertices is observed. In both situations it is assumed that the degrees are observed without error. Note that the random walk allows a vertex to be visited multiple times, whereas it is typically not allowed to be sampled several times in an RDS study.

3.1. Estimating Selection Probabilities from Full Degrees

In order to estimate $\{\pi_i\}$, we assume that $X(t_0) = i$, where t_0 is sufficiently large for the stationary distribution to be reached. We evaluate the frequency with which $X(t)$ visits i in the subsequent times. If $X(t)$ leaves i through an undirected edge $e_{i\cdot}^{(\mathrm{un})}$, where $e_{i\cdot}^{(\mathrm{un})}$ is one of the $d_i^{(\mathrm{un})}$ undirected edges owned by i, $X(t)$ may return to i after two steps using the same edge and repeat the same type of returns m times in total, perhaps using different undirected edges $e_{i\cdot}^{(\mathrm{un})}$. Then, $X(t_0) = X(t_0 + 2) = \cdots = X(t_0 + 2m) = i$ and $X(t_0 + 2m + 2) = k$ for some $k \neq i$.

If $X(t_0 + 2) = i$, the walk first moves from i through an undirected edge to vertex j at $t = t_0 + 1$ and returns to i through the same edge at $t = t_0 + 2$. The probability of this event is given by $d_i^{(\mathrm{un})}/\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right) \cdot 1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$. Because the out-degree of vertex j, that is, $d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}$, is unknown, we approximate $1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$ by $E(1/(\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}))$. Here $\tilde{D}^{(\mathrm{un})}$ denotes the undirected degree distribution under the condition that the vertex is reached by following an undirected edge. This yields a *size-biased* distribution for the undirected degree, given by $P(\tilde{D}^{(\mathrm{un})} = d) \propto dP(D^{(\mathrm{un})} = d)$ (Newman 2010). It is also possible to estimate $1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$ by $1/E(\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})})$, which however proved to have very little effect in our simulations, and if any, a slightly worse one. Thus, we estimate the probability of returning to vertex i after two steps by

$$p_i^{(\mathrm{ret})} = \frac{d_i^{(\mathrm{un})}}{d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}} E\left(\frac{1}{\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}}\right). \tag{5}$$

When $t \geq t_0 + 2m + 3$, we use Equation (4) to estimate the probability of visiting vertex i at any time as being proportional to $d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}$, that is,

$$p_i^{(\mathrm{vis})} = \frac{d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}}{N(E(D^{(\mathrm{un})}) + E(D^{(\mathrm{in})}))}. \tag{6}$$

Under these estimates, the number of returns after two steps to vertex i, counting the starting point $X(t_0) = i$ as the first return to i, is geometrically distributed with expected value $1/\left(1 - p_i^{(\mathrm{ret})}\right)$, and the number of steps starting from $t = t_0 + 2m + 2$, including this step, and ending at the time immediately before visiting i with probability $p_i^{(\mathrm{vis})}$ is geometrically distributed with the expected value $1/p_i^{(\mathrm{vis})}$.

We then have a renewal process, that is, a process which repeatedly regenerates at random times such that the intervals between them are of independent and identically distributed lengths. These random times are called renewals. We denote our process $\left\{R_i^n; n \geq 1, R_i^0 = 0\right\}$, with the nth renewal occurring at the random time $R_i^n = \sum_{k=1}^{n}\left(2Z_i^k + Y_i^k\right)$, where $Z_i^k \sim Ge\left(1 - p_i^{(\mathrm{ret})}\right)$ and $Y_i^k \sim Ge\left(p_i^{(\mathrm{vis})}\right)$. In Figure 2, the behavior of the process during the kth renewal period is shown schematically. Figure 2(a) shows the behavior of the walk when it makes consecutive returns to i. During this time, the walker always leaves i through an undirected edge, which is not necessarily the same edge each time (left part of Figure 2(a)), and returns after two time steps to i via the same edge (right part of Figure 2(a)). This is repeated such that the walker makes in total Z_i^k consecutive returns to i. The duration of this is $2Z_i^k$ time steps. Figure 2(b) shows the behavior of the walk when it leaves i and does not return after two time steps. This occurs

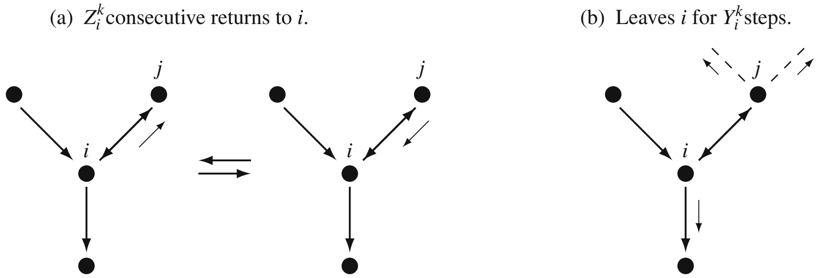

(a) Z_i^kconsecutive returns to i.
(b) Leaves i for Y_i^ksteps.
j
i
j
i
j
i

Fig. 2. Schematic illustration of the kth renewal period.

if either the walker leaves i by an outgoing edge, through which it is impossible to return directly to i, or if the walker leaves i by an undirected edge but does not return to i through this edge in the next time step. When the walker has left i, the time until its return to i is Y_i^k time steps. The average time step between consecutive renewal events is equal to $2E\left(Z_i^k\right) + E\left(Y_i^k\right)$. The average number of visits to i between two renewal events, with the visit to i at $t = t_0$ included, is equal to $E\left(Z_i^k\right)$. Therefore, from renewal theory (see e.g., Feller 1950), we obtain an estimate of π_i as

$$\pi_i \approx \frac{E\left(Z_i^k\right)}{2E\left(Z_i^k\right) + E\left(Y_i^k\right)} = \frac{\dfrac{1}{1 - p_i^{(\mathrm{ret})}}}{2\dfrac{1}{1 - p_i^{(\mathrm{ret})}} + \dfrac{1}{p_i^{(\mathrm{vis})}}} = \frac{p_i^{(\mathrm{vis})}}{2p_i^{(\mathrm{vis})} + 1 - p_i^{(\mathrm{ret})}}. \tag{7}$$

Because $p_i^{(\mathrm{ret})} = O(1)$ and $p_i^{(\mathrm{vis})} = O(1/N)$, removing higher-order terms in Equation (7) yields

$$\hat{\pi}_i \approx \frac{p_i^{(\mathrm{vis})}}{1 - p_i^{(\mathrm{ret})}} \propto \frac{d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}}{1 - \dfrac{d_i^{(\mathrm{un})}}{d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}} E\left(\dfrac{1}{\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}}\right)}. \tag{8}$$

The proportionality constant is given by imposing that $\sum_{i=1}^{N} \hat{\pi}_i = 1$. If the network is undirected, we obtain $\hat{\pi}_i \propto d_i^{(\mathrm{un})}$, such that $\hat{\pi}_i$ coincides with the exact solution used in Equation (1). If the network is without reciprocal edges, the estimator is proportional to incoming degree $d_i^{(\mathrm{in})}$.

3.2. Estimating Selection Probabilities from Out-Degrees

A common situation in RDS is that only the out-degrees (i.e., $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$) of respondents are observed. Then, the estimator of the selection probabilities given by Equation (8) cannot be used directly. To cope with this situation, we estimate the number of undirected, incoming, and outgoing edges from the observed out-degrees and substitute the estimated $\left(\hat{d}_i^{(\mathrm{un})}, \hat{d}_i^{(\mathrm{in})}, \hat{d}_i^{(\mathrm{out})}\right)$ in Equation (8).

Assume that we have observed the out-degree $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ of vertex i. We estimate $d_i^{(\mathrm{un})}$ and $d_i^{(\mathrm{out})}$ by their expected proportions of the observed out-degree, and the incoming degree by its expectation, as follows:

$$\begin{cases} \hat{d}_i^{(\mathrm{un})} = \dfrac{E(D^{(\mathrm{un})})}{E(D^{(\mathrm{un})}) + E(D^{(\mathrm{out})})} \left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right), \\ \hat{d}_i^{(\mathrm{out})} = \dfrac{E(D^{(\mathrm{out})})}{E(D^{(\mathrm{un})}) + E(D^{(\mathrm{out})})} \left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right), \\ \hat{d}_i^{(\mathrm{in})} = E(D^{(\mathrm{in})}). \end{cases} \tag{9}$$

The expectations used in Equation (9) rely on the assumption that we have a random sample from the network, which is not true in this case. We have evaluated the case of a size-biased distribution for incoming and/or undirected degrees; however, our numerical results suggest that this makes little difference, and if any, increases the bias of selection probability estimators. Therefore, we stay with the estimators given by Equation (9).

When $\left(\hat{d}_i^{(\mathrm{un})}, \hat{d}_i^{(\mathrm{in})}, \hat{d}_i^{(\mathrm{out})}\right)$ is substituted in Equation (8) in place of $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$, the term $\hat{d}_i^{(\mathrm{un})} / \left(\hat{d}_i^{(\mathrm{un})} + \hat{d}_i^{(\mathrm{out})}\right)$ in the denominator is constant. Therefore, the estimator is proportional to $\hat{d}_i^{(\mathrm{un})} + \hat{d}_i^{(\mathrm{in})}$ and hence equivalent to Equation (4) calculated with the estimated degrees.

3.3. Estimating Expectations of Degree Distributions

The degree estimators in Equation (9) rely on $E(D^{(\mathrm{un})})$, $E(D^{(\mathrm{in})})$, and $E(D^{(\mathrm{out})})$, which are not estimable from a typical RDS sample, where only the out-degrees $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ of respondents are observed. In order to extend the estimation procedure to handle these unknown expectations, we assume a model for the network by which they can be estimated.

Specifically, it is assumed that the observed network is a realization of a directed equivalent of the simple $G(N, p = \lambda/(N-1))$ random graph (Erdős and Rényi 1960). This graph has N vertices and hence $\binom{N}{2}$ pairs of vertices. Given parameters $\alpha \in [0, 1]$ and $\lambda \in [0, N-1]$, each pair of vertices independently forms an edge with probability $\lambda/(N-1)$, which is undirected with probability $(1-\alpha)$ and directed with probability α. When the edge is directed, the direction is selected with equal probability. Because each vertex may connect to each of the other $N-1$ vertices, it follows that λ is the expected total degree of a vertex. We also have that α is the fraction of directed edges as $N \rightarrow \infty$.

Because edges are formed independently of each other, vertex degrees are binomially distributed. Hence, if N is large, $D^{(\mathrm{un})}$, $D^{(\mathrm{in})}$, and $D^{(\mathrm{out})}$ approximately follow independent Poisson distributions with parameters $(1-\alpha)\lambda$, $\alpha\lambda/2$, and $\alpha\lambda/2$, respectively. It follows that the out-degree $D^{(\mathrm{un})} + D^{(\mathrm{out})}$ and the in-degree $D^{(\mathrm{un})} + D^{(\mathrm{in})}$ are both Poisson distributed with parameter $(2-\alpha)\lambda/2$. Consequently, to estimate the unknown expectations, it is enough to estimate α and λ, and substitute the estimated $\hat{\alpha}$ and $\hat{\lambda}$ in the expectations of the (Poissonian) degree distributions.

To find an estimator of α, we again consider the random walk $X = \{X(t)\}$ on the network. Assume that $e_{ij} = 1$, $X(t_0) = i$, and $X(t_0 + 1) = j$, for a large t_0. If $X(t_0 + 2) = i$, the edge between i and j is undirected, that is, $e_{ij} = e_{ji} = 1$, and the random walk leaves vertex j via e_{ji}. The probability that the edge is undirected is set to $(1-\alpha)/(1-\alpha/2)$, that is, the probability that an edge selected uniformly at random among all undirected and incoming edges is undirected. This will only approximately hold for the random walk, but simulations show that it is a reasonable approximation. If there is an undirected edge between i and j (i.e., $e_{ji} = 1$), the random walk leaves j via e_{ji} with probability $1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$. Thus, the random walk revisits vertex i at $t_0 + 2$ under the directed E-R random-graph model with probability

$$p_j^{(\mathrm{rev})} = \frac{1-\alpha}{1-\alpha/2} \cdot \frac{1}{d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}}. \tag{10}$$

Let M be the number of revisits, as described above, during l consecutive steps, where l is typically equal to the sample size. Then, we have $M = \sum_{k=2}^{l} M_k$, where $M_k = 1$ if a revisit occurs in step k and $M_k = 0$ otherwise. M_k is Bernoulli distributed, $M_k \sim \mathrm{Be}\left(p_{j_{k-1}}^{(\mathrm{rev})}\right)$, where j_{k-1} is the vertex visited in step $k-1$. We obtain the expected number of revisits as

$$E(M) = \frac{1-\alpha}{1-\alpha/2} \sum_{k=1}^{l-1} \frac{1}{d_{j_k}^{(\mathrm{un})} + d_{j_k}^{(\mathrm{out})}}. \tag{11}$$

If m is the observed number of revisits, we set $m = E(M)$ in Equation (11) to obtain the moment estimator

$$\hat{\alpha} = \frac{m - \sum_{k=1}^{l-1}\left(d_{j_k}^{(\mathrm{un})} + d_{j_k}^{(\mathrm{out})}\right)^{-1}}{m/2 - \sum_{k=1}^{l-1}\left(d_{j_k}^{(\mathrm{un})} + d_{j_k}^{(\mathrm{out})}\right)^{-1}}. \tag{12}$$

If the estimated $\hat{\alpha} < 0$, we force $\hat{\alpha} = 0$.

Given $\hat{\alpha}$, we estimate λ as follows. If $\alpha = 0$, the network contains only undirected edges, and the observed out-degree equals the observed undirected degree, which has a size-biased distribution with $E(\tilde{D}^{(\mathrm{un})}) = \lambda + 1$. If $\alpha = 1$, the network has only directed edges, and the expected observed out-degree is well approximated by the expected number of outgoing edges, $\lambda/2$. By linearly interpolating the expected observed out-degree between $\alpha = 0$ and $\alpha = 1$, and substituting it with the mean sample out-degree $\bar{u}$, we obtain $\bar{u} = \lambda/2 + (1 - \alpha)(1 + \lambda/2)$, which yields an estimator of λ as

$$\hat{\lambda} = \frac{\bar{u} + \hat{\alpha} - 1}{1 - \hat{\alpha}/2}. \tag{13}$$

Using $\hat{\alpha}$ and $\hat{\lambda}$, we can estimate the expectations of the degree distributions under the random-graph model. For example, $E(D^{(\mathrm{un})})$ is estimated by $(1 - \hat{\alpha})\hat{\lambda}$. By substituting these estimated expectations in Eqs. (8) and (9), we obtain an estimator of the selection probability of vertex i as

$$\hat{\pi}_i \propto \hat{d}_i^{(\mathrm{un})} + \hat{d}_i^{(\mathrm{in})} = \frac{1-\hat{\alpha}}{1-\hat{\alpha}/2}\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right) + \frac{\hat{\alpha}\hat{\lambda}}{2}. \tag{14}$$

When $\alpha = 0$ is assumed known and used in place of $\hat{\alpha}$, the estimator in Equation (14) is equivalent to that used in Equation (1). When $\hat{\alpha} = \alpha = 1$, the estimator is proportional to one, and thus equivalent to the sample mean. This reflects the fact that, if $\alpha = 1$, the network has no undirected edges, and the out-degree is equal to the outgoing degree, which does not provide any information on the selection probability of a vertex in this case.

It should be noted that the construction of the directed Erdős-Rényi graphs results in vertices having the same out-degree and in-degree on average, which is not likely to occur in actual populations where RDS is used. This makes estimation of in-degree using only the observed out-degree feasible, and might possibly favor the performance of the estimator in Equation (14) for networks generated by this model.

4. Simulation Setup

We numerically examine the accuracy of our estimation schemes on directed Erdős-Rényi graphs, a model of directed power-law networks (i.e., networks with a power-law degree distribution), and an online MSM social network. We evaluate both the estimated selection probabilities and corresponding estimates of p_A. As described in Section 1, real-life directed social networks show a varying fraction of directed edges, corresponding to a diversity of α values. Therefore, α is varied in the model networks. We also vary λ and other network parameters. We study the performance of our estimators when the full degree is observed and when only the out-degree is observed, and compare them with existing estimators. We do not consider RDS estimators that are not based on the random walk framework because they fall outside the scope of this study.

4.1. *Network Models and Empirical Network*

The first model network that we use is a variant of the Erdős-Rényi graph with a mixture of undirected and directed edges, as described in Section 3. We generate the networks with $\alpha \in \{0.25, 0.5, 0.75\}$ and $\lambda \in \{5, 10, 15\}$. We then extract the largest strongly connected component of the generated network, which has $O(N)$ vertices for all combinations of α and λ.

The directed Erdős-Rényi networks have Poisson degree distributions with quickly decaying tails. In fact, many empirical networks have heavy-tailed degree distributions as represented by the power law (Newman 2010). In other words, there are typically small numbers of vertices whose degrees are huge, and a majority of vertices have small degrees. To mimic heavy-tailed degree distributions, we also use a variant of a power-law network model (Goh et al. 2001; Chung and Lu 2002; Chung et al. 2003). The original algorithm for generating undirected power-law networks presented in Goh et al. (2001) is as follows.

We fix the number of vertices N and expected degree $E(D)$. Then, we set the weight of vertex i $(1 \leq i \leq N)$ to be $w_i = i^{-\tau}$. As specified in the following, w_i represents the extent to which vertex i attracts edges; a large w_i value yields a large degree. Parameter $0 \leq \tau \leq 1$ controls the power-law exponent of the degree distribution. If $\tau = 0$, all w_i are equal such that each vertex is statistically the same. In this case, the degree distribution produced by the following algorithm will not be heavy-tailed. When $\tau > 0$, a vertex with small i possesses large w_i and will in fact have a very large degree. Then, we select a pair of vertices i and j $(1 \leq i \neq j \leq N)$ with probability proportional to $w_i w_j$. If the two vertices are not yet connected, we connect them by an undirected edge. We repeat the procedure until the network has $E(D)N/2$ edges such that the expected degree is equal to $E(D)$. The expected degree of vertex i is proportional to w_i, and the degree distribution is given by $p(d) \propto d^{-\gamma}$, where $\gamma = 1 + \frac{1}{\tau}$ (Goh et al. 2001).

To generate a power-law network in which undirected and directed edges are mixed with a desired fraction, we extend the algorithm as follows. First, we specify the expected undirected degree $E(D^{(\mathrm{un})})$ and generate an undirected network. Second, we define $w_i^{\mathrm{in}} = (\sigma^{\mathrm{in}}(i))^{-\tau^{\mathrm{in}}}$ $(1 \le i \le N)$, where σ^{in} is a realization of the random permutation on $1, \ldots, N$. Parameter τ^{in} specifies the power-law exponent of the incoming degree distribution. Similar to the undirected case, a vertex with a small $\sigma^{\mathrm{in}}(i)$ value will have a large in-degree. Similarly, we set $w_i^{\mathrm{out}} = (\sigma^{\mathrm{out}}(i))^{-\tau^{\mathrm{out}}}$ $(1 \le i \le N)$, where σ^{out} is another realization of the random permutation on $1, \ldots, N$. Third, we select a pair of vertices with probability proportional to $w_i^{\mathrm{in}} w_j^{\mathrm{out}}$. If $i \neq j$ and there is no directed edge from j to i yet, we place a directed edge from j to i. We repeat the procedure until a total of $E(D^{(in)})N/2$ edges are placed. It should be noted that $E(D^{(\mathrm{in})}) = E(D^{(\mathrm{out})})$. The incoming degree distribution is given by $p(d^{\mathrm{in}}) \propto (d^{\mathrm{in}})^{-\gamma^{\mathrm{in}}}$, where $\gamma^{\mathrm{in}} = 1 + \frac{1}{\tau^{\mathrm{in}}}$, and similar for the outgoing degree distribution. Finally, we superpose the obtained undirected network and directed network to make a single graph. If the combined graph is not strongly connected, we discard it and start over until a strongly connected network is generated. By construction, a network constructed from this model is devoid of degree correlation.

In both network models, we vary the probability of a vertex being assigned property A as proportional to six different combinations of its degree: in-degree, out-degree, undirected degree, incoming degree, outgoing degree, and directed degree, that is, the sum of incoming and outgoing degree. Formally, if $P(\text{vertex } i \text{ has } A) \propto g\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$, we let g be equal to $\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}\right)$, $\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right)$, $d_i^{(\mathrm{un})}$, $d_i^{(\mathrm{in})}$, $d_i^{(\mathrm{out})}$, and $\left(d_i^{(\mathrm{in})} + d_i^{(\mathrm{out})}\right)$, respectively. We refer to these as different ways to allocate property A. We also examined the case in which we assigned the property uniformly over all vertices. However, because the performance of the different estimators is similar in this case, we do not show the results in the following. For all allocations of A, the property is assigned in such a way that the expected proportion of vertices being assigned A is equal to some fixed value p. Because A is stochastically assigned, the actual proportion p_A of vertices with A will vary between realized allocations.

We also evaluate our estimators on an online MSM social network, extracted during Dec 2005-Jan 2006 from www.qruiser.com, which is the Nordic region's largest community for lesbian, gay, bisexual, transgender, and queer persons (Rybski et al. 2009). In this network, an edge represents that at least one message has been sent between the two vertices connected by that edge. A directed edge occurs if messages have only been sent in one direction between two vertices. The data set considered here was first described in Lu et al. (2012) and represents a subpopulation of the original data set consisting of 16,082 male homosexual members in a directed social network that is made up of one strongly connected component. This network represents the social structure of a hidden population and makes it possible to evaluate the effect of the presence of nonreciprocal edges in RDS. It has previously been used to evaluate the performance of RDS estimators under different violations of Assumptions (i)-(vi) in Lu et al. (2012) and in directed social networks in Lu et al. (2013). The data set also includes users' profiles, which are seldom available. From these, we obtain four dichotomous individual properties: age (born before 1980 or not), county (live in Stockholm or not), civil status (married or unmarried), and profession (employed or unemployed). This makes it possible to evaluate the performance of RDS estimators of population proportions on this network. The fraction of directed edges in the network is equal to $\alpha = 0.76$. The in-degree and out-degree distributions are skewed, and the mean number of edges λ is equal to 27.74 (Lu et al. 2012). Preferably, RDS would be evaluated on a network which is known to depict that on which the recruitment process in RDS takes place. Such network data is rare, however, and in its absence, the considered network is a good option for RDS evaluation.

4.2. *Evaluation of Estimators*

We compared the performance of our estimators of the selection probabilities with three other estimators. We refer to our estimator $\{\hat{\pi}_i\}$ obtained from Equation (8) as $\left\{\hat{\pi}_i^{(\mathrm{ren})}\right\}$, where *ren* stands for renewal. This estimator is compared to $\left\{\hat{\pi}_i^{(\mathrm{uni})}\right\}$, which assigns a uniform probability $\hat{\pi}_i^{(\mathrm{uni})} = 1/N$ for all i, $\left\{\hat{\pi}_i^{(\mathrm{outdeg})}\right\}$, for which $\hat{\pi}_i^{(\mathrm{outdeg})} \propto d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$, that is, proportional to out-degree (Equation 1), and $\left\{\hat{\pi}_i^{(\mathrm{indeg})}\right\}$, where $\hat{\pi}_i^{(\mathrm{indeg})}$ $\propto d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}$, that is, proportional to in-degree (Equation 4). Note that if the network is undirected, $\left\{\hat{\pi}_i^{(\mathrm{outdeg})}\right\}$ and $\left\{\hat{\pi}_i^{(\mathrm{indeg})}\right\}$ are equal. However, typically in RDS the out-degree is observed and $\left\{\hat{\pi}_i^{(\mathrm{outdeg})}\right\}$, which is used in the current RDS estimator in Equation (1), is the estimator that should be considered in the undirected case. In the following, we suppress the $\{\}$ notation.

To assess the performance of an estimator we calculated its estimated selection probabilities $\hat{\pi}_i$ and the true stationary distribution π_i for all the vertices in the given network. Then, we calculated the *total variation distance* defined by

$$D_{TV} = \frac{1}{2}\sum_{i=1}^{N} |\hat{\pi}_i - \pi_i| \tag{15}$$

(Levin et al. 2009). The stationary distribution π_i was obtained using the power method, which is an iterative method that works as follows (Langville and Meyer 2006). Starting from an arbitrary nonzero vector of size N, in each iteration the resulting vector is multiplied with the matrix $\{e_{ij}\}$, where $e_{ij} = 1$ if there is a directed edge from i to j. Then, under some conditions that hold for the networks used in this study, the resulting vector converges to the stationary distribution, which is the eigenvector corresponding to the largest eigenvalue of $\{e_{ij}\}$. We use an accuracy of 10^{-10} in terms of the total variation distance for the two distributions given in the successive two steps of the power iteration.

For $\hat{\pi}^{(\mathrm{ren})}$, we considered three variants depending on the information available from observed degree and knowledge of the expectations of the degree distributions. When the full degree $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$ was observed, we used Equation (8) to calculate $\hat{\pi}^{(\mathrm{ren})}$, where $E(1/(\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}))$ is estimated by the mean of the inverse sample out-degrees. We denote the corresponding estimator with $\hat{\pi}_{\mathrm{f.d.}}^{(\mathrm{ren})}$, where *f.d.* stands for "full degree". When only the out-degree was observed and the expectations of the degree distributions were known, we used Equation (9). This case is only evaluated for the directed Erdős-Rényi graphs, and the corresponding estimator is denoted by $\hat{\pi}_{\alpha,\lambda}^{(\mathrm{ren})}$. If only the out-degree was observed and the expectations of the degree distributions were unknown, we used Equations (12), (13), and (14), and the estimator is denoted $\hat{\pi}^{(\mathrm{ren})}$.

We obtained a sample of size n_s from each generated network by means of a random walk starting from a randomly selected vertex. In the random walk, we collected the degree of visited vertices and observed whether they had property A or not. We estimated the population proportion p_A from the sample by replacing π in Equation (2) by either $\hat{\pi}^{(\text{uni})}$, $\hat{\pi}^{(\text{outdeg})}$, $\hat{\pi}^{(\text{indeg})}$, or any of the variants of $\hat{\pi}^{(\text{ren})}$, yielding estimates $\hat{p}_A^{(\text{uni})}$, $\hat{p}_A^{(\text{outdeg})}$, $\hat{p}_A^{(\text{indeg})}$, or $\hat{p}_A^{(\text{ren})}$, respectively. Note that $\hat{p}_A^{(\text{uni})}$ yields the sample proportion suggested as an estimator for RDS in Heckathorn (1997), $\hat{p}_A^{(\text{outdeg})}$ yields the RDS estimator from Volz and Heckathorn (2008), where the direction of edges is ignored, and $\hat{p}_A^{(\text{indeg})}$ gives the RDS estimator for directed networks from Lu et al. (2013).

5. Numerical Results

5.1. *Directed Erdős-Rényi Graphs*

In Table 2, we show the mean of the total variation distance D_{TV} between the true stationary distribution and $\hat{\pi}^{(\text{uni})}$, $\hat{\pi}^{(\text{outdeg})}$, $\hat{\pi}^{(\text{indeg})}$, and $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$, calculated on the basis of 1,000 realizations of the largest strongly connected component of the directed random graph having $N = 1{,}000$ vertices. Because the standard deviation of D_{TV} is similar

Table 2. Mean and average standard deviation (s.d.) of D_{TV} for the directed random graph when $\left(d_i^{(un)}, d_i^{(in)}, d_i^{(out)}\right)$ is observed and moments of the degree distributions are known. The lowest D_{TV} value is marked in boldface. We set $N = 1{,}000$

(a) $\alpha = 0.1$

λ	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
5	0.185	0.074	0.042	**0.041**	0.004
10	0.131	0.045	0.017	**0.016**	0.002
15	0.106	0.036	**0.010**	**0.010**	0.001

(b) $\alpha = 0.25$

	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
	0.203	0.134	0.077	**0.075**	0.005
	0.140	0.081	0.031	**0.030**	0.002
	0.112	0.063	**0.019**	**0.019**	0.002

(c) $\alpha = 0.5$

λ	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
5	0.247	0.225	0.138	**0.133**	0.009
10	0.160	0.136	0.056	**0.055**	0.004
15	0.126	0.105	0.034	**0.033**	0.002

(d) $\alpha = 0.75$

	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
	0.303	0.319	0.207	**0.201**	0.014
	0.188	0.201	0.090	**0.088**	0.005
	0.144	0.156	**0.055**	**0.055**	0.003

between the estimators, we show an average over the four estimators. The sample size n_s used in $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$ is 500. We also tried $n_s = 200$, which gave similar results. The D_{TV} value of $\hat{\pi}^{(\mathrm{indeg})}$ and $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$ is much smaller than that of $\hat{\pi}^{(\mathrm{uni})}$ and $\hat{\pi}^{(\mathrm{outdeg})}$ for all values of α and λ. Furthermore, $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d}}$ always gives smaller D_{TV} than $\pi^{(\mathrm{indeg})}$ although the two values are similar for many combinations of the parameters.

In Table 3, we show the mean and average s.d. of D_{TV} when the out-degree, that is, $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$, is observed but the individual $d_i^{(\mathrm{un})}$ and $d_i^{(\mathrm{out})}$ values are not. The assumptions underlying the network generation are the same as those for Table 2, and $n_s = 500$. We consider two cases. In the first case, the expectations of the degree distributions are known, and we use the estimator $\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$. In the second case, they are not known, and we use $\hat{\pi}^{(\mathrm{ren})}$. Results for $\hat{\pi}^{(\mathrm{indeg})}$ are not shown in Table 3 because in-degree is not observed. Table 3 indicates that D_{TV} for $\hat{\pi}^{(\mathrm{ren})}$ is smaller than that for $\hat{\pi}^{(\mathrm{uni})}$ and $\hat{\pi}^{(\mathrm{outdeg})}$ when α is 0.5 and 0.75. When $\alpha = 0.75$, $\hat{\pi}^{(\mathrm{outdeg})}$ yields the largest D_{TV}. For $\alpha = 0.1$ and 0.25, $\hat{\pi}^{(\mathrm{ren})}$ and $\hat{\pi}^{(\mathrm{outdeg})}$ yield similar results. For all parameter values $\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$ slightly outperforms $\hat{\pi}^{(\mathrm{ren})}$. We tried $n_s = 200$ (not shown), which gave similar s.d. for $\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$, and similarly for $\hat{\pi}^{(\mathrm{ren})}$, except for $\alpha = 0.1$, where, for example, $\lambda = 15$ yielded the s.d. values of 0.0039 and 0.0073 for $n_s = 500$ and $n_s = 200$, respectively.

Table 3. Mean and average s.d. of D_{TV} for the directed random graph when $d_i^{(un)} + d_i^{(out)}$ is observed. We set $N = 1{,}000$

(a) $\alpha = 0.1$

λ	$\hat{\pi}^{(\mathrm{uni})}$	$\hat{\pi}^{(\mathrm{outdeg})}$	$\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\mathrm{ren})}$	s.d.
5	0.185	**0.074**	**0.074**	0.075	0.004
10	0.131	**0.045**	**0.045**	0.047	0.003
15	0.106	0.036	**0.035**	0.037	0.002

(b) $\alpha = 0.25$

	$\hat{\pi}^{(\mathrm{uni})}$	$\hat{\pi}^{(\mathrm{outdeg})}$	$\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\mathrm{ren})}$	s.d.
	0.203	0.135	**0.132**	0.133	0.006
	0.140	0.081	**0.079**	0.080	0.003
	0.112	0.063	**0.061**	0.063	0.002

(c) $\alpha = 0.5$

λ	$\hat{\pi}^{(\mathrm{uni})}$	$\hat{\pi}^{(\mathrm{outdeg})}$	$\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\mathrm{ren})}$	s.d.
5	0.246	0.225	**0.214**	0.215	0.010
10	0.160	0.136	**0.127**	0.128	0.004
15	0.125	0.105	**0.098**	0.099	0.003

(d) $\alpha = 0.75$

	$\hat{\pi}^{(\mathrm{uni})}$	$\hat{\pi}^{(\mathrm{outdeg})}$	$\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\mathrm{ren})}$	s.d.
	0.303	0.318	**0.294**	0.295	0.014
	0.188	0.201	**0.177**	0.178	0.006
	0.144	0.156	**0.135**	0.135	0.004

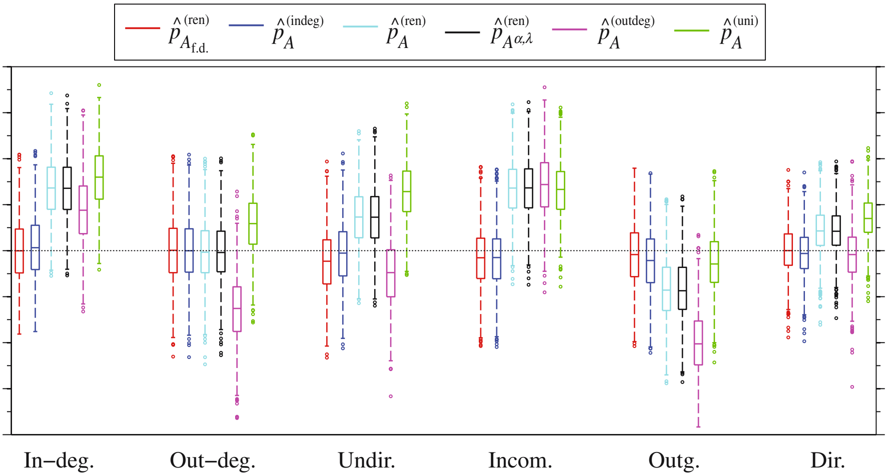


Fig. 3. Deviations of estimated $\hat{p}_A$ from the true value in the directed Erdős-Rényi graphs with $N = 1{,}000$, $\alpha = 0.75$, $\lambda = 10$, $p = 0.5$, and $n_s = 500$. Each group of boxplots corresponds to $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_{A}$, $\hat{p}^{(ren)}_{A}$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, $\hat{p}^{(outdeg)}_{A}$, and $\hat{p}^{(uni)}_{A}$ for one allocation of the individual property A. The abbreviations for the allocations corresponds to the function g, that is, In-deg. equals $\left(d^{(un)}_i + d^{(in)}_i\right)$, Out-deg. $\left(d^{(un)}_i + d^{(out)}_i\right)$, Undir. $d^{(un)}_i$, Incom. $d^{(in)}_i$, Outg. $d^{(out)}_i$, and Dir. $\left(d^{(in)}_i + d^{(out)}_i\right)$.

To compare estimated p_A, we generated 1,000 networks for each combination of the parameters $\alpha \in \{0.25,\ 0.5,\ 0.75\}$ and $\lambda = 10$. On each of these networks we in turn allocate the property A in each of the six ways described in Section 1. The probability of a vertex having A is denoted by $p \in \{0.2,\ 0.5\}$. For each network and allocation, we simulate a random walk with length $n_s \in \{200,\ 500\}$ and calculate the differences between the estimated and the actual proportions of the population with property A. In Figure 3, results for $\alpha = 0.75$, $p = 0.5$, and $n_s = 500$ are shown. The six groups of boxplots correspond to the six different ways of allocating A (see Section 1). The six boxplots in each group correspond to $\hat{p}^{(\mathrm{ren})}_{A_{\mathrm{f.d.}}}$, $\hat{p}^{(\mathrm{indeg})}_{A}$, $\hat{p}^{(\mathrm{ren})}_{A}$, $\hat{p}^{(\mathrm{ren})}_{A_{\alpha,\lambda}}$, $\hat{p}^{(\mathrm{outdeg})}_{A}$, and $\hat{p}^{(\mathrm{uni})}_{A}$, respectively.

We see that the bias of $\hat{p}^{(\mathrm{ren})}_{A_{\mathrm{f.d}}}$ and $\hat{p}^{(\mathrm{indeg})}_{A}$ is small for all allocations, as is to be expected. For the estimators utilizing the out-degree, $\hat{p}^{(\mathrm{ren})}_{A}$, $\hat{p}^{(\mathrm{ren})}_{A_{\alpha,\lambda}}$, and $\hat{p}^{(\mathrm{outdeg})}_{A}$, Figure 3 indicates that the choice of how to allocate A has a significant impact on the performance of the estimators. When A is allocated proportional to the out-degree (Out-deg. in Figure 3), $\hat{p}^{(\mathrm{ren})}_{A}$ and $\hat{p}^{(\mathrm{ren})}_{A_{\alpha,\lambda}}$ yield the most accurate result, and when A is allocated proportional to the number of directed edges (Dir. in Figure 3), $\hat{p}^{(\mathrm{outdeg})}_{A}$ is most accurate. This is true for almost all parameter combinations. In general, the bias and variance increase with both α and p for all estimators, and a small n_s results in an increased variance, as is to be expected. In the supplemental data, these findings are further illustrated by numerical results with (α, p, s) equal to (0.5, 0.2, 500), (0.25, 0.5, 500), and (0.75, 0.5, 200). The supplemental file is available at: http://dx.doi.org/jos-2016-0023

5.2. *Networks With Power-Law Degree Distributions*

To generate power-law networks, we set the expected total number of edges for each vertex to 16, while we set the expected number of undirected and directed edges equal to $(E(D^{\,(\mathrm{un})}),$

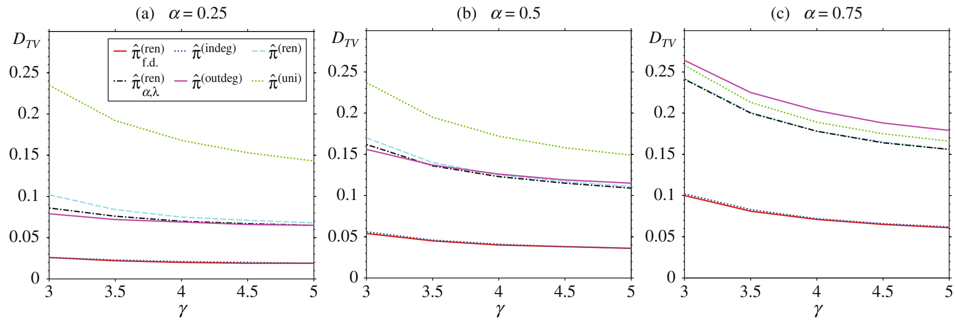


Fig. 4. Average D_{TV} between the true stationary distribution and $\hat{\pi}^{(ren)}_{f.d.}$, $\hat{\pi}^{(indeg)}$, $\hat{\pi}^{(ren)}$, $\hat{\pi}^{(ren)}_{\alpha,\lambda}$, $\hat{\pi}^{(outdeg)}$, and $\hat{\pi}^{(uni)}$ in the power-law networks with $N = 1{,}000$, α equal to a) 0.25, b) 0.5, and c) 0.75, and $n_s = 500$.

$E(D^{(in)} + D^{(out)}))$ = (12, 4), (8, 8), and (4, 12). The three cases yield $\alpha = 0.25$, 0.5, and 0.75, respectively. For each combination of the parameters, we generate 1,000 networks of size $N = 1{,}000$ and calculate the mean of the D_{TV}. We also calculate the s.d., which is of magnitude 10^{-3} and therefore not shown. The sample size n_s is set to 200 and 500.

The average D_{TV} values for $\hat{\pi}^{(ren)}_{f.d.}$, $\hat{\pi}^{(indeg)}$, $\hat{\pi}^{(ren)}$, $\hat{\pi}^{(ren)}_{\alpha,\lambda}$, $\hat{\pi}^{(outdeg)}$, and $\hat{\pi}^{(uni)}$ are shown in Figure 4 for various α and γ values. Figure 4 suggests that $\hat{\pi}^{(ren)}_{f.d}$ and $\hat{\pi}^{(indeg)}$ are the most accurate among the four estimators, with $\hat{\pi}^{(ren)}_{f.d}$ being slightly better. When $\alpha = 0.25$ and 0.5, $\hat{\pi}^{(ren)}_{\alpha,\lambda}$ has a lower mean D_{TV} than $\hat{\pi}^{(ren)}$, but this difference is not seen when $\alpha = 0.75$. $\hat{\pi}^{(outdeg)}$ performs better than $\hat{\pi}^{(ren)}$ for all values of γ when $\alpha = 0.25$, and the opposite result holds true when $\alpha = 0.75$.

In Figure 5, the results for $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_{A}$, $\hat{p}^{(ren)}_{A}$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, $\hat{p}^{(outdeg)}_{A}$, and $\hat{p}^{(uni)}_{A}$ when $\gamma = 3$, $E(D^{(un)}) = 4$, $E(D^{(in)} + D^{(out)}) = 12$, $p = 0.2$, and $n_s = 500$ are shown. The figure indicates that $\hat{p}^{(ren)}_{A_{f.d.}}$ and $\hat{p}^{(indeg)}_{A}$ have small bias across different allocations of A. In contrast, the magnitude of the bias of $\hat{p}^{(ren)}_{A}$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, and $\hat{p}^{(outdeg)}_{A}$ depends on the allocation type; $\hat{p}^{(ren)}_{A}$ has the smallest bias when A is allocated proportional to the undirected degree, and $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$

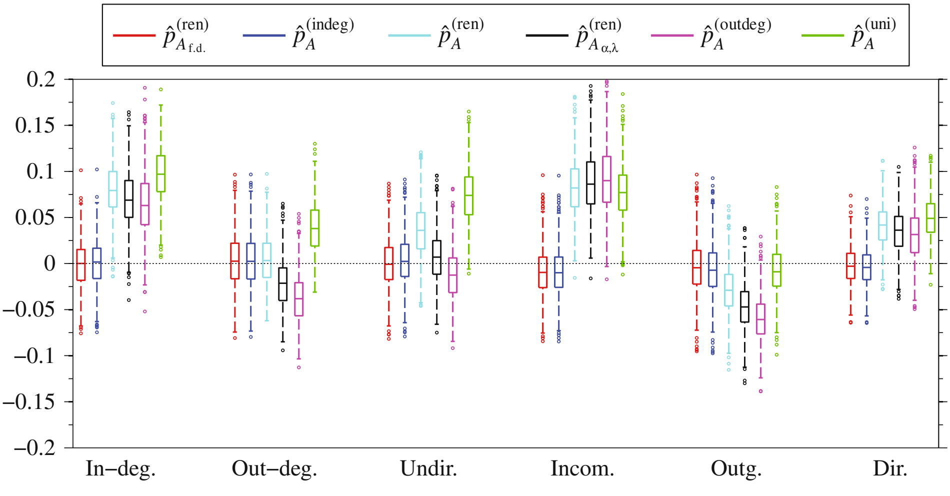


Fig. 5. Deviations of estimated p_A from the true population proportion in the power-law networks for $\gamma = 3$, $E(D^{(un)}) = 4$, $E(D^{(in)} + D^{(out)}) = 12$, $p = 0.2$, and $n_s = 500$. Each group of boxplots corresponds to $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_{A}$, $\hat{p}^{(ren)}_{A}$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, $\hat{p}^{(outdeg)}_{A}$, and $\hat{p}^{(uni)}_{A}$, for one allocation of A.

Table 4. D_{TV} between the true stationary distribution and $\hat{\pi}^{(ren)}_{f.d.}$, $\hat{\pi}^{(indeg)}$, $\hat{\pi}^{(ren)}$, $\hat{\pi}^{(outdeg)}$ and $\hat{\pi}^{(uni)}$. S.d. is shown in the second row, but only applies to $\hat{\pi}^{(ren)}_{f.d.}$ and $\hat{\pi}^{(ren)}$.

$\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$	$\hat{\pi}^{(\mathrm{indeg})}$	$\hat{\pi}^{(\mathrm{ren})}$	$\hat{\pi}^{(\mathrm{outdeg})}$	$\hat{\pi}^{(\mathrm{uni})}$
0.2198	0.2248	0.4057	0.4290	0.4484
0.0004	–	0.0048	–	

and $\hat{p}_A^{(\mathrm{outdeg})}$ when A is allocated proportional to the out-degree. Their relative performance is hard to assess for other allocations. In general, a large fraction of directed edges, small γ, and large p increase bias and variance, and variance decreases with n_s. The supplemental data contains numerical results for $(\gamma, E(D^{(\mathrm{un})}), E(D^{(\mathrm{in})} + D^{(\mathrm{out})}), p, s) = (4.5, 4, 12, 0.2, 500), (4.5, 4, 12, 0.5, 500), (4.5, 12, 4, 0.5, 500)$, and $(3, 4, 12, 0.2, 200)$ to further support these results.

5.3. Online MSM Network

For the Qruiser online MSM network, we first evaluate $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$, $\hat{\pi}^{(\mathrm{indeg})}$, $\hat{\pi}^{(\mathrm{ren})}$, $\hat{\pi}^{(\mathrm{outdeg})}$ and $\hat{\pi}^{(\mathrm{uni})}$. The results are shown in Table 4. Note that $\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$ is not evaluated because α and λ are not known beforehand. For $\hat{\pi}^{(\mathrm{uni})}$, $\hat{\pi}^{(\mathrm{outdeg})}$, and $\hat{\pi}^{(\mathrm{indeg})}$, D_{TV} to the true selection probabilities is exactly calculated. For $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$ and $\hat{\pi}^{(\mathrm{ren})}$, we show the mean and s.d. of D_{TV} on the basis of 1,000 samples of size 500. We see that $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$ has smaller D_{TV} than $\hat{\pi}^{(\mathrm{indeg})}$, and that the mean D_{TV} of $\hat{\pi}^{(\mathrm{ren})}$ is smaller than that of $\hat{\pi}^{(\mathrm{uni})}$ and $\hat{\pi}^{(\mathrm{outdeg})}$.

In Figure 6, we show estimates of the population proportions of the age, county, civil status, and profession properties. The true population proportions are shown by the dashed lines. The sample size is 500. Figure 6 indicates that $\hat{p}^{(\mathrm{ren})}_{A_{\mathrm{f.d.}}}$ performs best of all estimators. Among the estimators utilizing $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$, $\hat{p}_A^{(\mathrm{ren})}$ has the smallest overall bias. Moreover, the variance of $\hat{p}_A^{(\mathrm{ren})}$ is smaller than for $\hat{p}_A^{(\mathrm{outdeg})}$ for all properties, in particular the civil status.

6. Conclusion and Discussion

We developed statistical procedures for the random walk on directed networks to account for the empirical fact that social networks generally include nonreciprocal edges. The proposed estimation procedures typically outperformed the considered existing methods that neglect directed edges in the scenarios investigated in the simulations. In the present study, the best accuracy of estimation was obtained when undirected, incoming, and outgoing degree are observed separately for sampled individuals. In this case, our estimator $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$ should be compared to $\hat{\pi}^{(\mathrm{indeg})}$ when the expectations of the degree distributions are known. In Tables 2 and 4, and Figure 4, it is seen that $\hat{\pi}^{(\mathrm{ren})}_{\mathrm{f.d.}}$ performs slightly better than $\hat{\pi}^{(\mathrm{indeg})}$ in all the studied situations. The corresponding estimated proportions given by $\hat{p}^{(\mathrm{ren})}_{A_{\mathrm{f.d.}}}$ and $\hat{p}_A^{(\mathrm{indeg})}$ in Figures 3, 5, and 6 are very similar. In the more realistic scenario in which only the sum of undirected and outgoing edges of sampled individuals is known, all estimation procedures are less precise. In this situation, we compare our new estimator $\hat{\pi}^{(\mathrm{ren})}$ with the estimator $\hat{\pi}^{(\mathrm{outdeg})}$ that one would use if ignoring the direction of edges (Tables 3 and 4, and Figure 4). We also include $\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$ in the

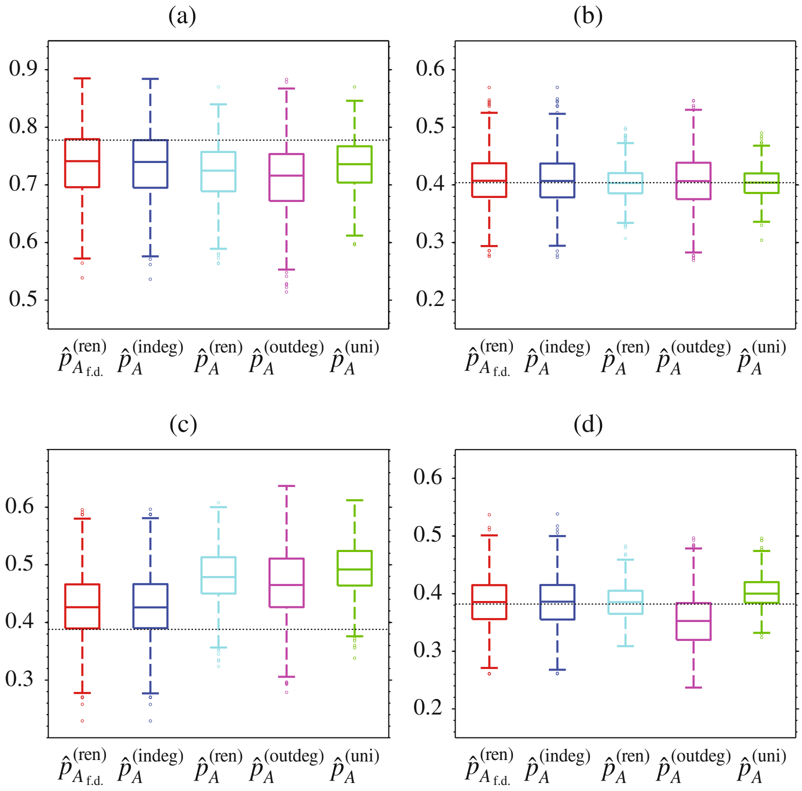


Fig. 6. Estimates of population proportions in the Qruiser network for a) age, b) civil status, c) county, and d) profession. Each figure shows $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_{A}$, $\hat{p}^{(ren)}_{A}$, $\hat{p}^{(outdeg)}_{A}$, and $\hat{p}^{(uni)}_{A}$. The true population proportions are shown by the dashed lines and are equal to 0.77, 0.40, 0.39, and 0.38 for age, civil status, county, and profession, respectively.

comparison for the generated networks, and it can be seen that the performance of $\hat{\pi}^{(\mathrm{ren})}_{\alpha,\lambda}$ is only slightly better than that of $\hat{\pi}^{(\mathrm{ren})}$. Because $\hat{\pi}^{(\mathrm{ren})}$ will deviate further from $\hat{\pi}^{(\mathrm{outdeg})}$ when $\hat{\alpha}$ increases, as seen in Equation (14), it outperforms $\hat{\pi}^{(\mathrm{outdeg})}$ except when the fraction of directed edges α is small (0.1 in Table 3 and 0.25 in Figure 4). Our simulations showed that estimators of population proportions were highly sensitive to how the property of interest is allocated in the social network. For example, Figures 3 and 5 indicate that the results of the estimators $\hat{p}^{(\mathrm{ren})}_{A}$, $\hat{p}^{(\mathrm{ren})}_{A_{\alpha,\lambda}}$, and $\hat{p}^{(\mathrm{outdeg})}_{A}$ depend strongly on the allocation of the property A. We believe that the question of how properties are distributed in empirical social networks is of interest to further study.

It is generally believed that recruitment does not happen over nonreciprocal edges in RDS, which is however refuted by the examples in Section 1. Furthermore, recruitment over nonreciprocal edges may occur on a relatively large scale in the presence of coupon selling, that is, when respondents trade coupons instead of randomly distributing them among their peers in order to increase their personal profit from study participation. Coupon selling is a side effect of the dual incentive system of RDS. It has been observed by, for example, Scott (2008) in an RDS study of IDUs in Chicago, where interviews with participants indicated that coupon selling was common and also that it had side effects such as increased risk exposure and violence among participants (see also Broadhead 2008; Ouellet 2008). For additional examples of coupon selling, see Johnston et al. (2008) and the references therein, where implications of the size of the incentives and the practical study setup on coupon selling are also discussed. In RDS studies where there is evidence of coupon selling, it might be difficult to obtain valid information on the occurrence of nonreciprocal recruitments, and then the possibility of such recruitments should be taken into account for estimation.

Information on the nonreciprocal edges in the network can be obtained from several sources. The fraction of directed edges, α, may be known for some social networks, and then we can estimate the total mean degree λ using only the mean sample out-degree in Equation (13). If α is not known, it may be estimated by utilizing additional information from an RDS sample. As previously discussed in Section 1, in the majority of RDS studies respondents quantify the nature of the relationship with their recruiter. Through this, the proportion of recruitments that occur over nonreciprocal edges (i.e., coupons passed from strangers) can be obtained and used as an estimate of α in Equation (13). In Gile, Johnston, and Salganik (2015), an alternative estimation procedure is given. This procedure utilizes several questions on respondents' degrees that serve to calculate the differences between the number of incoming and outgoing edges, which are then used to produce an estimate. However, as the authors point out, this procedure may be subject to large reporting errors. Additionally, an alternative to the standard formulation for assessing reciprocation is given in the same paper. It is also possible to estimate α through information on the number of revisits m used in Equation (12). This could be done by asking, for example, "Would you give a coupon to the person who gave you a coupon if he or she had not yet participated in the study?". This has been done in RDS studies (e.g., Bui et al. 2014), but the question may be cognitively difficult for respondents.

An alternative strategy would be to develop a sampling procedure that accounts for a directed social network of the population, that is, in which it is possible to determine whether an edge is undirected, incoming, or outgoing from a vertex, and then utilize this information for estimation. For example, in some RDS studies, the characteristics of neighbors of respondents have been collected (see Lu 2013 and the references therein). If such data were also to include, for example, the number of undirected, incoming, and outgoing edges of an individual, they could be useful in RDS estimation. As previously noted, however, it is difficult for respondents to provide such data. Alternatively, the sampling procedure could be adapted to the case of directed social networks by encouraging respondents to recruit people that are less known to them. Then, one could expect that recruitment takes place on nonreciprocal edges to a larger extent and possibly more easily identify and account for these recruitments in estimation. However, such a sampling scheme may reduce the ability of RDS to successfully penetrate the population, and may also suffer from difficulties in deciding on edge directions from sampled data.

In the present study, we considered RDS estimators that are based on the random walk framework for estimation. It could also be of interest to consider the RDS estimators of Gile (2011), Gile and Handcock (2015), Lu et al. (2013), and Lu (2013) mentioned in Section 1 for the situations studied in this article. The estimator of Gile (2011), while not adapted to the case of directed networks, is in a sense a combination of the $\hat{p}_A^{(\text{outdeg})}$ and

$\hat{p}_A^{(\mathrm{uni})}$ estimators. Hence, it can be expected to perform better than our estimator in cases where a combination of these two estimators would be favorable (e.g., when A is allocated proportional to out-degree in Figure 5 in the supplemental data), given that prior information on the population size is available. The model-assisted approach of Gile and Handcock (2015) incorporates network structural properties through an exponential random-graph model (ERGM) (e.g., Robins et al. 2007) for the network. Hence, it might be less sensitive to the different allocations of the property A that were seen to have relatively large effects on the estimators considered in our simulations. Additionally, the ERGM should not be difficult to extend to the case of directed networks. The estimator in Lu et al. (2013) is similar to our estimators in that it is developed for directed networks and could be expected to perform similarly to $\hat{p}_A^{(\mathrm{indeg})}$ given that prior information on the ratio of average in-degrees of groups in the network is available. The estimator of Lu (2013) has performed well in a recent evaluation (Verdery et al. 2015) and it could be of interest to extend it to the case of directed networks. In future work, it would be of interest to make a comprehensive evaluation of the performance of the estimators presented in this article as well as other RDS estimators, both random walk-based and nonrandom walk-based, on simulated RDS samples and data from actual RDS studies.

The main focus of the present article was on accounting for directed edges in a social network when performing RDS. There are also other assumptions in existing estimation procedures (including the current one) worthy of relaxing. For example, the methods typically assume that participants choose coupon recipients uniformly at random among their neighbors in the social network. In reality, they probably are more likely to sample closely connected neighbors, which may bias estimators of selection probabilities. Extending the RDS methods by allowing weighted edges warrants future work. It should be noted that our methods allow the two weights on the same undirected edge in the opposite directions to be different, because our framework targets directed networks. Alternatively, it is also possible that some of the previously mentioned recently developed estimators could be extended to the case of directed weighted networks.

7. References

Abramovitz, D., E.M. Volz, S.A. Strathdee, T.L. Patterson, A. Vera, and S.D. Frost. 2009. "Using Respondent Driven Sampling in a Hidden Population at Risk of HIV Infection: Who Do HIV-Positive Recruiters Recruit?" *Sexually Transmitted Diseases* 36: 750–756. Doi: http://dx.doi.org/10.1097/OLQ.0b013e3181b0f311.

Bernhardt, A., M.W. Spiller, and D. Polson. 2013. "All Work and No Pay: Violations of Employment and Labor Laws in Chicago, Los Angeles and New York City." *Social Forces* 91: 725–746. Doi: http://dx.doi.org/10.1093/sf/sos193.

Boldi, P., M. Rosa, M. Santini, and S. Vigna. 2011. "Layered Label Propagation: A Multiresolution Coordinate-Free Ordering for Compressing Social Networks." In Proceedings of the 20th International Conference on World Wide Web. 587–596. Available at: dl.acm.org/citation.cfm?id=1963405. (accessed Feb 2014).

Boldi, P. and S. Vigna. 2004. "The Webgraph Framework I: Compression Techniques." In Proceedings of the 13th International Conference on World Wide Web. 595–602. Available at: dl.acm.org/citation.cfm?id=988672. (accessed Feb 2014).

Broadhead, R.S. 2008. "Notes on a Cautionary (Tall) Tale About Respondent-Driven Sampling: A Critique of Scott's Ethnography." *The International Journal of Drug Policy* 19: 235–237. Doi: http://dx.doi.org/10.1016/j.drugpo.2008.02.014.

Bui, T., J. Nyoni, M. Ross, J. Mbwambo, C. Markham, and S. McCurdy. 2014. "Sexual Motivation, Sexual Transactions and Sexual Risk Behaviors in Men Who Have Sex with Men in Dar es Salaam, Tanzania." *AIDS and Behavior* 18: 2432–2441. Doi: http://dx.doi.org/10.1007/s10461-014-0808-x.

Chung, F. and L.Y. Lu. 2002. "The Average Distances in Random Graphs with Given Expected Degrees." *Proceedings of the National Academy of Sciences of the United States of America* 99: 15879–15882. Doi: http://dx.doi.org/10.1073/pnas.252631999.

Chung, F., L.Y. Lu, and V. Vu. 2003. "Spectra of Random Graphs with Given Expected Degrees." *Proceedings of the National Academy of Sciences of the United States of America* 100: 6313–6318. Doi: http://dx.doi.org/10.1073/pnas.0937490100.

Deaux, E. and J. Callaghan. 1985. "Key Informant Versus Self-Report Estimates of Health-Risk." *Evaluation Review* 9: 365–368. Doi: http://dx.doi.org/10.1177/0193841X8500900308.

Dombrowski, K., B. Khan, J. Moses, E. Channell, and E. Misshula. 2013. "Assessing Respondent Driven Sampling for Network Studies in Ethnographic Contexts." *Advances in Anthropology* 3: 1–9. Doi: http://dx.doi.org/10.4236/aa.2013.31001.

Donato, D., L. Laura, S. Leonardi, and S. Millozzi. 2004. "Large Scale Properties of the Webgraph." *European Physical Journal B* 38: 239–243. Doi: http://dx.doi.org/10.1140/epjb/e2004-00056-6.

Doyle, P.G. and J.L. Snell. 1984. *Random Walks and Electric Networks*. The Mathematical Association of America: Washington.

Erdős, P. and A. Renyi. 1960. "On the Evolution of Random Graphs." *Publications of the Mathematical Institute of the Hungarian Academy of Science* 5: 17–61.

Erickson, B.H. 1979. "Some Problems of Inference from Chain Data." *Sociological Methodology* 10: 276–302. Doi: http://dx.doi.org/10.2307/270774.

Feller, W. 1950. *An Introduction to Probability Theory and Its Applications*, Vol. 1. New York: Wiley.

Fortunato, S., M. Boguñá, A. Flammini, and F. Menczer. 2008. "Approximating PageRank from In-Degree." In *Algorithms and Models for the Web-Graph*, edited by W. Aiello, A. Broder, J. Janssen, and E. Milios, 59–71. Heidelberg: Springer.

Freeman, L.C., C.M. Webster, and D.M. Kirke. 1998. "Exploring Social Structure Using Dynamic Three-Dimensional Color Images." *Social Networks* 20: 109–118. Doi: http://dx.doi.org/10.1016/S0378-8733(9700016-6).

Ghoshal, G. and A.L. Barabási. 2011. "Ranking Stability and Super-Stable Nodes in Complex Networks." *Nature Communications* 2: 394. Doi: http://dx.doi.org/10.1038/ncomms1396.

Gile, K.J. 2011. "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation." *Journal of the Americal Statistical Association* 106: 135–146. Doi: http://dx.doi.org/10.1198/jasa.2011.ap09475.

Gile, K.J. and M.S. Handcock. 2010. "Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological Methodology* 40: 285–327. Doi: http://dx.doi.org/10.1111/j.1467-9531.2010.01223.x.

Gile, K.J. and M.S. Handcock. 2015. “Network Model-Assisted Inference from Respondent-Driven Sampling Data.” *Journal of the Royal Statistical Society A* 178: 619–639. Doi: http://dx.doi.org/10.1111/rssa.12091.

Gile, K.J., L.G. Johnston, and M.J. Salganik. 2015. “Diagnostics for Respondent-Driven Sampling.” *Journal of the Royal Statistical Society A* 178: 241–269. Doi: http://dx.doi.org/10.1111/rssa.12059.

Gleiser, P. and L. Danon. 2003. “Community Structure in Jazz.” *Advances in Complex Systems* 6: 565–573. Doi: http://dx.doi.org/10.1142/S0219525903001067.

Goel, S. and M.J. Salganik. 2010. “Assessing Respondent-Driven Sampling.” *Proceedings of the National Academy of Sciences of the United States of America* 107: 6743–6747. Doi: http://dx.doi.org/10.1073/pnas.1000261107.

Goh, K.I., B. Kahng, and D. Kim. 2001. “Universal Behavior of Load Distribution in Scale-Free Networks.” *Physical Review Letters* 87: 278701-4. Doi: http://dx.doi.org/10.1103/PhysRevLett.87.278701.

Gong, N.Z. and W. Xu. 2014. “Reciprocal Versus Parasocial Relationships in Online Social Networks.” *Social Network Analysis and Mining* 4: 1–14. Doi: http://dx.doi.org/10.1007/s13278-014-0184-6.

Hakre, S., G. Arteaga, A. Núñez, N. Arambu, B. Aumakhan, M. Liu, S. Peel, J. Pascale, and P. Scott. 2014. “Prevalence of HIV, Syphilis, and Other Sexually Transmitted Infections among MSM from Three Cities in Panama.” *Journal of the Urban Health* 91: 793–808. Doi: http://dx.doi.org/10.1007/s11524-014-9885-4.

Heckathorn, D.D. 1997. “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations.” *Social Problems* 44: 174–199. Doi: http://dx.doi.org/10.2307/3096941.

Hobkirk, A.L., M.H. Watt, K.T. Green, J.C. Beckham, D. Skinner, and C.S. Meade. 2015. “Mediators of Interpersonal Violence and Drug Addiction Severity Among Methamphetamine Users in Cape Town, South Africa.” *Addictive Behaviors* 42: 167–171. Doi: http://dx.doi.org/10.1016/j.addbeh.2014.11.030.

Johnston, L.G., M. Malekinejad, C. Kendall, I.M. Iuppa, and G.W. Rutherford. 2008. “Implementation Challenges to Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance: Field Experiences in International Settings.” *AIDS and Behavior* 12: 131–141. Doi: http://dx.doi.org/10.1007/s10461-008-9413-1.

Kazerooni, P.A., N. Motazedian, M. Motamedifar, M. Sayadi, M. Sabet, M.A. Lari, and K. Kamali. 2013. “The Prevalence of Human Immunodeficiency Virus and Sexually Transmitted Infections Among Female Sex Workers in Shiraz, South of Iran: By Respondent-Driven Sampling.” *International Journal of STD and AIDS* 25: 155–161. Doi: http://dx.doi.org/10.1177/0956462413496227.

Killworth, P.D. and H.R. Bernard. 1976. “Informant Accuracy in Social Network Data.” *Human Organization* 35: 269–286.

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. “What is Twitter, a Social Network or a News Media?” In Proceedings of the 19th International Conference on World Wide Web. 591–600. Available at: dl.acm.org/citation.cfm?id=1772690 (accessed Feb 2014).

Langville, A.N., and C.D. Meyer. 2006. *Google’s PageRank and Beyond*. Princeton: Princeton University Press.

Levin, D.A., Y. Peres, and E.L. Wilmer. 2009. *Markov Chains and Mixing Times*. Providence: American Mathematical Society.

Lovász, L. 1993. "Random Walks on Graphs: A Survey." *Bolyai Society Mathematical Studies* 2: 1–46.

Lu, X. 2013. "Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-Driven Sampling." *Social Networks* 35: 669–685. Doi: http://dx.doi.org/10.1016/j.socnet.2013.10.001.

Lu, X., L. Bengtsson, T. Britton, M. Camitz, B.J. Kim, A. Thorson, and F. Liljeros. 2012. "The Sensitivity of Respondent-Driven Sampling." *Journal of the Royal Statistical Society A* 175: 191–216. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00711.x.

Lu, X., J. Malmros, F. Liljeros, and T. Britton. 2013. "Respondent-Driven Sampling on Directed Networks." *Electronic Journal of Statistics* 7: 292–322. Doi: http://dx.doi.org/10.1214/13-EJS772.

Magnani, R., K. Sabin, T. Saidel, and D. Heckathorn. 2005. "Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance." *AIDS* 19 (Supplement 2): 67–72. Doi: http://dx.doi.org/10.1097/01.aids.0000172879.20628.e1.

Marsden, P.V. 1990. "Network Data and Measurement." *Annual Review of Sociology* 16: 435–463. Doi: http://dx.doi.org/10.1146/annurev.so.16.080190.002251.

Masuda, N. and H. Ohtsuki. 2009. "Evolutionary Dynamics and Fixation Probabilities in Directed Networks." *New Journal of Physics* 11: 033012. Doi: http://dx.doi.org/10.1088/1367-2630/11/3/033012.

McCreesh, N., S.D.W. Frost, J. Seeley, J. Katongole, M.N. Tarsh, R. Ndunguse, F. Jichi, N.L. Lunel, D. Maher, L.G. Johnston, P. Sonnenberg, A.J. Copas, R.J. Hayes, and R.G. White. 2012. "Evaluation of Respondent-Driven Sampling." *Epidemiology* 23: 138–147. Doi: http://dx.doi.org/10.1097/EDE.0b013e31823ac17c.

Mislove, A., M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. 2007. "Measurement and Analysis of Online Social Networks." In Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement. 29–42. October 23–26, 2007 San Diego, CA, USA. Available at: dl.acm.org/citation.cfm?id=1298306 (accessed Feb 2014).

Moreno, J.L. 1960. *The Sociometry Reader*. New York: Free Press.

Muhib, F.B., L.S. Lin, A. Stueve, R.L. Miller, W.L. Ford, W.D. Johnson, and P.J. Smith, Community Intervention Trial for Youth Study Team. 2001. "A Venue-Based Method for Sampling Hard-to-Reach Populations." *Public Health Reports* 116 (Suppl. 1): 216–222.

Newman, M. 2010. *Networks: an Introduction*. Oxford: Oxford University Press.

Newman, M.E., S. Forrest, and J. Balthrop. 2002. "Email Networks and the Spread of Computer Viruses." *Physical Review E* 66: 035101. Doi: http://dx.doi.org/10.1103/PhysRevE.66.035101.

Ouellet, L.J. 2008. "Cautionary Comments on an Ethnographic Tale Gone Wrong." *International Journal of Drug Policy* 19: 238–240. Doi: http://dx.doi.org/10.1016/j.drugpo.2008.02.013.

Paquette, D.M., J. Bryant, and J.D. Wit. 2011. "Use of Respondent-Driven Sampling to Enhance Understanding of Injecting Networks: A Study of People Who Inject Drugs in

Sydney, Australia." *International Journal of Drug Policy* 22: 267–273. Doi: http://dx.doi.org/10.1016/j.drugpo.2011.03.007.

Phillips II, G., L.M. Kuhns, R. Garofalo, and B. Mustanski. 2014. "Do Recruitment Patterns of Young Men Who Have Sex With Men (YMSM) Recruited Through Respondent-Driven Sampling (RDS) Violate Assumptions?" *Journal of Epidemiology and Community Health* 68: 1207–1212. Doi: http://dx.doi.org/10.1136/jech-2014-204206.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher. 2007. "An Introduction to Exponential Random Graph (p*) Models for Social Networks." *Social Networks* 29: 173–191. Doi: http://dx.doi.org/10.1016/j.socnet.2006.08.002.

Rybski, D., S.V. Buldyrev, S. Havlin, F. Liljeros, and H.A. Makse. 2009. "Scaling Laws of Human Interaction Activity." *Proceedings of the National Academy of Sciences of the United States of America* 106: 12640–12645. Doi: http://dx.doi.org/10.1073/pnas.0902667106.

Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–240. Doi: http://dx.doi.org/10.1111/j.0081-1750.2004.00152.x.

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

Schwitters, A., M. Swaminathan, D. Serwadda, M. Muyonga, R. Shiraishi, I. Benech, S. Mital, R. Bosa, G. Lubwama, and W. Hladik. 2012. "Prevalence of Rape and Client-Initiated Gender-Based Violence Among Female Sex Workers: Kampala, Uganda, 2012." *AIDS and Behavior* 19: 68–76. Doi: http://dx.doi.org/10.1007/s10461-014-0957-y.

Scott, G. 2008. "'They Got Their Program, and I Got Mine': A Cautionary Tale Concerning the Ethical Implications of Using Respondent-Driven Sampling to Study Injection Drug Users." *International Journal of Drug Policy* 19: 42–51. Doi: http://dx.doi.org/10.1016/j.drugpo.2007.11.014.

Solomon, S.S., S.H. Mehta, A.K. Srikrishnan, S. Solomon, A.M. McFall, O. Laeyendecker, D.D. Celentano, S.H. Iqbal, S. Anand, C.K. Vasudevan, S. Saravanan, G.M. Lucas, H.R. Kumar, M.S. Sulkowski, and T.C. Quinn. 2015. "Burden of Hepatitis C Virus Disease and Access to Hepatitis C Virus Services in People Who Inject Drugs in India: A Cross-Sectional Study." *The Lancet Infectious Diseases* 15: 36–45. Doi: http://dx.doi.org/10.1016/S1473-3099(14)71045-X.

Tomas, A. and K.J. Gile. 2011. "The Effect of Differential Recruitment, Non-Response and Non-Recruitment on Estimators for Respondent-Driven Sampling." *Electronic Journal of Statistics* 5: 899–934. Doi: http://dx.doi.org/10.1214/11-EJS630.

Van de Bunt, G., M. van Duijn, and T. Snijders. 1999. "Friendship Networks Through Time: An Actor-Oriented Dynamic Statistical Network Model." *Computational and Mathematical Organization Theory* 5: 167–192. Doi: http://dx.doi.org/10.1023/A:1009683123448.

Verdery, A.M., M.G. Merli, J. Moody, J.A. Smith, and J.C. Fisher. 2015. "Brief Report: Respondent-Driven Sampling Estimators Under Real and Theoretical Recruitment Conditions of Female Sex Workers in China." *Epidemiology* 26: 661–665. Doi: http://dx.doi.org/10.1097/EDE.0000000000000335.

Volz, E. and D.D. Heckathorn. 2008. “Probability Based Estimation Theory for Respondent Driven Sampling.” *Journal of Official Statistics* 24: 79–97.

Wasserman, S. and K. Faust. 1994. *Social Network Analysis*. New York: Cambridge University Press.

Zhang, S.X., M.W. Spiller, B.K. Finch, and Y. Qin. 2014. “Estimating Labor Trafficking among Unauthorized Migrant Workers in San Diego.” *The Annals of the American Academy of Political and Social Science* 653: 65–86. Doi: http://dx.doi.org/10.1177/0002716213519237.

Received February 2014
Revised October 2015
Accepted October 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 461–486, http://dx.doi.org/10.1515/JOS-2016-0024

Modernizing a Major Federal Government Survey: A Review of the Redesign of the Current Population Survey Health Insurance Questions

Joanne Pascale[1]

Measurement error can be very difficult to assess and reduce. While great strides have been made in the field of survey methods research in recent years, many ongoing federal surveys were initiated decades ago, before testing methods were fully developed. However, the longer a survey is in use, the more established the time series becomes, and any change to a questionnaire risks a break in that time series. This article documents how a major federal survey – the health insurance module of the Current Population Survey (CPS) – was redesigned over the course of 15 years through a systematic series of small, iterative tests, both qualitative and quantitative. This overview summarizes those tests and results, and illustrates how particular questionnaire design features were identified as problematic, and how improvements were developed and evaluated. While the particular topic is health insurance, the general approach (a coordinated series of small tests), along with the specific tests and methods employed, are not uniquely applicable to health insurance. Furthermore, the particular questionnaire design features of the CPS health module that were found to be most problematic are used in many other major surveys on a range of topic areas.

Key words: Health reform; questionnaire redesign; health insurance; CPS.

1. Introduction

Measurement error – the difference between the "true value" of a concept being measured and the survey estimate that represents that concept – can be very difficult to assess and reduce. While there have been great strides in the field of survey methods research in recent years, many ongoing federal surveys were initiated decades ago before testing methods were fully developed. As such, the degree and nature of measurement error associated with these surveys is often unknown. However, the longer a survey is in use, the more established the time series becomes, and any change to a questionnaire risks a break in that time series. Thus, a comprehensive survey redesign is generally approached with caution for a number of reasons. First, demonstrating that any changes actually do reduce measurement error (and do not inadvertently introduce other problems) can be elusive. Second, even if a statistic is imperfect (e.g., a point estimate is biased), if the cause of the imperfection does not interact with time, then the trend line of that point estimate can still be valid and informative. Third, the loss of the time trend to the data-user community due

[1] Center for Survey Measurement, US Census Bureau, 4600 Silver Hill Road, Suitland, MD 20233, U.S.A. Email: joanne.pascale@census.gov

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

to a break in series is sometimes untenable. Finally, research to bridge the break in series is costly and sometimes data do not exist for the task. On balance, continuing with the status quo can be the best course of action in many cases.

This article documents a middle-ground approach in which a key component of a major federal survey – the health insurance module of the Current Population Survey Annual Social and Economic Supplement (CPS ASEC, aka CPS) – was redesigned over the course of 15 years through a series of small, iterative tests, both qualitative and quantitative. The CPS is one of several surveys (federal, state, and private) that measure health insurance coverage, but the CPS is the most widely cited and used source of estimates on health coverage in the United States (Blewett and Davern 2006). This is partly due to its trend line on coverage going back to the 1980s, and because the sample size is large enough to make state-level estimates (Farley-Short 2001). It also offers rich and detailed auxiliary data on income and employment, making it a particularly valuable data source for researchers investigating connections between economic well-being, health coverage and health status. The CPS, however, has also been the subject of widespread criticism because its estimate of the uninsured tracks higher than that of other major surveys (Bhandari 2004; Congressional Budget Office 2003). For example, the estimate of those uninsured throughout calendar year 2012 was 15.4 percent in the CPS and 11.1 percent in the National Health Interview Survey (NHIS). The NHIS also produces an estimate of those uninsured at a single point in time, and in 2012, it was 14.7 percent, which happens to be close to the CPS 2012 calendar-year estimate of 15.4 percent (State Health Access Data Center 2013). Indeed, other studies have also found that the CPS calendar-year estimate was very similar to point-in-time estimates from other surveys (Rosenbach and Lewis 1998; Congressional Budget Office 2003), leading to widespread speculation over what the CPS estimate really represented – calendar year or point-in-time coverage or something in between (Lewis et al. 1998; Swartz 1986).

Due to the divergent estimates across surveys, and the persistent criticism of the CPS, in 1999 the Census Bureau began a comprehensive research program to examine and reduce measurement error associated with health insurance estimates from the CPS, focusing on the role of the questionnaire. Numerous small-scale studies – both qualitative and quantitative – were fielded and analyzed, and results were fed into subsequent small-scale tests in an iterative fashion. The research approach was to identify features of the questionnaire that were potential candidates for contributing to measurement error, explore and modify those features, and test against the status quo to assess empirical evidence for improvements due to the changes. After a decade of this testing, a fundamentally redesigned questionnaire was crafted. A formal pretest of the redesign was conducted in 2009 and the basic approach was found to be sound (Pascale 2009a). Minor refinements were made and a large-scale split-ballot experiment was conducted in March 2010 with promising results (Boudreaux et al. 2013). By chance, the very day the 2010 test was launched (March 23, 2010), the Patient Protection and Affordable Care Act (ACA) was passed. In response to that, the draft redesign was then adapted with questions specific to health reform and tested in 2011–2012 with residents of Massachusetts, given its passage of ACA-like state-level legislation in 2006. That testing proved successful (Pascale et al. 2013), and the adaptations were integrated into the redesigned questionnaire for a follow-up large-scale field test in 2013. Results were favorable to the redesign (Pascale et al. 2015), and it was officially implemented in the March 2014 CPS ASEC data collection. See Figure 1 for a side-by-side display of the basic question structure for the old and new CPS.

This article documents the 15-year history of the CPS ASEC health insurance module redesign. The tests and results are summarized and synthesized with other relevant survey methods literature to illustrate how particular questionnaire design features were identified as problematic, and how improvements were developed and evaluated. In terms of evaluation criteria, it is important to note here that the CPS and most surveys define health

Old CPS

1. In 2013 was anyone in hh covered by job plan?
   - Yes ➔2
   - No ➔6
2. Who in hh were policyholders?
   ➔3
3. Who else in hh was covered?
   ➔4
4. Did plan cover anyone outside hh?
   ➔5
5. Did emp/union pay all/part/none of premium
   ➔6
6. In 2013 was anyone in hh covered by direct plan?
   - Yes ➔7
   - No ➔10
7. Who in hh were policyholders?
   ➔8
8. Who else in hh was covered?
   ➔9
9. Did plan cover anyone outside hh?
   ➔10
10. In 2013 was anyone in hh covered by outside hh?
    - Yes ➔Who? ➔11
    - No ➔11
11. In 2013 was anyone in hh covered by Medicare?
    - Yes ➔Who? ➔12
    - No ➔12
12. In 2013 was anyone in hh covered by Medicaid/state-specific program?
    - Yes ➔Who? ➔13
    - No ➔14
13. How many months in 2013 was NAME covered?
    ➔14
14. In 2013 was anyone in hh covered by CHIP?
    - Yes ➔Who? ➔15
    - No ➔15
15. In 2013 was anyone in hh covered by military?
    - Yes ➔Who? ➔Plan type ➔16
    - No ➔16
16. In 2013 was anyone in hh covered by other/state-specific program?
    - Yes ➔Who? ➔Plan type ➔17
    - No ➔ END
17. I have recorded NAME(s) not covered; correct?
    - Yes ➔END
    - No ➔Who was covered? ➔Plan type ➔END

New CPS

CK1: If disabled or age=65+ ➔1; else ➔2

1. Are you covered by Medicare?
   - Yes ➔11
   - No ➔2
2. Are you NOW covered by any type of health plan?
   - Yes ➔3
   - No ➔Qs on Medicaid, CHIP, state-specific program names, verification of uninsured ➔15
3. Is it provided thru a job, govt, or other way?
   - Job ➔6
   - Government ➔4
   - Other way ➔7
4. Is that plan related to a JOB with the government?
   - Yes ➔6
   - No ➔5
5. Is that Medicare, Medicaid/CHIP, military, other?
   - Medicaid/CHIP/other/DK ➔9
   - Military ➔[type of military plan] ➔10
   - Medicare ➔11
6. Is the plan related to military service in any way?
   [if yes, type of military plan] ➔10
7. How is it provided – parent/spouse, direct, other?
   - Parent/spouse/direct ➔10
   - Other ➔8
8. Is it thru former emp, union, group, assn, school?
   - Former emp/union/group/assn/school ➔10
   - Other ➔9
9. What do you call the program?
   [pick list of state-specific program names] ➔11
10. Who is the policyholder? [hh roster; outside hh]
    [If Q7=parent/spouse: thru their job or direct?] ➔11
11. Did coverage start before January 1, 2013?
    - Yes ➔12
    - No ➔Qs on start month/year ➔12
12. Has coverage been continuous since then?
    [if no ➔Qs on months of spells] ➔13
13. Is anyone else in hh covered on same plan?
    [if yes, who?] ➔14
14. Were they covered same months?
    - Yes ➔15
    - No ➔What months were they covered? ➔15
15. Any [other] coverage Jan 2013 till now?
    - Yes ➔loop thru series again, starting with 3
    - No ➔CK2 for next person on roster

CK2: if any coverage already reported for person ➔15; else ➔ CK1; If no more on roster ➔END

Fig. 1. Old versus New CPS Questionnaire Structure (mimicked for March 2014 administration)

insurance by asking about coverage through a range of different sources or types of comprehensive coverage, both public (e.g., Medicaid) and private (e.g., employer-sponsored insurance). This is sometimes referred to as the "laundry-list" approach. Respondents are asked about each type of coverage and those without any of the listed types of coverage are then tabulated as uninsured (Lewis et al. 1998). While the particulars of this list vary across surveys, most exclude specialty plans (e.g., dental, vision, accident plans) and definitional differences have been found to have negligible empirical effects on the uninsured estimate (Farley-Short 2001). In the course of the CPS redesign development, some of the tests included a validation component for reports of particular *types* of coverage, but there simply is no single, comprehensive, accurate source of data on those with and without coverage in the U.S. that could serve as a gold standard for validating the uninsured estimate. Given the literature cited above suggesting that the CPS is particularly prone to underreporting of coverage (Congressional Budget Office 2003; Bhandari 2004; Lewis et al. 1998; Swartz 1986), a default "more is better" model was used to assess reduction in measurement error in the uninsured estimate.

The particular subject of this research is health insurance, but the general approach (a coordinated series of small, iterative tests), along with the specific tests and methods employed, is not uniquely applicable to health insurance. Furthermore, the particular questionnaire design features of the CPS health module that were found to be most problematic are used in many other major surveys on a range of topic areas. Thus, the article concludes with a discussion of the broader implications of this research for other topic areas and for survey redesign efforts in general. For more details, see Pascale (2015).

2. Methodological Variations in Data Sources

For many major social indicators – for example, poverty, disability, and labor force participation – there are multiple data sources, each with their own purpose and constraints in terms of data collection. That is, the purpose and budget of the survey determine key design features such as content, sample size, mode, timing, and frequency of data collection. Variation in these design features can lead to variation in the estimates, leaving researchers to weigh out the strengths and weaknesses of the measures, as well as other factors for analysis such as sample size and auxiliary variables on the dataset. As a result, often the data source that researchers come to rely on for a particular analysis is not purpose built for the specific concept under study.

In the case of health insurance, several major federal surveys came to include questions on health coverage gradually, as the health care system in the United States changed and as data needs developed. Taking them chronologically, the National Health Interview Survey (NHIS), sponsored by the National Center for Health Statistics, has been fielded since 1957 for monitoring the health of the U.S. population. In 1959, questions on private coverage were added, and in the early 1970s questions were added on large public programs that had been introduced in the 1960s – Medicaid (primarily for low-income individuals), and Medicare (primarily for those 65+) (Blumberg 2014). Starting in 1981, the CPS ASEC began including questions on certain types of health insurance and by 1987 the questions collected data on Medicaid, employer-sponsored insurance (ESI) and other noncash benefits in order to inform the poverty measure (US Census Bureau 2015a). In 1983, the Census Bureau launched the Survey of Income and Program Participation (SIPP), the main purpose of which is to measure the dynamics of economic and social well-being over time. As such, the focus is on measuring household income and benefits from a comprehensive range of sources, including health insurance (US Census Bureau 2015b). In 1997, the Agency for Healthcare Research and Quality launched the National Medical Care and Expenditure Survey (NMCES), which was replaced by the Medical Expenditure Panel Survey (MEPS) in 1996. The MEPS collects data on the usage, cost and financing of health services, as well as health insurance (Agency for Healthcare Research and Quality 2015). Finally, the American Community Survey (ACS), designed to replace the decennial census long form, began in 2005 and health insurance questions were added in 2008 (US Census Bureau 2015b).

While the breadth of data sources on health insurance has its advantages, one downside is that the methods and the estimates they produce vary across surveys and it is not clear which estimates are most accurate (Office of the Assistant Secretary for Planning and Evaluation 2005). Research going back to the 1980s suggests that much of the variation in estimates across these surveys is rooted in subtle differences in the questionnaires (Swartz 1986). This notion prompted the current line of inquiry investigating the association between questionnaire design features and measurement error. The overarching finding from this inquiry is that within the questionnaire, three fundamental design features are particularly important in terms of their potential for driving the estimates. One is the "reference period" – the time period specified in the survey question. Some surveys ask about current coverage status, while others ask about coverage over a certain time span. The reference period is then intertwined with the definition of the uninsured. For example, the CPS is administered in March and asks respondents if they had coverage "at any time" during the previous calendar year. The uninsured are then defined as those uninsured throughout the entire calendar year. The NHIS and ACS, on the other hand, both ask about current coverage and define the uninsured as those without coverage at a particular point in time (i.e., the day of the interview). The MEPS and the SIPP, both longitudinal surveys that follow respondents for a number of years, use a reference period somewhere in between. Both surveys ask about monthly coverage during the reference period and the uninsured can then be defined in several different ways – uninsured in any given month, throughout the calendar year, throughout an entire three-year panel, or any number of months in between.

Another difference across surveys is the specificity of questions about household members. For most plan types, the CPS, NHIS, and MEPS use a "household-level" approach: ". . .was anyone in this household covered by [plan type X]?" If yes, follow-up questions are asked to determine which household members have the coverage. The SIPP and ACS, on the other hand, employ a "person-level" approach and ask about each household member by name (e.g., "Does [name] have [plan type X]?"). The SIPP, furthermore, attempts self-response for all household members age 15 and older, while in the ACS, CPS, MEPS, and NHIS a single household respondent is asked questions on behalf of all household members.

A third major design feature that varies is the specificity of questions concerning the type of health coverage. The CPS, MEPS, and ACS ask a series of yes/no questions, each on a particular type of health coverage. The NHIS takes a different approach and asks a global question about any coverage and, if yes, a single follow-up question is asked to determine the particular type of coverage.

3. Problematic Design Features and Prototype Improvements

3.1. Reference Period

For at least two decades, researchers have been trying to understand the source of variation in health-coverage estimates. For example, Swartz's seminal article in 1986 examined the sampling framework, weighting procedures, adjustments for nonresponse and attrition, and questionnaire design across four different surveys (CPS, SIPP, NHIS, and the NMCES). Swartz concluded that the largest contributor to variation in the estimates was differences in the questionnaire, particularly differences in the reference period (Swartz 1986). Research comparing the calendar-year and current reference periods corroborated Swartz's findings (Rosenbach and Lewis 1998; Pascale 2001a), and prompted qualitative research to understand more about measurement error associated with the reference period. This research found that some respondents simply do not hear or do not focus on the reference period stated in the question, and instead report on their current situation (Pascale 2008). A similar investigation of the twelve-month reference period in the context of a CPS supplement on Food Stamps receipt also found a lack of attentiveness to the reference period (Hess and Singer 1995). Related literature showed that respondents tend to underreport receipt of benefits from the more distant past (Lynch 2006; National Research Council 2006; Resnick et al. 2004; Ringel and Klerman 2005), which is consistent with findings from the survey methods literature more generally (Schaeffer and Presser 2003). Quantitative studies on Medicaid corroborated the finding on underreporting of coverage in the more distant past, and also found that accuracy of past coverage improved if respondents were currently covered (Pascale et al. 2009; Research Project to Understand the Medicaid Undercount 2008). In sum, it appears there were several factors indicating that the CPS phrase "at any time during [past year]" often failed to prompt recall of events going back to the beginning of the calendar year.

A subtle but potentially important point about the reference period has to do with "lag time." This is the length of time between the interview date and the time period of coverage asked about in the survey question. In surveys with a current reference period, obviously, there is no lag time. However, in the CPS there is roughly a three-month lag time which, in light of research findings that some respondents focus on their current situation, could compound the problems with recalling and reporting past coverage. Respondents are never anchored in their current day-of-interview status but are given the task of thinking back over 15 months (from January of the previous year until March of the current year), focusing on the first twelve of those months while "subtracting out" the most recent three months. Furthermore, they are asked about coverage "at any time" during those twelve months which, technically speaking, includes coverage for as little as one day. Thus the relatively long duration of the reference period, the three-month lag time, and the fact that respondents are not asked about their current situation may all be working against the CPS.

To address these issues, an exploration of ways of asking about both current and past calendar-year coverage within the same set of questions was undertaken. The rationale was to accommodate respondents' tendencies to focus on their current status (even if instructed otherwise) and then leverage that current status as an anchor to ask about retrospective coverage. Memory and recall literature suggested that providing multiple time frames could enhance the reporting accuracy of past events (Crespi and Swineheart 1982; Sudman et al. 1984; Loftus et al. 1990; Blair and Ganesh 1991; Martin et al. 2002; Pascale 2009b). Putting the health-coverage findings together with the memory literature resulted in a question series with two time frames: the date of the interview, and the beginning of the reference period (January 1 of the previous calendar year). This approach both anchors the respondent in the present day and frames the full 15-month time span of interest.

Very generally, the series begins by asking about current coverage, and then asking whether that coverage started before January 1 of the previous year. If it did, respondents are asked if the coverage has been continuous since January 1. If so, it is inferred that the coverage was held all 15 months. If the coverage began after January 1, respondents are asked in what month the current coverage started. These respondents, along with those whose coverage was not continuous, are asked about specific months within the reference period for which they were not yet reported to have coverage. This is a key departure from the much more general line of questioning asking "At any time during [past year] . . ." that was found to be problematic in the recall literature.

3.2. Household- Versus Person-Level Questions

While reference-period issues have tended to dominate the literature on health insurance measurement error, as a design feature, reference period alone does not explain all the variation observed in the estimate of the uninsured. For example, both the SIPP and the MEPS employed reference periods shorter than the calendar year, yet the difference in their uninsured-throughout-the-year estimates for 2002 was quite striking: 8.1 percent in the SIPP and 12.9 percent in the MEPS (Davern 2009). Key differences between these surveys are the household- versus person-level approach (used in the MEPS versus SIPP, respectively), and the fact that the SIPP is a self-response survey while the MEPS asks one household member to answer questions about all other household members. The combination of the person-level approach and the self-response could account for SIPP's higher reporting of coverage relative to the MEPS.

Most surveys do not have the luxury of self-response, but the difference in the MEPS and SIPP uninsured estimates suggests there could be some benefit to "naming names" – that is, asking the household respondent about household members by name, versus a more general "anyone in the household." There is, indeed, some evidence that a failure to name each household member individually risks the respondent failing to report coverage for some members, particularly in larger or complex households (Blumberg et al. 2004; Hess et al. 2001). On the other hand, administering the entire series for each household member individually risks respondent fatigue and associated underreporting (Blumberg et al. 2004; Pascale 2001b). So while there are pros and cons to both the household- and person-level approaches, it is not entirely clear how the overall estimates are affected across all plan types and across households of various sizes and complexity.

The first experiment on the CPS redesign integrated both the household/person-level design and reference-period features in order to isolate the effects of each. This 1999 experiment employed a two-by-two design in which respondents were randomly assigned to one of four treatments: (A) calendar year/household-level (this is the status quo CPS ASEC); (B) calendar year/person-level; (C) current coverage/household-level; and (D) current coverage/person-level. In the absence of measurement error, there should be no difference in estimates within the calendar-year designs or within the current-year designs. That is, reporting should be the same whether questions are asked at the household- or person-level. However, asking about coverage "at any time" during the past calendar year should, theoretically, result in higher reporting than asking about current coverage, since the former allows for coverage on ANY of 365 days of the year and the latter only includes one – the day of the interview. Results from the experiment were telling. Within the current reference period versions, the difference between person- and household-level designs was nonsignificant – that is, there was no evidence of measurement error ascribable to the household/person-level design when asking about current coverage status. However, when asking about past calendar-year coverage, the uninsured rate in the household-level design was 5.1 percentage points higher than in the person-level designs ($p < 0.01$). This suggested that respondents could report on all household members, whether or not they were provided with the individual names, equally well when they were only asked about current coverage, but when asked about coverage during the past calendar year they reported more coverage when provided with individual household members' names (Pascale 2001a).

Within the person-level versions, the difference in reporting was observed in the expected direction. The insured rate was higher in the calendar-year design than in the current design by 3.4 percentage points ($p < 0.05$). However, within the household-level designs, the insured rate was identical – at 12.0 percent – for both calendar-year and current designs. This suggested the reference-period wording – "at any time during [previous calendar year]" – was effective at eliciting reports of past coverage when respondents were asked to think about only one person at a time, but not when they were asked to think about "anyone in this household." In short, there appeared to be some kind of cognitive overtaxing going on within the household/calendar-year design in the CPS.

To address this, a hybrid household-person-level approach was developed. Each household member is asked about by name, but whenever a specific plan or plan type is reported, a follow-up question is asked to determine whether anyone else in the household is also covered by that same plan or plan type. That information is stored and harnessed so that when subsequent household members are asked about, if they were already reported to have coverage, the question series is significantly abbreviated and only asks about any additional coverage. This allows the questions to reference each household member by name (rather than the more general "anyone in the household"), but it avoids the tedium of repeating the entire series for each household member. It also avoids redundant reporting of health plans in the vast majority of cases where multiple household members share the same type of coverage. Finally, this change, in combination with asking about current and past coverage in separate questions, greatly reduces the risk of overtaxing the respondent by loading a single question with too many demands (thinking about ALL household members, across ALL 15 months).

3.3. *Questions on Individual Coverage Types*

With regard to the specificity of questions on coverage type (the laundry-list approach noted above), a number of reporting problems has been identified in the literature (Beatty and Schechter 1998; Loomis 2000; Roman et al. 2002; Pascale 2009c; Pascale 2008; Schaeffer and Presser 2003; Willson 2005). First, the list itself is often not mutually exclusive. For example, employer-sponsored insurance (ESI) and military coverage can overlap, and coverage from someone outside the household can overlap with ESI and/or directly purchased coverage. Second, respondents can have difficulty figuring out which category best fits their coverage. For example, self-employed individuals who get coverage through a trade association are sometimes torn between the ESI and directly purchased category. In addition, dependents on a spouse or parent's job-based plan are sometimes reluctant to report their coverage as ESI because the coverage is not through *their* job but that of the spouse or parent. Third, individual questions on plan type are often too detailed and complex for respondents to grasp with confidence, or they fail to tap into the respondent's understanding of the coverage. This proved particularly problematic when respondents were answering for other household members, about whom they had only limited knowledge of personal circumstances like health coverage. For example, in many cases, respondents knew another household member was covered, and that it was a government-sponsored plan, but they were unclear on the distinction between Medicaid and Medicare (Loomis 2000; Roman et al. 2002; Willson 2005). All these factors led respondents to misreport one plan type as another, report the same plan twice, or fail to report the plan altogether (Loomis 2000; Pascale 2008).

One of the most compelling examples of this type of misreporting is demonstrated in two related studies. Loomis (2000) conducted cognitive testing of the CPS series but used two alternative versions – one with the status-quo method in which Medicare was asked about prior to Medicaid and one that reversed just the order of questions on those two plan types within the series. She found that respondents confused the two programs, which sometimes meant they reported the same coverage twice (e.g., Medicaid enrollees reported their coverage at both the Medicare and the subsequent Medicaid question). Loomis concluded it ". . .seems quite possible for Medicaid recipients to simply respond 'yes' to the first question that sounds familiar to them." (Loomis 2000, 16).

Following on from these findings, a split-ballot field test on sequencing effects of Medicare and Medicaid was conducted in 2003 in which the standard CPS series of questions was asked, but in half the sample the Medicare question preceded the Medicaid item and in the other half the sequence of these two plan types was reversed. Results suggested there was false-positive reporting of Medicare when that plan type was asked first. Among low-income households (in which individuals are more likely to be covered by Medicaid), when Medicare was asked first, the Medicare estimate was 24.8 percent, but when Medicaid was asked first the Medicare estimate was 18.6 percent. The Medicaid estimate was unaffected by question sequencing (Pascale 2004). A later validation study of people known through Blue Cross/Blue Shield records to be enrolled in Medicaid found that "a fair number of Medicaid enrollees – when asked both the Medicaid and Medicare question – answered yes to both" (Davern et al. 2008). Though this one example is suggestive of a certain pattern of misreporting, overall it is difficult to gauge the magnitude and direction of misreporting stemming from the laundry-list approach. However, there is ample evidence from this and numerous other studies of the potential for underreporting, double reporting and misreporting of plan type (Pascale 2009c).

The redesign addresses the problems with the laundry-list approach by starting with a relatively simple question about whether the respondent has coverage or not, and following up with questions that go from general to specific to identify plan type. The objective was to make each individual question easier for respondents to understand and answer, and to tap into the level of knowledge they did have. For respondents who report some kind of coverage, a follow-up question first determines the general source of coverage – through a job, the government or state, or some other way – and tailored questions from each response category obtain the necessary detail. For job-based plans, subsequent questions identify policyholders and dependents, and determine if the coverage is related to military service in any way. For government plans a follow-up question asks about the type of government plan, presenting both Medicare and Medicaid in the same list of response categories, so that respondents can assess which plan type is closest to their understanding of the coverage they have. For respondents who choose "other" as their general source of coverage, follow-up questions ask about plans obtained in the next most common ways (other than employment and government) – through direct purchase and, to accommodate dependents on someone else's private plan, through a parent or spouse. In the end, the same level of detail (in fact, more detail in some cases) is captured, but in a way that is more respondent friendly. The questions go from very general to very specific in terms of plan type, and the series accommodates respondents' sometimes-limited knowledge of other household members' health-coverage situation.

4. Testing of the Redesign

4.1. Cognitive Testing and Pretest (2008–2009)

The initial prototype of the redesigned questionnaire addressed the reference period, household-level design and the laundry-list approach. Informal testing of this experimental draft was conducted with an unpaid convenience sample in order to correct any fatal flaws before conducting cognitive testing with paid respondents. Cognitive testing was carried out in the spring of 2008 with 36 household respondents (Pascale 2009a), followed by a larger pretest in March 2009 with 54 household respondents. Minor refinements were made along the way and iteratively tested, and there was no evidence of problems with these changes (Pascale 2009b). Only one substantive design issue was identified. Technically speaking the aim of the CPS ASEC is to capture coverage during the previous calendar year, and one major goal of the redesign was to address the distinction between current and past-year coverage. Thus initial redesign questions captured only past calendar year and day-of-interview coverage. However, pretest results suggested it was, perhaps ironically, less burdensome to collect data across the entire 15-month time span with wording such as "And was [NAME] also covered from January, 2008 up until now?" than it was to ask about current coverage and then only the previous calendar year. The post-pretest questionnaire, then, was modified slightly for the next field test to ask about continuous coverage – from January 1 of the previous calendar year up to the day of the interview (that is, including the first few months of the current year).

4.2. Survey of Health Insurance and Program Participation (2010)

The 2010 split-ballot field test of the old vs new CPS, called the Survey of Health Insurance and Program Participation or SHIPP, was carried out in the spring of 2010 by the Census Bureau's telephone-interviewing staff. The sample was drawn from two sources – an RDD frame and Medicare enrollment files. Medicaid records were sought but unavailable, so Medicare records were used given the scarcity of validation studies in health-measurement research. Individuals under 65 and those recently enrolled were oversampled under the assumption that these groups were more vulnerable to misreporting – and hence more could be learned from them – than those 65 and over and/or enrolled for longer periods of time. Data were collected on 8,507 individuals, split about 60/40 across the RDD/Medicare samples. Response rates (based on the AAPOR RR4 definition) were 47.6 percent and 61.4 percent for the RDD and Medicare samples, respectively.

Results showed that there were no statistically significant differences in calendar-year estimates of the uninsured, or estimates of coverage by plan type, between the old and new CPS (Boudreaux et al. 2013). However, the direction of differences favored the new CPS, and coverage and sample bias could have been major contributors to the lack of statistical significance in the differences. For cost reasons, the SHIPP survey was entirely landline-telephone-based and did not include a cell-phone-only or face-to-face component, introducing an overall bias to both questionnaire panels. In the latter half of 2010 almost 30 percent of households in the U.S. were cell-phone only, and individuals living in those households were more likely to be young adults (in particular, those aged 25–29), living in or near poverty, living with unrelated adult roommates, Hispanic, male, and uninsured (Blumberg and Luke 2011). This makes it difficult to draw conclusions from the unweighted SHIPP data with regard to any absolute estimates, but relative comparisons should be valid since assignment to treatment was randomized. However, these relative comparisons could be compromised if the methods operate differently among subgroups that were particularly lacking in the overall sample. For example, one of the goals of the redesign was to encourage more accurate reporting of past coverage – particularly relatively short spells of coverage – by prompting respondents with the specific months for which no coverage had yet been reported. If respondents with short, intermittent spells of coverage were underrepresented in the SHIPP sample overall, then this potential advantage of the redesign would be statistically imperceptible.

The SHIPP data were also evaluated for face validity of person/month/plan-level data. Statistical analysis was limited by the small sample size, coverage bias, and the low prevalence of change in coverage status during the 15- to 17-month reference period. However, the patterns exhibited were informative. Among the nonelderly sample who were administered the new CPS ($n = 2,882$), the vast majority (80.3 percent) were insured throughout the entire reference period, 11.2 percent were uninsured throughout, and another 2.5 percent had coverage in January 2009 but lost it and never regained it by the end of the reference period. The remaining 6.1 percent ($n = 176$) began a spell of insurance during the reference period, and five of these individuals began two spells of

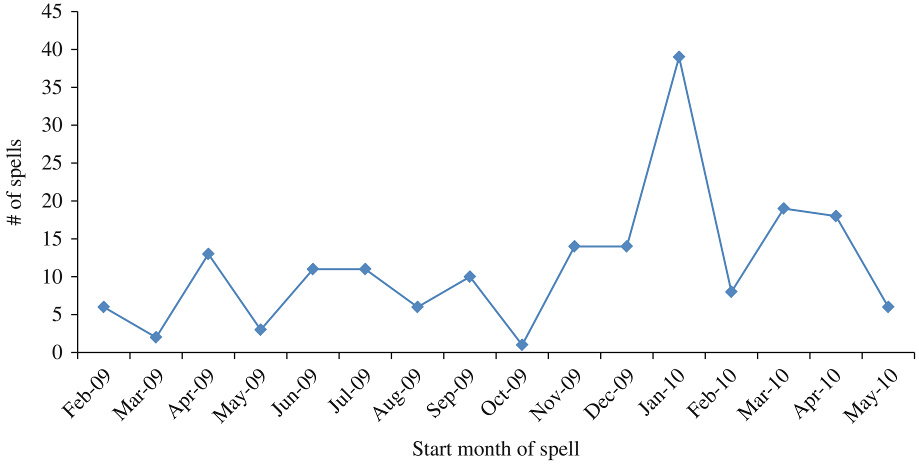


Fig. 2. SHIPP 2010 Start Month of Insured Spells (n = 181 spells). Source: 2010 Survey of Health Insurance and Program Participation

coverage during that time. Figure 2 plots the start month of these 181 spells. On average, just over eleven spells began in any given month. The range went from a low of one spell (beginning in October 2009) up to a high of 39 spells (beginning in January 2010) (Pascale 2011). The uptick in January 2010 may reflect actual behavior due to open enrollment, New Year's resolutions and/or other events tied to the start of the calendar year. This uptick notwithstanding, there does not appear to be evidence of seam bias or other obvious systematic biases in terms of reported start date of spells across the 15-month reference period.

The Medicare portion of the SHIPP survey data was matched back to the Medicare records from which that sample was drawn originally, and indicators of enrollment from the two data sources were compared. Results were favorable to the redesign in terms of both underreporting and overreporting, but for the latter results were particularly robust. The false-positive (overreporting) error rate in the old CPS was 6.9 percent and in the new CPS the error rate was 2.0% — a 4.9 percentage-point difference ($p = 0.01$) (Resnick 2013).

It was hoped that the new CPS would garner higher reports of past coverage than the old CPS. However, the new CPS did not lose ground compared to the status-quo design, and coverage bias could explain the lack of differences in the estimates. Furthermore, the person/month/plan-level data had face validity, and the Medicare match-back study favored the redesign. All these factors warranted a follow-up test.

4.3. CPS ASEC Content Test (2013)

In March 2013 the Census Bureau carried out the CPS ASEC 2013 Content Test to evaluate the CPS redesign in a production environment with a larger sample ($n = 29{,}629$) more representative of the CPS. The study was a comparison of estimates from the old and new CPS questionnaire. Estimates for the old CPS were derived by subsetting the CPS production sample to those interviewed by phone in March 2013. The new CPS was administered in parallel with the production CPS, by phone, to a sample of households that had already completed the final rotation of the CPS (this is known as "retired" sample).

Weights, multivariate modeling, and a separate analysis of a subset of the retired sample from the production side were all used to assess the effects of the nature of differences across samples. Results showed that the odds of coverage being reported for the past calendar year were higher under the new CPS than the old, and that within the new CPS, calendar-year estimates of coverage were higher than and distinct from point-in-time estimates. There were few statistically significant differences in coverage across demographic subgroups. See Pascale et al. (2015) for a thorough description of the weighting, methodology and results. These results demonstrated that the redesign did represent an improvement in the estimates, and there was no evidence that subgroups were unevenly affected by the design difference.

Further methodological analysis of the same dataset was conducted in an attempt to assess the impact of the individual features of the questionnaire that were changed. This was not a straightforward process, however, because all three questionnaire design features were modified at the same time in order to evaluate the final questionnaire as a whole. Two characteristics in particular, however, were tied to some of the modified questionnaire features: household size and “social distance” – the relationship between the household respondent and the person for whom he/she was reporting (aka self/proxy). The hybrid person-household-level design in the new CPS could reduce the chances that certain individuals are forgotten, especially in larger households, because it provides names of each household member. The new CPS structure of questions on coverage type (general to specific) allows respondents to provide basic information on whether other household members are covered or not (even if their knowledge of coverage type is limited), while the old CPS laundry list asks very detailed questions on coverage type but no general yes/no question on coverage. The more social distance, the less knowledge respondents may have on the details of coverage type for other household members. This could lead to more underreporting for more socially distant individuals in the old CPS than the new. While the reference-period changes could not be completely disentangled from the other two questionnaire features, comparing estimates from single-person households at least removes the effect of the household/person-level design for this subset. In both the old and the new CPS, the questions refer to only the household respondent (“you”) and no names are used (see Figure 3, upper panel). Thus, any observed differences in single-person households could be attributed to either the reference-period change, the change to the laundry-list approach, or some combination.

In Table 1, two distinct variables are examined. The first is a measure of household size, where households were categorized as small (single-person), medium (two to four people) or large (five or more people). The second is a measure of social distance, which categorizes individuals as having either a self- or proxy report of their health coverage. Within proxy reports, individuals were classified based on their relationship to the “reference person,” who is usually the household respondent or their spouse. The proxies were divided into two groups: the child of the reference person, or someone other than the child of the reference person. Chi-squared tests of association were used to test the statistical significance of bivariate comparisons of coverage across test and control panels. A logistic regression model was used to test within-treatment comparisons between small, medium, and large households and between self- and proxy reports. For all comparisons, a significance threshold of 0.05 was used.

	Old CPS	New CPS
Single-person households	1. At any time in 2009 were you covered by [plan type x]? 2. At any time in 2009 were you covered by [plan type y]? 3. [repeat for all eight plan types]	1. Do you NOW have any coverage? Yes ➔ [Qs to identify plan type] ➔ Q2 No ➔ next section 2. Did your coverage from [plan type x] start before or after January 1, 2009? Before ➔ Q3 After ➔ Q4 3. And has it been continuous since January 2009? Yes ➔ CK No ➔ Q4 4. In what month did that coverage start? ➔ Qs on gaps in coverage prior to start month ➔ CK CK: If single-person household ➔ next section; else ➔ Q5
Multi-person households	1. At any time during 2009 was anyone in this household covered by [plan type x]? Yes ➔ go to Q1a No ➔ go to Q2 1a. Who was covered? ➔ Q2 2. At any time during 2009 was anyone in this household covered by [plan type y]? Yes ➔ go to Q2a No ➔ go to Q3 2a. Who was covered? ➔ Q3 [repeat for all eight plan types]	[Same routine as in single-person households, with the addition of Q5-Q8]: 5. Is anyone else in this household also covered by [plan type x]? Yes ➔ Q6 No ➔ next section 6. Who is covered? ➔ Q7 7. Were they also covered in [months first person had the coverage]? Yes ➔ apply same months to this person as first person No ➔ Q8 8. What months between January 2009 and now was this person covered by [plan type x]?

Fig. 3. Simplified Question Flow by Household Size in Old versus New CPS

Regarding household size, Table 1 shows that among single-person and large households, there was no difference between the old and new CPS insured estimate. However, in medium households, the new CPS insured rate was 2.4 percentage points higher than in the old CPS design ($p = 0.004$). Within the new CPS there was no difference in the estimates between small and medium households, but within the old CPS the insured estimate in small households was 2.7 percentage points higher than in medium households ($p = 0.021$). In both the old and new CPS, reporting of coverage was much lower in large than in small households, by almost eight percentage points. Reporting of coverage was also lower in large households compared to medium households in both treatments ($p < 0.001$ in the new CPS; $p = 0.006$ in the old CPS).

Turning to self/proxy results, Table 1 shows there was no treatment difference in the insured estimate for self-responses, but for proxy responses overall the new CPS insured rate was two percentage points higher than the old CPS ($p = 0.041$). Among proxies, there was no treatment difference in reporting of coverage for the child of the reference person, but proxy reporting of coverage for any other household member was 3.6 percentage points higher in the new versus old CPS ($p = 0.003$). Within the new CPS, there was no difference between self and proxy reporting, whether the proxy was the child of the reference person or someone else in the household. But, within the old CPS the estimate for self-reports was 2.5 percentage points higher than for proxy reports overall ($p < 0.001$). Furthermore, among proxies, the old CPS reporting was

Table 1. Weighted Old versus New CPS Insured Estimates, by Household Size and Self/Proxy Status, 2013 CSP ASEC Content Test

		New CPS Insured		Old CPS Insured		
		N = 243,474,924	% (SE)	N = 243,474,924	% (SE)	New-Old % insured (p)
Household Size	**Small**	23,878,488	90.74	27,571,316	91.24	−0.50
			(1.0152)		(0.9246)	(0.7136)
	Medium	155,833,875	90.96	150,507,500	88.57	**2.39**
			(0.5383)		(0.5784)	**(0.0043)**
	Large	37,932,138	82.76	36,233,716	83.62	−0.87
			(1.7761)		(1.8462)	(0.7532)
Self/Proxy	**Self**	86,334,860	89.98	88,093,242	89.52	0.45
			(0.5656)		(0.5553)	(0.5676)
	Proxy	131,309,641	89.01	126,219,290	87.00	**2.00**
			(0.6735)		(0.6535)	**(0.0408)**
	Child	66,524,680	89.72	63,078,540	89.50	0.22
			(0.9987)		(0.8721)	(0.8743)
	Other	64,784,961	88.29	63,140,750	84.65	**3.64**
			(0.7498)		(0.8696)	**(0.0026)**

SE are clustered on household. Unweighted n = 29,629 (new CPS sample n = 16,401; old CPS sample n = 13,228). Weighted totals do not sum to population estimates due to the sample design of the 2013 content test and drop-out rates within the ASEC portion of the questionnaire. See Pascale et al. (2015) for more details.
Source: 2013 CPS-ASEC Content Test.

4.9 percentage points higher for child of the reference person than others in the household ($p < 0.001$). These results suggest that the new CPS closed the social-distance gap that existed in the old CPS.

In sum, if we assume higher reporting is more accurate reporting, the old and new CPS both do equally well in single-person households and when the respondent has to report only for him/herself and/or his/her child. Given the large gap in reporting between small and large households, one could also say both treatments do equally poorly once the household gets to be five or more people in size. However, it is unknown whether coverage levels are simply lower in larger households, or whether both treatments are missing reports of coverage for some individuals in large households. There is some evidence to support the former. For example, Hispanics as a group have higher uninsured rates than the overall population (19.9 percent versus 10.4 percent) (US Census Bureau 2014). And where the householder is of Hispanic origin the household size is larger, on average, than the national average (3.54 versus 2.59 people per household) (US Census Bureau 2012).

Where the new CPS demonstrates significant, measurable improvements is in medium-sized households and with proxy reports for individuals who are the most socially distant from the respondent. To estimate how many additional people were reported as insured under the new CPS solely due to its different impact on these two subgroups, a rough calculation is offered using the 2013 content test data. For reasons described elsewhere (Pascale et al. 2015), the 2013 test sample was weighted to a total somewhat lower than the U.S. population – about 243 million – and it is on that base population that these calculations are offered. Individuals in medium-sized households constituted 70 percent of the sample, those in the proxy/other category constituted 30 percent of the sample, and 25 percent of the total sample ($n = 60{,}868{,}731$) was in both groups. The new-old CPS difference in the insured estimate was 2.4 percentage points for individuals in medium households, and 3.6 percentage points for proxy/other individuals. To be conservative, if we allocate all of the 25-percent overlapping sample to medium households (where the new-old CPS difference is lower), the calculation would be:

Medium-sized households in the old CPS

- 169,925,637 individuals * 2.39 ppt new-old CPS difference ⟶ 4,061,223

Proxy/Other individuals in the old CPS

- 74,590,668 individuals – 60,868,731 (the 25% overlapping sample) = 13,721,937
- 13,721,937 individuals * 3.64 ppt new-old CPS difference ⟶ 499,479

Based on these rough calculations, a total of about four and a half million individuals reported as uninsured in the old CPS would have been reported as insured in the new CPS. Taking this a step further, if we add this four and a half million to the total reported as insured under the old CPS, the new hypothetical insured total comes up to 218,873,234, or 89.9 percent of the total old CPS population. This compares to the insured rate of 88.0 percent under the old CPS and represents a 1.9 percentage-point difference. This difference is in line with findings from the 2013 content test, which showed a difference of 1.4 percentage points in the uninsured rate between the old and new CPS. It is also in line with a comparison of the 2013 and 2014 production estimates, which employed the old and new CPS, respectively. While there were many caveats to that comparison, a 1.6 percentage-point difference was approximated to be attributable to the change in questionnaire design (Pascale et al. 2015).

4.4. Questionnaire Administration Time in the Old Versus New CPS

Timers were built into both the 2010 and 2013 tests. In 2010, the old CPS health insurance module took two minutes and 25 seconds to administer, while the new CPS took on average three minutes 56 seconds – about a minute and a half longer. The median measure showed a gap of about one minute between the old and new CPS. The 2013 results were almost identical, indicating that the new CPS took 1.5 minutes longer than the old CPS health insurance module (Bee and Cantu 2013). While the added duration in the new CPS is concerning, there are some important mitigating factors. For example, the average SHIPP interviewer had 8.39 years of field interviewing experience at the Census Bureau, and 5.41 years of experience on surveys that include questions about health insurance (the old CPS being among the most common of these). For the new CPS, interviewers only had the benefit of a two-day training and a ten-day field period. As interviewers become more familiar with the new CPS – which is a tremendous departure from the old design – the learning curve benefits should thus develop and reduce administration time. Other important considerations are the benefits of the redesign discussed below.

5. Discussion

When this research began in 1999, the objective was to identify key questionnaire design features that were associated with measurement error, develop improvements, and test their effectiveness. The 2013 test results indicated that the new CPS did reduce measurement error in estimates, across demographic subgroups, resulting in a lower estimate of the uninsured. Further analysis of the 2013 test, focusing on household size and self/proxy reporting, shed some light on the reasons for the observed improved reporting. While long-standing speculation suggested that recall error induced by the reference period was the main contributor to reporting error, results indicate that much more than recall error was at play. If recall were the only factor, then we would expect improved reporting under the new CPS even in single-person households, but we see no differences there.

What seems to have been driving lower reporting in the old CPS was not primarily recall error but a confluence of factors related to household size and composition, and cognitive difficulty of the reporting task. Consider that when asked in the old CPS, "At any time in 2012 did anyone in the household have [Plan Type X]?," respondents who fail to report coverage are not saying "no" – that an individual household member does NOT have that coverage type – they are just not saying "yes." Furthermore, they are failing to say "yes" to two different kinds of questions. First, in cases where the coverage type specified in the household-level question is unfamiliar or obscure to them (e.g., Medicaid and its state-specific program names, Tricare and other military plans) they may not say "yes" simply because they do not recognize the coverage type. Second, if they do answer "yes" to the household-level question, they then have to actively report individual household members who have that coverage type, and do not have the benefit of being asked about those individuals by name. In these cases, the more socially distant household members may not come to mind as readily. For example, a respondent reporting ESI coverage may be focused mostly on their own family members who are dependents on their particular plan, and less focused on other, more socially distant household members (e.g., a cousin, boarder), who may have their own ESI plans. Indeed, social distance has been found to affect data quality in other contexts as well. For example, Grieco and Armstrong (2014) reported that social distance between the respondent and others in the household was likely a factor in item nonresponse to questions on year of naturalization.

Social distance, however, is only part of the reporting problem. Results indicate that when the reporting task becomes particularly challenging – calling to mind several other household members at once, some of whom are more socially distant, AND thinking back over 15 months, sometimes about obscure coverage types – the old CPS design is less equipped than the new design to prompt reporting. These results hearken back to the 1999 study on reference period and household versus person-level design. That study suggested that respondents had difficulty when asked about both the past calendar year *and* all household members at once. They had less difficulty when either one person at a time was asked about (by name) for the calendar year, or when only the current time period was asked about for "anyone in the household" (with no names). All three revised questionnaire features in the new CPS play a role in simplifying the reporting task and collectively reducing the cognitive burden on the respondent. The series asks about only one person at a time, and it starts with a simple yes/no question on current coverage status – not specific coverage type. Furthermore, the series starts with questions about not just any household member but the respondent him/herself. This allows the respondent to focus on what are likely the more difficult aspects of the questions – coverage type and months of coverage – under (arguably) the simplest circumstances first, without being burdened with *also* thinking about those same aspects for other people over a 15-month time period.

Apart from addressing measurement error, the new CPS renders all the same data as the old CPS but with much more detailed information on monthly coverage. When health insurance questions were first added to the CPS in the 1980s, the strategy was to be consistent with the unit of measurement used for income questions. Since then the landscape of health insurance has changed, which raises the question: what kind of definition of the uninsured "should" be used? That is, what degree of "uninsured" makes most sense as an analytic category? In the old CPS, the only information collected was whether individuals had coverage "at any time" during the previous calendar year. Thus, the only way to define the uninsured was those without coverage throughout the entire year because that was the only data rendered by the questions. As such, one day of coverage could separate the insured from the uninsured. This is a rather dubious distinction, since a person who was uninsured for, say, eleven out of twelve months, or who had spells of noncoverage, is likely to have a profile more like a person uninsured throughout the year than someone who was insured for the entire twelve months. Indeed, according to the Congressional Budget Office, "Policies aimed at increasing coverage are most likely to be effective if they consider the distinction between the short-term and long-term uninsured." (Congressional Budget Office 2003, viii).

The original redesign strategy of asking about current coverage, framing the 15-month time period and, in some cases, prompting respondents with month-level questions on coverage, was all in the service of improving reports of retrospective coverage. However, as a side benefit, the questionnaire renders person/month/plan-level data, from the beginning of the previous calendar year all the way through and including the interview day. Thus, the redesign enhances the value of the dataset by enabling a more thorough analysis of the dynamics of coverage over a continuous 15-month reference period.

For example, analysts can examine the number and duration of spells, and the specific months of coverage, which allows for studies on topics such as seasonality and the effects of major external events, such as a recession. The data also enable analysis of churning within plan type (e.g., people who are on and off Medicaid), transitions from one plan type to another, and the direction of those transitions (e.g., from Medicaid to ESI, or vice versa). And while the traditional CPS does allow for the reporting of multiple plan types, because it only asks about coverage "at any time" during the past year it is impossible to know the duration of coverage, and whether individuals were covered by multiple plan types at the same time or transitioned from one plan type to another. Having the data to study source of coverage and transitions will likely become more valuable in the coming years. As the uninsured rate drops, the research community focus may shift from measuring the uninsured to exploring how individuals obtain their coverage, and the dynamics of coverage – when and why they shift from one source to another.

Like all studies that aim to measure the uninsured, definitive conclusions regarding measurement error are saddled by lack of a truth source. The U.S. health care system, both pre- and postreform, is a patchwork of private and public sources of coverage and no centralized database on individuals with and without health coverage exists. However, some limited conclusions can be drawn about the data quality of the new CPS. First, studies on the old CPS indicated the calendar-year estimate of the uninsured was in line with estimates of those uninsured at a point in time. In the new CPS, the finding that the estimate of calendar-year coverage is higher than and distinct from the estimate of current coverage represents an improved metric, at least in the relative sense. Second, the new CPS appears to be less prone to measurement error in relation to household size and social distance. Third, to the extent that relevance is a measure of data quality, the improved precision and detail in terms of person/month/plan-type-level variables, and the additional ACA-specific measures in the new CPS, represents a marked improvement over the old CPS.

The next steps in this line of research include continued study of public coverage reporting. The new CPS was designed to address the chronic and persistent underreporting of public coverage, but appears not to have made any gains in that area. Analysis of the 2013 test results suggests this could be a result of multiple factors all revolving around the theme of overreporting (Pascale et al. 2015). Results from the 2010 test showed that reporting of both public and private coverage was more than double in the old CPS than it was in the new CPS (Boudreaux et al. 2013), and that Medicare overreporting was higher in the old than in the new CPS (Resnick 2013). Furthermore, although the Medicaid undercount is well documented (Klerman et al. 2005; Blumberg and Cynamon 1999; Czajka and Lewis 1999; Lewis et al. 1998), there is some evidence of Medicaid overreporting (Klerman et al. 2009; Davern et al. 2008). The latter study showed that roughly 20 percent of the self-reported Medicaid estimate was comprised of enrollees in private coverage who also reported Medicaid. Finally, the dearth of evidence on overreporting is not indicative of a lack of overreporting; rather, it is much more difficult to measure due to state-level variation and the absence of a truly comprehensive dataset of enrollees.

Another key area for more research is validation. The accuracy of the uninsured estimate is intertwined with the accuracy of reporting on individual plan types. Medicaid reporting has received substantial study and attention, in part due to the existence and accessibility of fairly high-quality records. Yet even within the Medicaid reporting literature it is not entirely clear how misreporting of Medicaid affects estimates of other plan types, and the ultimate measure of the uninsured, at the national level. The accuracy of reporting of private plans has received less rigorous study, in part due to less accessible, more disparate sources of validation data. Finally, only two studies to date (Davern et al. 2008 and Nelson et al. 2003) have examined reporting accuracy of multiple sources of coverage – both public and private – within the same questionnaire.

To address this gap, the Census Bureau collaborated with other agencies to conduct a study in 2015 called CHIME (Comparing Health Insurance Measurement Error) that compared data from enrollment records across multiple markets – including ESI, Medicaid and the new marketplace plans – to survey reports from health insurance modules from the CPS ASEC redesign and the ACS. Analysis will include an assessment of both "absolute" reporting accuracy (the extent to which the survey data matches the record data) and "relative" reporting accuracy (comparing absolute reporting accuracy across questionnaire treatments) of particular plan types, both private and public.

6. Implications Beyond Health Insurance Measurement

These results have broad implications for both questionnaire redesign strategies and for topic areas beyond health insurance. Many redesign efforts involve a long-term plan whereby each phase of the research and the schedule is set years in advance. This often means the timeline is insufficient to fully digest the results of one study in order for the findings to be fed in to the design of the next study. More important, it can often mean that results from an early study raise questions that warrant a subsequent test of a certain nature and scope, but a different type of test was predefined in the research plan. Some degree of planning and scheduling is obviously necessary for staffing and budgeting purposes. However, if considerable flexibility is built in to the research plan, the value of each test can be enhanced with sufficient time for analysis, and each test can be adjusted to address the emerging research questions.

Results also provide guidance for questionnaire design, as the specific features found to be problematic are not unique to health insurance questions. The overarching finding is that if a reporting task is particularly demanding, it can pay to decompose complex questions into their simpler parts. Knowing what constitutes "particularly demanding" for all respondents is certainly not possible. However, results provide some guidance on how to design questions so that reporting under the more challenging conditions does not suffer relative to the simpler conditions. Numerous surveys ask retrospective questions using wording like "At any time during [time period X] did [you/anyone] receive [subject Y]?" These surveys may benefit by decomposing the task and asking first about the present (i.e.,: "Do you receive [subject Y]?"), which allows respondents to focus on subject Y without the compound tasks of thinking about duration and any change over time within the same question. In terms of subject matter, many topic areas lend themselves to being asked at a general level (e.g.,: retirement plans) or a specific level (e.g., 401(k), 403(b), IRA, Roth IRA). Results suggest asking about the general level first and then drilling down to the specifics, particularly when a single respondent is answering questions about all household members. Decomposing complex questions has been demonstrated to result in higher data quality in contexts beyond health coverage (Loftus et al. 1990; Redline 2013; Fowler 2004). Note, however, that the evidence is somewhat more mixed when the measures in question relate to frequency of behaviors (Beatty 2010; Belli et al. 2000; Schaeffer and Presser 2003).

Finally, to reduce the risk of underreporting by providing individual names, while also reducing burden, results suggest that a hybrid person-household-level approach can be effective. This could be especially relevant when asking about a topic area that may be shared across household members, such as jointly owned assets, or household-level receipt of public benefits.

7. References

Agency for Healthcare Research and Quality. 2015. "Medical Expenditure Panel Survey, MEPS Background." Available at: http://meps.ahrq.gov/mepsweb/communication/household_participant_back.jsp (accessed on October 5, 2015).

Beatty, P. 2010. "Considerations Regarding the Use of Global Survey Questions." Paper presented at the Consumer Expenditures Survey Methods Workshop, Hyattsville, MD, December 8–9, 2010. Available at: http://www.bls.gov/cex/methwrkshp_pap_beatty.pdf (accessed on February 5, 2016).

Beatty, P. and S. Schechter. 1998. "Questionnaire Evaluation and Testing in Support of the Behavioral Risk Factor Surveillance System (BRFSS), 1992–98." Office of Research and Methodology, NCHS, Working paper series, 26: 12–17.

Bee, C.A. and A. Cantu. 2013. "Using Timer Data to Evaluate the Respondent Burden of the 2013 CPS ASEC Content Test." Proceedings of the Federal Committee on Statistical Methodology Research Conference, November 4–6, 2013, Washington, DC. Available at: http://fcsm.sites.usa.gov/files/2014/07/H3_Bee_2013FCSM.pdf (accessed on February 5, 2016).

Belli, R.F., N. Schwarz, E. Singer, and J. Talarico. 2000. "Decomposition Can Harm the Accuracy of Behavioural Frequency Reports." *Applied Cognitive Psychology* 14: 295–308.

Bhandari, S. 2004. "People with Health Insurance: A Comparison of Estimates from Two Surveys." Survey of Income and Program Participation (SIPP) Working Paper No. 243. Washington, D.C.: U.S. Census Bureau. Available at: https://www.census.gov/sipp/workpapr/wp243.pdf (accessed on February 2, 2016).

Blair, E.A. and G.K. Ganesh. 1991. "Characteristics of Interval-Based Estimates of Autobiographical Frequencies." *Applied Cognitive Psychology* 5: 237–250. Doi: http://dx.doi.org/10.1002/acp.2350050306.

Blewett, L.A. and M.E. Davern. 2006. "Meeting the Need for State-Level Estimates of Health Insurance Coverage: Use of State and Federal Survey Data." *Health Services Research* 41: 946–975. Doi: http://dx.doi.org/10.1111/j.1475-6773.2006.00543.x.

Blumberg, S. 2014. "National Health Interview Survey Since 1957." Federal Statistics on Health Insurance Coverage: Technical Meeting on Methods Used in Household Surveys, August 18, 2014, Washington, DC. Available at: http://www.census.gov/newsroom/releases/pdf/20140818_gw_final.pdf (accessed on February 5, 2016).

Blumberg, S.J. and M.L. Cynamon. 1999. "Misreporting Medicaid Enrollment: Results of Three Studies Linking Telephone Surveys to State Administrative Records." Proceedings of the Seventh Conference on Health Survey Research Methods. 189–195. Available at: http://www.cdc.gov/nchs/data/hsrmc/hsrmc_7th_proceedings_1999.pdf. (accessed on February 5, 2016).

Blumberg, S.J. and J.V. Luke. 2011. "Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July-December 2010." National Center for Health Statistics. Available at: (http://www.cdc.gov/nchs/nhis.htm) (accessed on February 5, 2016).

Blumberg, S.J., L. Osborn, J.V. Luke, L. Olson, and M.R. Frankel. 2004. Estimating the Prevalence of Uninsured Children: An Evaluation of Data from the National Survey of Children with Special Health Care Needs, 2001. Vital Health Statistics 2 (136). Hyattsville, MD: National Center for Health Statistics.

Boudreaux, M.H., B. Fried, J. Turner, and K.T. Call. 2013. "SHADAC Analysis of the Survey of Health Insurance and Program Participation." State Health Assistance Data Center. Available at: http://www.shadac.org/files/shadac/publications/SHIPP_final_report.pdf (accessed on August 4, 2014).

Congressional Budget Office. 2003. "How Many People Lack Health Insurance and for How Long?" A CBO Report. The Congress of the United States. Available at: https://www.cbo.gov/publication/14426 (accessed on February 5, 2016).

Crespi, I. and J.W. Swineheart. 1982. "Some Effects of Sequenced Questions Using Different Time Intervals on Behavioral Self-Reports: A Field Experiment." Paper presented at the American Association for Public Opinion Research. Hunt Valley, MD, May 1982.

Czajka, J. and K. Lewis. 1999. "Using Universal Survey Data to Analyze Children's Health Insurance Coverage: An Assessment of Issues." Washington, DC: Mathematica Policy Research, Inc.

Davern, M. 2009. "Unstable Ground: Comparing Income, Poverty & Health Insurance Estimates from Major National Surveys." Paper presented at the Academy Health Annual Research Meeting, June 29, 2009, Chicago.

Davern, M., K.T. Call, J. Ziegenfuss, G. Davidson, T. Beebe, and L. Blewett. 2008. "Validating Health Insurance Coverage Survey Estimates: A Comparison of Self-Reported Coverage and Administrative Data Records." *Public Opinion Quarterly* 72: 241–259.

Farley-Short, P. 2001. "Counting and Characterizing the Uninsured." University of Michigan, Ann Arbor, MI: Economic Research Initiative on the Uninsured Working Paper Series. Available at: http://www.umich.edu/~eriu/pdf/wp2.pdf. (accessed on August 25, 2005).

Fowler, F.J. 2004. "The Case for More Split-Sample Experiments in Developing Survey Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer. 173–188. Hoboken, NJ: Wiley.

Grieco, E.M. and D.M. Armstrong. 2014. "Assessing the 'Year of Naturalization' Data in the American Community Survey: Characteristics of Naturalized Foreign Born Who Report – and Don't Report – the Year They Obtained Citizenship." Paper presented at the Applied Demography Conference, April 30 – May 3, 2014, San Antonio, TX.

Hess, J., J. Moore, J. Pascale, J. Rothgeb, and C. Keeley. 2001. "The Effects of Person-level vs. Household-level Questionnaire Design on Survey Estimates and Data Quality." *Public Opinion Quarterly* 65: 574–584.

Hess, J. and E. Singer. 1995. "The Role of Respondent Debriefing Questions in Questionnaire Development". In Proceedings of the Section on Survey Research Methods, August 13–17, 1995. 1075–1080. Washington, DC: American Statistical Association. Available at: http://www.amstat.org/sections/srms/proceedings/papers/1995_187.pdf (accessed on February 5, 2016).

Klerman, J.A., M. Davern, K.T. Call, V. Lynch, and J. Ringel. 2009. "Understanding the Current Population Survey's Insurance Estimates and the Medicaid 'Undercount'" *Health Affairs* web exclusive. Doi: http://10.1377/hlthaff.28.6.w991.

Klerman, J.A., J.S. Ringel, and B. Roth. 2005. "Under-Reporting of Medicaid and Welfare in the Current Population Survey". RAND Working Paper WR-169-3. Santa Monica, CA: RAND.

Lewis, K., M. Ellwood, and J. Czajka. 1998. *Counting the Uninsured: A Review of the Literature*. Washington, DC: The Urban Institute.

Loftus, E.F., M.R. Klinger, K.D. Smith, and J. Fiedler. 1990. "A Tale of Two Questions: Benefits of Asking More Than One Question." *Public Opinion Quarterly* 54: 330–345.

Loomis, L. 2000. "Report on Cognitive Interview Research Results for Questions on Welfare Reform Benefits and Government Health Insurance for the March 2001 Income Supplement to the CPS." Washington, DC: Center for Survey Methods Research, Statistical Research Division, U.S. Census Bureau.

Lynch, V. 2006. "Causes of Error in Survey Reports About Who in the Household Gets Welfare." Unpublished Paper, Joint Program in Survey Methodology. College Park, MD.

Martin, E.A., R.E. Fay, and E.A. Krejsa. 2002. "Analysis of Questionnaire Errors in Survey Measurements of Census Coverage." In Proceedings of the American Statistical Association Survey Research Methods Section, August 10–15, New York, NY. 2260–2265. Alexandria: VA American Statistical Association.

National Research Council. 2006. *Food Insecurity and Hunger in the United States: An Assessment of the Measure. Panel to Review the U.S. Department of Agriculture's Measurement of Food Insecurity and Hunger*, edited by G.S. Wunderlich and J.L. Norwood. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nelson D.E., E. Powell-Griner, M. Town, M.G. Kovar. 2003. A Comparison of National Estimates From the National Health Interview Survey and the Behavioral Risk Factor Surveillance System. *American Journal of Public Health* 93:1335–1341.

Office of the Assistant Secretary for Planning and Evaluation, Health and Human Services. 2005. "Understanding Estimates of the Uninsured: Putting the Differences in Context." Updated September 5. Available at: http://aspe.hhs.gov/health/reports/05/uninsured-understanding-ib/#estimates (accessed on April 9, 2016).

Pascale, J. 2001a. "Methodological Issues in Measuring the Uninsured." Proceedings of the Seventh Health Survey Research Methods Conference, Williamsburg, VA. 167–173. Available at: http://www.cdc.gov/nchs/data/hsrmc/hsrmc_7th_proceedings_1999.pdf (accessed on February 5, 2016).

Pascale, J. 2001b. “The Role of Questionnaire Design in Medicaid Estimates: Results from an Experiment.” Talk presented to the Washington Statistical Society, March 21, 2001.

Pascale, J. 2004. “Medicaid and Medicare Reporting in Surveys: An Experiment on Order Effects and Program Definitions.” Proceedings of the American Association for Public Opinion Research, American Statistical Association, May 13–16, 2004. pp. 4976–4983. Available at: http://www.amstat.org/sections/SRMS/Proceedings/ (accessed on April 2, 2015).

Pascale, J. 2008. “Measurement Error in Health Insurance Reporting.” *Inquiry* 45(4): 422–437. doi: 10.5034/inquiryjrnl_45.04.422 Available at: http://inq.sagepub.com/content/45/4/422.full.pdf+html (accessed on February 4, 2015).

Pascale, J. 2009a. “Findings from a Pretest of a New Approach to Measuring Health Insurance in the Current Population Survey.” Paper prepared for the Federal Committee on Statistical Methodology Research Conference, November 2–4, 2009. Available at: https://fcsm.sites.usa.gov/files/2014/05/2009FCSM_Pascale_VIII-A.pdf (accessed on February 5, 2016).

Pascale, J. 2009b. “Survey Measurement of Health Insurance Coverage: Cognitive Testing Results of Experimental Questions on Integrated Current and Calendar Year Coverage.” Available through “Q-BANK,” a database of questions and reports on cognitive testing maintained by the National Center for Health Statistics. Available at: http://wwwn.cdc.gov/qbank/report/Pascale_Census_2009_HealthInsurance.pdf (accessed on February 5, 2016).

Pascale, J. 2009c. “Health Insurance Measurement: A Synthesis of Cognitive Testing Results.” Paper presented at the Questionnaire Evaluation Standards (QUEST) meeting, May 18–20, 2009, Bergen, Norway.

Pascale, J. 2015. “Moderniziing a Major Federal Government Survey: A Review of the Redesign of the Current Population Survey Health Insurance Questions. Study Series, Survey Methodology #2015–03. Available at: https://www.census.gov/srd/papers/pdf/SSM2015-03.pdf (accessed April 2016).

Pascale, J., M. Boudreaux, and R. King. 2015. “Understanding the New Current Population Survey Health Insurance Questions.” *Health Services Research* 51: 240–261. Doi: http://dx.doi.org/10.1111/1475-6773.12312.

Pascale, J., J. Rodean, J. Leeman, C. Cosenza, and A. Schoua-Glusberg. 2013. “Preparing to Measure Health Coverage in Federal Surveys Post-Reform: Lessons from Massachusetts.” *Inquiry: The Journal of Health Care Organization, Provision, and Financing* 50: 106–123. Doi: http://dx.doi.org/10.1177/0046958013513679.

Pascale, J. 2011. “Findings from a Split-Ballot Experiment on a New Approach to Measuring Health Insurance in the Current Population Survey.” Report prepared for the Center for Survey Measurement, US Census Bureau.

Pascale, J., M.I. Roemer, and D.M. Resnick. 2009. “Medicaid Underreporting in the CPS: Results from a Record Check Study.” *Public Opinion Quarterly* 73: 497–520. Available at http://inq.sagepub.com/content/45/4/422.full.pdf+html (accessed on April 2, 2015).

Redline, C. 2013. “Clarifying Categorical Concepts in a Web Survey.” In “Topics in Survey Measurement and Public Opinion.” Special issue, *Public Opinion Quarterly* 77: 89–105.

Research Project to Understand the Medicaid Undercount. 2008. “Phase II Research Results: Examining Discrepancies between the National Medicaid Statistical Information System (MSIS) and the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC).” Available at: https://www.census.gov/did/www/snacc/docs/SNACC_Phase_II_Full_Report.pdf (accessed on October 5, 2015).

Resnick, D. 2013. “Microsimulation Support for Tax, Transfer & Health Insurance Policy Analysis Summary.” Submitted to David S. Johnson & Charles Nelson, U.S. Department of Commerce, U.S. Census Bureau, and Don Oellerich, U.S. Department of Health and Human Services, ASPE, on October 3, 2013.

Resnick, D., S. Love, C. Taeuber, and J. Staveley. 2004. “Analysis of ACS Food Stamp Program Participation Underestimate.” Paper presented at the 2004 Joint Statistical Meeting, August 8–12, 2004, Toronto.

Ringel, J.S. and J.A. Klerman. 2005. “Today or Last Year? How Do Interviewees Answer the CPS Health Insurance Questions?” RAND Labor and Population working paper series WR-288. Santa Monica, CA: RAND.

Roman, A.M., A. Hauser, and A. Lischko. 2002. “Measurement of the Insured Population: The Massachusetts Experience.” Paper presented at the 2002 Annual Meetings of the American Association for Public Opinion Research, May 16–19, 2002, St. Pete’s Beach, FL.

Rosenbach, M. and K. Lewis. 1998. “Estimates of Health Insurance Coverage in the Community Tracking Study and the Current Population Survey.” Document no. PR98-54. Washington, DC: Mathematica Policy Research, Inc.

Schaeffer, N.C. and S. Presser. 2003. “The Science of Asking Questions.” *Annual Review of Sociology* 29: 65–88. Doi: http://dx.doi.org/10.1146/annurev.soc.29.110702.110112.

Sudman, S., A. Fin, and L. Lannon. 1984. “The Use of Bounded Recall Procedures in Single Interviews.” *Public Opinion Quarterly* 48: 520–524.

Swartz, K. 1986. “Interpreting the Estimates from Four National Surveys of the Number of People Without Health Insurance.” *Journal of Economic and Social Measurement* 14: 233–242.

State Health Access Data Assistance Center (SHADAC). 2013. *Comparing Federal Government Surveys that Count the Uninsured*. State Health Access Data Center.

U.S. Census Bureau. 2012. “America’s Families and Living Arrangements: 2012: Average Number of People: AVG1. Average Number of People per Household, by Race and Hispanic Origin/1, Marital Status, Age, and Education of Householder: 2010.” Available at: https://www.census.gov/hhes/families/data/cps2012AVG.html (accessed on October 5, 2015).

U.S. Census Bureau. 2014. “Health Insurance in the United States: 2014 (P60-253), Detailed Tables, Table HI01. Health Insurance Coverage Status and Type of Coverage by Selected Characteristics: 2014.” Available at: http://www.census.gov/hhes/www/hlthins/data/index.html (accessed on October 5, 2015).

U.S. Census Bureau. 2015a. “Measuring Health Insurance Coverage with the Current Population Survey: A History of Improvement.” Available at: http://www.census.gov/content/dam/Census/library/infographics/measuring_health_insurance.pdf (accessed on October 2, 2015).

U.S. Census Bureau. 2015b. "About Health Insurance." Available at: https://www.census.gov/hhes/www/hlthins/about/ (accessed on October 5, 2015).

Willson, S. 2005. *Cognitive Interviewing Evaluation of the National Immunization Survey Insurance Module: Results of Fieldwork and Laboratory Interviews*. Unpublished report. Hyattsville, MD. National Center for Health Statistics.

Received September 2014
Revised February 2016
Accepted February 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 487–505, http://dx.doi.org/10.1515/JOS-2016-0025

Misspecification Effects in the Analysis of Panel Data

Marcel de Toledo Vieira[1], *Peter W.F. Smith*[2], *and Maria de Fátima Salgueiro*[3]

Misspecification effects (*meffs*) measure the effect on the sampling variance of an estimator of incorrect specification of both the sampling scheme and the model considered. We assess the effect of various features of complex sampling schemes on the inferences drawn from models for panel data using *meffs*. Many longitudinal social survey designs employ multistage sampling, leading to some clustering, which tends to lead to *meffs* greater than unity. An empirical study using data from the British Household Panel Survey is conducted, and a simulation study is performed. Our results suggest that clustering impacts are stronger for longitudinal studies than for cross-sectional studies, and that *meffs* for the regression coefficients increase with the number of waves analysed. Hence, estimated standard errors in the analysis of panel data can be misleading if any clustering is ignored.

Key words: Longitudinal survey; sampling variance; multistage sampling; stratification; weighting.

1. Introduction

Interest in fitting models to longitudinal complex survey data has grown in the last decade. Longitudinal surveys often make use of complex sampling procedures, such as unequal selection probabilities, stratification and multistage sampling, to select the initial panel sample at the first wave in order to best use the available resources (e.g., Smith et al. 2009). Nevertheless, to our knowledge, insufficient attention is still paid to the impacts of sampling complexities on the regression analysis of panel data in the survey-sampling literature.

Researchers and other users of panel data often make use of standard statistical techniques, which in most of the cases do not take account of the complex sample designs. These techniques may assume that the data are (after conditioning on some covariates) realizations of independent and identically distributed random vectors, which is rare in practice. The standard formulation of inference methods is often not valid when analysing data collected using a complex sampling scheme. According to Chambers and Skinner

[1] Departamento de Estatística e Programa de Pós-Graduação em Economia, Universidade Federal de Juiz de Fora (UFJF), Rua José Lourenço Kelmer, s/n, Campus Universitário, Bairro São Pedro, 36036-900, Juiz de Fora, MG, Brazil. Email: marcel.vieira@ice.ufjf.br
[2] Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, Southampton, SO17 1BJ, United Kingdom. Email: P.W.Smith@soton.ac.uk
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Av. Forças Armadas, 1649-026, Lisbon, Portugal. Email: fatima.salgueiro@iscte.pt

Acknowledgments: This work was supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais [grant numbers CEX-APQ-01799-10, CEX-APQ-02071-12 to the first author]; and the Fundação para a Ciência e a Tecnologia [grant number PTDC/GES/72784/2006 to the third author]. Visits to Southampton were partially supported by the Southampton Statistical Sciences Research Institute for the first and third authors.

(2003), even when the sampling design is considered ignorable, the standard inferential procedures may not satisfactorily reproduce the population complexities underlying the sampling mechanism. For a discussion on design-based and model-based methods for estimating model parameters under both ignorable and nonignorable sampling designs, see also Binder and Roberts (2003).

Moreover, complex sampling schemes may induce a correlation structure among observations, as elements in the same cluster are likely to be more similar than elements in different clusters. Therefore, when a sample is selected by complex sampling at Wave 1, a correlation structure among the observations, additional to the longitudinal correlation, may be induced. Under this situation, the use of standard statistical techniques with complex sampling data may lead to seriously biased point and standard-error estimates (see e.g., Nathan and Holt 1980). Ignoring clustering and weighting effects, for example, tends to lead to the underestimation of standard errors, and therefore to narrowed confidence intervals and to the incorrect rejection of null hypotheses. Stratification normally affects the analysis in an opposite direction. Thus ignoring clustering, weighting and stratification effects may lead to inappropriate statistical inference.

There is a well substantiated literature on methods for taking account of complex sampling schemes in the analysis of survey data. Skinner et al. (1989), Chambers and Skinner (2003), and Pfeffermann (2011), for example, provide further information and references. For cross-sectional data, Kish and Frankel (1974), Holt and Scott (1981), Scott and Holt (1982), Skinner (1986, 1989a, b), and Feder (2011), for example, have considered the effects of complex sampling on regression model parameters estimation.

Furthermore, Feder et al. (2000) proposed combining multilevel modelling, time-series modelling and survey-sampling methods for panel data analysis; Sutradhar and Kovacevic (2000) developed a generalised estimating equations approach by considering an autocorrelation structure in multivariate polytomous panel data models. In addition, Skinner and Holmes (2003) studied two approaches for dealing with sampling effects, either by taking the repeated observations as multivariate outcomes and utilising weighted estimators that account for the correlation structure, or by considering a two-level longitudinal model.

Skinner and Vieira (2007) presented some empirical and theoretical evidence that the variance-inflating impacts of clustering may be higher for longitudinal analyses than for corresponding cross-sectional analyses and that those effects may increase with the number of waves considered in some types of analysis. Moreover, Vieira and Skinner (2008) considered parametric models for panel data and have proposed methods of estimating model parameters that allow for complex schemes by incorporating survey weights into alternative point estimation procedures and using linearisation methods for variance estimation (see Vieira 2009 for further references).

Large-scale longitudinal studies usually involve the selection of a probability sample from a population at the time the panel starts. Weighting in the panel data context has three main aims. If we consider, for example, a survey with two waves, then the longitudinal weight at Wave 2 would: (i) account for unequal selection probabilities at Wave 1, (ii) adjust for unit nonresponse which may occur at Waves 1 and 2, and (iii) adjust (via poststratification, raking or calibration) so that weighted sample estimates for certain auxiliary variables match their respective known population parameters. Longitudinal weights, therefore, allow for different selection probabilities and nonresponse at Wave 1 and attrition, and are adjusted, at each wave, to take account of previous wave respondents' absence through refusal at the current wave or through some other way of sample attrition. Longitudinal weights are calculated in order to guarantee the property that weighted sample moments are consistent for population moments with respect to the joint sampling/nonresponse probability distribution.

The current article further examines the impacts of clustering in panel data analysis, previously investigated by Skinner and Vieira (2007). Moreover, the impacts of survey weighting and stratification are studied by comparing these with the impact on corresponding cross-sectional analyses and by examining how these effects behave with increases in the number of survey waves considered in the analysis. Misspecification effects (*meffs*) for parameter estimates in regression models for (i) the logarithm of household income and (ii) a material satisfaction score are used to evaluate the impact of various features of complex designs on inference. The data are taken from Waves 12 to 15 of the British Household Panel Study (BHPS). To validate the conclusions from an empirical study, a simulation study is also performed, where the use of the *meffs* as a measure of incorrect specification of the model considered is also extensively explored in the longitudinal data analysis context.

The contribution of the current article, when compared to Skinner and Vieira (2007), is (i) the investigation of the impacts of survey weighting and stratification, (ii) the consideration of alternative *meff* measures, (iii) the undertaking of a detailed simulation study, and (iv) the use of the *meffs* as a measure of the impact of incorrect specification of longitudinal models.

This article is organised in six sections. In Section 2 we introduce the panel data under analysis. Section 3 introduces the models, point and variance estimation procedures, and describes the various *meffs*. In Section 4 we present our motivating application and empirical results obtained from real panel data. In Section 5, the simulation study conducted is described and its results are presented. The concluding discussion is presented in Section 6.

2. Data and Sampling Design

The empirical evidence presented in this article is based upon data from the BHPS, which was a large nationally representative household panel survey of individuals in private domiciles in Great Britain (see Taylor et al. 2010). This survey had the main objective of providing information about social and economic change at the individual and household levels.

The BHPS is a longitudinal survey and adopts a complex multistage sampling scheme for collecting data. In addition, it has a multiple-cohort prospective panel design. At Wave 1, in 1991, the survey design involved (i) a multistage stratified clustered probability design with systematic sampling and (ii) approximately equal probability selection of households. As primary sampling units (PSUs or clusters), 250 postcode sectors were selected, with replacement, and with probability of selection proportional to size, using a systematic sampling procedure. The final strata are the result of several stratification stages, which may be summarised as follows:

(a) First, the population was divided into 18 implicit regional strata (regions).
(b) Within each region, PSUs were ranked and then split into major strata of approximately equal size based on the proportion of heads of households in professional or managerial positions.
(c) Within major strata, PSUs were reranked by the proportion of their population in pensionable age.
(d) Major strata were then split into two minor strata: a nonmetropolitan area, with PSUs sorted by their proportion of employed population in agriculture; and a metropolitan area, with PSUs sorted by their population both under pensionable age and living in single-person households. For further details on the BHPS sampling design, see Taylor et al. (2010).

Our analyses are based upon a subset of 2,255 men and women aged 16 or more, clustered in 234 PSUs, who were original sample members, who gave a full interview in Waves 12 to 15 (collected from 2002 until 2005), and who were employed throughout the period. This results in a balanced panel. Note that we study the same subsample considered by Salgueiro et al. (2013), which does not include the BHPS extension samples selected from Scotland, Wales, and Northern Ireland. Therefore, $T = 4$, where T is the number of waves considered. BHPS respondents were asked to answer several questions related to sociodemographic, economic, and attitudinal characteristics. The following variables are considered in our analysis: gender, age category, number of children in the household, education level, social class, marital status, health status, hours normally worked per week, and the logarithm of the household income.

The BHPS data set includes longitudinal weights w_i, which are provided for individual cases that have responded at each wave up to and including the latest wave (Wave 15 in our analysis). The longitudinal weight at any wave generally accounts for losses between each immediate pair of waves up to that point and for the initial sampling design. For information regarding how the weights are defined for the BHPS, see Taylor et al. (2010), where further details about the sampling design of the BHPS are also given.

We have also included a material satisfaction score variable in our data set. Factor analysis, undertaken by Salgueiro et al. (2013), was used to assess which BHPS measures of subjective wellbeing could be combined into a measure of satisfaction with material dimensions of life. A material satisfaction score has subsequently been calculated for each respondent as the total sum of the responses to the following three satisfaction variables: (i) satisfaction with household income, (ii) satisfaction with house/flat, and (iii) satisfaction with job. These three variables were originally measured on a scale from 1 (not satisfied) to 7 (completely satisfied).

In our sample, the relative frequency for males and females is approximately 50%. The distribution of the age category variable is negatively skewed, as the frequencies for the older categories are larger. Most of the respondents were either married or living as a couple in 2002. Approximately 80% of the respondents considered themselves in either a good or excellent health condition. Furthermore, over 75% of the individuals worked at least 30 hours per week. About 55% of the individuals had a high level of education, and only 16% of them occupied a partly skilled or an unskilled position in their last job. Almost 62% of the respondents had no children in the household where they live.

Moreover, in 2002, the average household income of the sample members was approximately 3,365 British pounds in the month before the interview was made.

3. Models, Estimation Procedures, and Misspecification Effects

Regression models have found a wide range of useful applications with panel data (e.g., Diggle et al. 2002; Fitzmaurice et al. 2004). Such data consist of repeated observations on the same variables for the same individuals across equally spaced waves of data collection. The models considered here are concerned with representing the relationship between one of the variables, treated as dependent, and several other variables, treated as covariates. We shall adopt i to denote an individual and t to denote time. We denote the survey variable of interest as y_{it} for individual i at time t. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ be the vector of repeated measures. For the population, we consider linear models of the following form to represent the expectation of $\mathbf{y}_i$ given the values of covariates:

$$E(\mathbf{y}_i) = \mathbf{x}_i \boldsymbol{\beta}, \tag{1}$$

where $\mathbf{x}_i = (\mathrm{x}_{i1}, \ldots, \mathrm{x}_{iT})'$ is a $T \times q$ matrix, $\mathbf{x}_{it}$ is a vector of specified values of q covariates for individual i at time t, $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients, and the expectation is with respect to the model.

The estimation of $\boldsymbol{\beta}$ is based on data from the 'longitudinal sample', s, (i.e., the sample for which observations are available for each $t = 1, \ldots, T$). Following the pseudolikelihood approach (Skinner 1989b), the most general estimator of $\boldsymbol{\beta}$ considered in this article is (Skinner and Vieira 2007)

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \epsilon s} w_i \mathbf{x}_i{}' \mathbf{V}^{-1} \mathbf{x}_i \right)^{-1} \sum_{i \epsilon s} w_i \mathbf{x}_i{}' \mathbf{V}^{-1} \mathbf{y}_i, \tag{2}$$

where w_i is a longitudinal survey weight, $\mathbf{V}$ is a $T \times T$ estimated 'working' variance matrix of $\mathbf{y}_i$ given $\mathbf{x}_i$ (Diggle et al. 2002), taken as the exchangeable variance matrix with diagonal elements $\hat{\sigma}^2$ and off-diagonal elements $\hat{\rho}\hat{\sigma}^2$ and $(\hat{\rho}, \hat{\sigma}^2)$ is an estimator of $(\rho, \hat{\sigma}^2)$. Further details on the pseudolikelihood approach may be found in Vieira (2009). The parameter ρ is the intra-individual correlation and σ^2 is the variance of y_{it}. Further discussion on the estimation of $\boldsymbol{\beta}$ and ρ is presented in Skinner and Vieira (2007). Notice that $\hat{\boldsymbol{\beta}}$ would be fully efficient when the underlying working model holds. Furthermore, under (1), $\hat{\boldsymbol{\beta}}$ is approximately unbiased with respect to the model and to the survey design, and may still be expected to associate both within and between individual information in a reasonably efficient manner, even if the working model for the error structure does not hold exactly (Skinner and Vieira 2007). Without the weight terms and survey-sampling considerations, the form of $\hat{\boldsymbol{\beta}}$, given by (2), is motivated by the generalised estimating Equations (GEE) approach of Liang and Zeger (1986). We shall denote this *unweighted* version by $\hat{\boldsymbol{\beta}}^{\mathrm{u}}$. The following estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ allows for a stratified multistage sampling scheme and it is based upon the classical method of linearisation (Binder 1983; Skinner 1989b; Skinner and Vieira 2007)

$\mathbf{v}(\hat{\boldsymbol{\beta}}) = \left[\sum_{i \epsilon s} w_i \mathbf{x}_i{}' \mathbf{V}^{-1} \mathbf{x}_i\right]^{-1} \left[\sum_h n_h/(n_h - 1) \sum_a (\mathbf{z}_{ha} - \bar{\mathbf{z}}_h)(\mathbf{z}_{ha} - \bar{\mathbf{z}}_h)'\right] \left[\sum_{i \epsilon s} w_i \mathbf{x}_i{}' \mathbf{V}^{-1} \mathbf{x}_i\right]^{-1}$, where h denotes stratum, a denotes PSU, n_h is the number of PSUs in stratum h,

$\mathbf{z}_{ha}{=}\sum_i w_i \mathbf{x}_i{}' \mathbf{V}^{-1} \mathbf{e}_i$, $\bar{\mathbf{z}}_h{=}\sum_a \mathbf{z}_{ha}/n_h$ and $\mathbf{e}_i = \mathbf{y}_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$. If the weights, the sampling scheme and the difference between $n/(n - 1)$ and 1 are ignored, this estimator reduces to the 'robust' variance estimator presented by Liang and Zeger (1986), which is as consistent when (1) holds, even when the working variance matrix $\mathbf{V}$ does not reflect the true variance structure (Diggle et al. 2002). We shall consider three further alternatives for estimating the covariance matrix of $\hat{\boldsymbol{\beta}}$: (i) $\mathbf{v}_a(\hat{\boldsymbol{\beta}})$, which considers that the population consists of only one stratum ($h = 1$), and therefore ignores stratification but takes *area* clustering into account; (ii) $\mathbf{v}_h(\hat{\boldsymbol{\beta}})$, which considers that each individual i is a PSU, and therefore ignores clustering but takes stratification into account; and (iii) the *naive* $\mathbf{v}_n(\hat{\boldsymbol{\beta}})$, which considers that $h = 1$ and that each individual is a PSU, and therefore ignores both stratification and clustering. We shall also perform variance estimation for $\hat{\boldsymbol{\beta}}^{\mathrm{u}}$, which is the point estimator that ignores the weights and stratification, and considers each individual as a PSU.

We shall be concerned with the potential bias of $\mathbf{v}_a(\hat{\boldsymbol{\beta}})$, $\mathbf{v}_h(\hat{\boldsymbol{\beta}})$ and $\mathbf{v}_n(\hat{\boldsymbol{\beta}})$ when in fact the design is complex. Skinner (1989a) has proposed the *misspecification effect* (*meff*), which is designed to measure the effects of incorrect specification of both (i) all the features of the sampling scheme and (ii) the model considered. The effect of the complex sampling scheme on $\mathbf{v}_a(\hat{\boldsymbol{\beta}})$, $\mathbf{v}_h(\hat{\boldsymbol{\beta}})$ and $\mathbf{v}_n(\hat{\boldsymbol{\beta}})$ can be evaluated by considering alternative *meffs* estimators, such as

$$meff_a\left[\hat{\beta}_k, v_a(\hat{\beta}_k)\right] = v(\hat{\beta}_k)/v_a(\hat{\beta}_k),\ meff_h\left[\hat{\beta}_k, v_h(\hat{\beta}_k)\right]$$
$$= v(\hat{\beta}_k)/v_h(\hat{\beta}_k),\ \text{and}\ meff_n\left[\hat{\beta}_k, v_n(\hat{\beta}_k)\right] = v(\hat{\beta}_k)/v_n(\hat{\beta}_k),$$

where $\hat{\beta}_k$ denotes the kth element of $\hat{\boldsymbol{\beta}}$. The $meff_a$, $meff_h$, and $meff_n$ separately estimate the impacts of stratification, clustering, and both stratification and clustering, respectively, and therefore are particular cases of the original *meff* of Skinner (1989a). We shall also calculate all the versions of the *meff* measure considered for $\hat{\boldsymbol{\beta}}^u$. Furthermore, a general *meff*,

$$meff_g = v(\hat{\beta}_k)/v_n\left(\hat{\beta}_k^u\right),$$

with $\hat{\beta}_k^u$ denoting the kth element of $\hat{\boldsymbol{\beta}}^u$, defined above, shall be calculated in order to access the bias caused by ignoring all the sampling-scheme features.

4. Applications

We consider two applications of regression analysis for four waves of the BHPS data, which include (i) the logarithm of the household income and (ii) a material satisfaction score as the dependent variables. Covariates were selected on the basis of the discussion in Salgueiro et al. (2013) and include time, gender, age category, marital status, number of children in the household, education level, social class, health status, and number of hours normally worked per week. We first estimate *meffs* for the linearisation estimator, considering $\hat{\boldsymbol{\beta}}$, as discussed in Section 3. By using data from just the first wave and setting $x_i = 1$, the estimated $meff_n$ for this cross-sectional mean is given in Table 1 and equals 1.343. In order to evaluate the impact of the longitudinal aspect of the data, we estimated a sequence of each of the *meffs* discussed above, using data for time 1, . . . , t, for $t = 2, 3, 4$. It is important to note that the estimation of cross-sectional and longitudinal means is often

Table 1. Meff estimates for estimated longitudinal means.

Dependent Variable	*Meff*	Waves			
		12	12 and 13	12 to 14	12 to 15
Log of the household income	$meff_a[\hat{\beta}_k, v_a(\hat{\beta}_k)]$	0.971	0.965	0.965	0.963
	$meff_h[\hat{\beta}_k, v_h(\hat{\beta}_k)]$	1.490	1.653	1.699	1.695
	$meff_n[\hat{\beta}_k, v_n(\hat{\beta}_k)]$	1.282	1.431	1.474	1.458
	$meff_a\left[\hat{\beta}_k^u, v_a\left(\hat{\beta}_k^u\right)\right]$	0.969	0.963	0.961	0.960
	$meff_h\left[\hat{\beta}_k^u, v_h\left(\hat{\beta}_k^u\right)\right]$	1.572	1.795	1.830	1.870
	$meff_n\left[\hat{\beta}_k^u, v_n\left(\hat{\beta}_k^u\right)\right]$	1.343	1.504	1.575	1.653
	$meff_g$	1.494	1.598	1.778	1.706
Material satisfaction score	$meff_a[\hat{\beta}_k, v_a(\hat{\beta}_k)]$	0.994	0.997	0.993	0.889
	$meff_h[\hat{\beta}_k, v_h(\hat{\beta}_k)]$	1.075	1.125	1.190	1.197
	$meff_n[\hat{\beta}_k, v_n(\hat{\beta}_k)]$	1.087	1.104	1.135	1.132
	$meff_a\left[\hat{\beta}_k^u, v_a\left(\hat{\beta}_k^u\right)\right]$	1.000	1.000	0.996	0.996
	$meff_h\left[\hat{\beta}_k^u, v_h\left(\hat{\beta}_k^u\right)\right]$	1.079	1.113	1.182	1.199
	$meff_n\left[\hat{\beta}_k^u, v_n\left(\hat{\beta}_k^u\right)\right]$	1.119	1.155	1.207	1.203
	$meff_g$	1.306	1.309	1.328	1.297

the aim of official statistics agencies, and therefore we consider the following analysis to be of particular relevance.

Although these estimated *meffs* are subject to sampling error, there seems to be some evidence from Table 1 of a tendency for $meff_h$, $meff_n$, and $meff_g$ to increase with the number of waves. It therefore seems like it becomes more important to allow for clustering and for the complex sampling design in general when the number of waves in the analysis increases. This result agrees with Skinner and Vieira (2007). Furthermore, the stratification effects appear ($meff_a$) to remain constant as the number of waves increases.

The models with logarithm of the household income as the dependent variable appear to have larger values for $meff_h$, $meff_n$, and $meff_g$ than the models with a material satisfaction score as the dependent variable. This result was expected, as attitudinal variables tend to have small estimated intracluster (intra-postcode) correlations for variables in British surveys (Lynn and Lievesley 1991; Vieira and Skinner 2008).

We have elaborated the analysis by including educational level as a covariate and we present in Table 2 only *meff* estimates for the estimated coefficients for the constant term of the longitudinal models.

The main feature of these results is that, as before, there is some evidence that $meff_h$, $meff_n$, and $meff_g$ increase with the number of waves. The intercept term may be seen as a domain mean, and standard survey-sampling theory for a *meff* of a mean in a domain cutting across clusters (Skinner 1989b; Skinner and Vieira 2007) implies that it will be somewhat less than a *meff* for the mean in the whole sample, as we have generally observed when comparing the results in Table 2 with those from Table 1. Moreover, such a comparison also confirms the observation of Kish and Frankel (1974) and Skinner and Vieira (2007) that *meffs* for regression coefficients tend not to be greater than *meffs* for the

Table 2. Meff estimates for the estimated constant terms in the longitudinal models (with one education covariate).

Dependent Variable	*Meff*	Waves 12	 12 and 13	 12 to 14	 12 to 15
Log of the household income	$meff_a[\hat{\beta}_k, v_a(\hat{\beta}_k)]$	1.000	1.000	1.000	0.980
	$meff_h[\hat{\beta}_k, v_h(\hat{\beta}_k)]$	1.000	1.127	1.179	1.230
	$meff_n[\hat{\beta}_k, v_n(\hat{\beta}_k)]$	1.016	1.108	1.118	1.143
	$meff_a\left[\hat{\beta}_k^u, v_a\left(\hat{\beta}_k^u\right)\right]$	0.983	0.982	0.980	0.980
	$meff_h\left[\hat{\beta}_k^u, v_h\left(\hat{\beta}_k^u\right)\right]$	1.104	1.117	1.274	1.330
	$meff_n\left[\hat{\beta}_k^u, v_n\left(\hat{\beta}_k^u\right)\right]$	1.051	1.131	1.208	1.237
	$meff_g$	1.195	1.190	1.208	1.214
Material satisfaction score	$meff_a[\hat{\beta}_k, v_a(\hat{\beta}_k)]$	0.996	0.998	0.998	1.000
	$meff_h[\hat{\beta}_k, v_h(\hat{\beta}_k)]$	1.038	1.052	1.111	1.065
	$meff_n[\hat{\beta}_k, v_n(\hat{\beta}_k)]$	0.972	1.046	1.128	1.087
	$meff_a\left[\hat{\beta}_k^u, v_a\left(\hat{\beta}_k^u\right)\right]$	0.993	0.995	0.998	1.002
	$meff_h\left[\hat{\beta}_k^u, v_h\left(\hat{\beta}_k^u\right)\right]$	1.127	1.172	1.137	1.120
	$meff_n\left[\hat{\beta}_k^u, v_n\left(\hat{\beta}_k^u\right)\right]$	1.069	1.176	1.180	1.174
	$meff_g$	1.247	1.268	1.406	1.323

means of the dependent variable. Again the stratification effects appear to be constant with increases in the number of waves.

Although we have chosen not to present the *meffs* for the contrasts (coefficients for the education level covariate considered in the model), we have observed that they have varied in size and generally do not show any tendency to converge to one as the number of waves analysed increases, which would indicate no misspecification. As observed by Skinner and Vieira (2007), a *meff* for a contrast may be considered a combination of the traditional variance-inflating effect of clustering in surveys together with the variance-reducing effect of blocking in an experiment. Such variance reduction may be observed when the domains being contrasted share a common cluster effect that tends to cancel out in the contrasts, and therefore may imply that the actual variance of the contrast is lower than the expectation of the variance estimator which assumes independence between domains (Skinner and Vieira 2007).

The models have been further refined by the inclusion of additional covariates:

- time
- gender (g1 ¼ male, reference category; and g2 ¼ female)
- age category (ac1 ¼ 16 to 21 years, reference category; ac2 ¼ 22 to 29 years; ac3 ¼ 30 to 39 years; ac4 ¼ 40 to 49 years; and ac5 ¼ 50 years or older)
- number of children in the household
- education level (el1 ¼ first or higher degree, reference category; el2 ¼ other higher qualification; el3 ¼ nursing or A-levels; el4 ¼ other levels; and el5 ¼ no post-school qualification)

- social class (sc1 ¼ professional occupation, reference category; sc2 ¼ managerial or technical; sc3 ¼ skilled; and sc4 ¼ partly skilled or unskilled),
- health status (hs1 ¼ excellent, reference category; hs2 ¼ good; hs3 ¼ fair; and hs4 ¼ poor), numbers of hours normally worked per week (nh1 ¼ less than 16 hours, reference category; nh2 ¼ 16 to 29 hours; nh3 ¼ 30 to 40 hours; and nh4 ¼ more than 40 hours)
- and marital status (ms1 ¼ married or living as a couple, reference category, and ms2 ¼ widowed, divorced, separated or never married).

For the model with a material satisfaction score as the dependent variable, we have also added the logarithm of the household income as a covariate. As before, in Table 3 we present *meff* estimates only for the estimated coefficients for the constant term of the further elaborated longitudinal models.

There is some evidence of a tendency in the *meffs* for the constant to diverge from unity as the number of waves increases, especially for the model with a material satisfaction score as the dependent variable. Although we have not presented the *meffs* for the covariates, we have observed that $meff_h$, $meff_n$, and $meff_g$ generally have not shown any tendency to converge to one, for the same reasons as we have argued above. In general, when comparing the results in Table 3 with those in Tables 1 and 2, we have also confirmed the observation of Kish and Frankel (1974) and Skinner and Vieira (2007) that *meffs* for regression coefficients tend not to be greater than *meffs* for the means of the dependent variable, except for the estimated *meffs* for the constant term of the model with a material satisfaction score as dependent variable, which has presented surprisingly high *meffs* for the more elaborate model.

Table 3. Meff estimates for the estimated constant terms in the longitudinal models (with several covariates).

		Waves			
Dependent Variable	*Meff*	12	12 and 13	12 to 14	12 to 15
Log of the household income	$meff_a[\hat{\beta}_k,v_a(\hat{\beta}_k)]$	0.982	1.000	1.000	1.001
	$meff_h[\hat{\beta}_k,v_h(\hat{\beta}_k)]$	1.000	0.948	0.994	0.944
	$meff_n[\hat{\beta}_k,v_n(\hat{\beta}_k)]$	0.829	0.938	1.000	0.970
	$meff_a\left[\hat{\beta}_k^u, v_a\left(\hat{\beta}_k^u\right)\right]$	1.000	0.994	1.000	1.002
	$meff_h\left[\hat{\beta}_k^u, v_h\left(\hat{\beta}_k^u\right)\right]$	0.980	0.966	0.981	0.916
	$meff_n\left[\hat{\beta}_k^u, v_n\left(\hat{\beta}_k^u\right)\right]$	0.849	0.955	0.994	0.947
	$meff_g$	1.000	1.124	1.175	1.138
Material satisfaction score	$meff_a[\hat{\beta}_k,v_a(\hat{\beta}_k)]$	0.992	0.996	0.992	1.000
	$meff_h[\hat{\beta}_k,v_h(\hat{\beta}_k)]$	1.211	1.273	1.311	1.112
	$meff_n[\hat{\beta}_k,v_n(\hat{\beta}_k)]$	1.184	1.278	1.349	1.205
	$meff_a\left[\hat{\beta}_k^u, v_a\left(\hat{\beta}_k^u\right)\right]$	0.993	0.996	0.991	1.000
	$meff_h\left[\hat{\beta}_k^u, v_h\left(\hat{\beta}_k^u\right)\right]$	1.176	1.225	1.369	1.200
	$meff_n\left[\hat{\beta}_k^u, v_n\left(\hat{\beta}_k^u\right)\right]$	1.155	1.250	1.432	1.306
	$meff_g$	1.413	1.573	1.628	1.446

Table 4 presents coefficient, standard error ($\mathbf{se}(\hat{\boldsymbol{\beta}}) = \sqrt{\mathbf{v}(\hat{\boldsymbol{\beta}})}$ and $\mathbf{se}_n(\hat{\boldsymbol{\beta}}) = \sqrt{\mathbf{v}_n(\hat{\boldsymbol{\beta}})}$) and *meff* estimates for the model for Waves 12 to 15 with logarithm of the household income as the dependent variable and several covariates. The differences observed when we compare the point estimates produced by the standard Liang and Zeger (1986) estimator ($\hat{\boldsymbol{\beta}}_n$, given by Equation (2) without the weight terms) and the weighted pseudolikelihood estimator ($\hat{\boldsymbol{\beta}}$, given by Equation (2)) suggest that using standard statistical techniques with complex sampling data may lead to biased point estimates. Note the differences in the estimated coefficients for gender, age category, health status, and numbers of hours normally worked, confirming Nathan and Holt's (1980) results produced in a cross-sectional context. Moreover, the results in Table 4 also suggest that, in general, the BHPS complex sampling effects, if not taken into account in the estimation procedure, tend to lead to an underestimation of standard errors (compare columns labelled (1) and (2) and columns labelled (3) and (4)), and therefore to narrowed confidence intervals and possibly to the incorrect rejection of null hypotheses. In our application, complex sampling effects may lead to inappropriate statistical conclusions. This is confirmed by the estimated *meffs*, which are generally above one and even above two for gender. The $meff_n$ for $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\beta}}$ are similar, suggesting the impact of the complex sampling is the same irrespective of whether or not weights are used. However, $meff_g$ is nearly always larger than both these $meff_n$, suggesting that the effect of weighting is to further increase the estimated standard errors.

Figure 1 includes confidence intervals for both $\hat{\boldsymbol{\beta}}^u$ and $\hat{\boldsymbol{\beta}}$, considering both $\mathbf{se}_n(.)$ and $\mathbf{se}(.)$, for coefficients of covariates which had at least one $meff_g > 1.5$. Horizontal lines are represented both at $\beta = 0$ and $\hat{\boldsymbol{\beta}}$ for the plots on the left-hand side, and only $\beta = 0$ for the right-hand ones. Four different confidence intervals were calculated for each coefficient, labelled as: (a) confidence interval for $\hat{\boldsymbol{\beta}}^u$ based on $\mathbf{se}_n(.)$, (b) confidence interval for $\hat{\boldsymbol{\beta}}^u$ based on $\mathbf{se}(.)$, (c) confidence interval for $\hat{\boldsymbol{\beta}}$ based on $\mathbf{se}_n(.)$, and (d) confidence interval for $\hat{\boldsymbol{\beta}}$ based on $\mathbf{se}(.)$. Note, therefore, that: (a) does not allow for any sampling design features, (b) allows for clustering and stratification, (c) allows for weighting, and (d) allows for clustering, stratification, and weighting. The comparison of (a), (b), (c), and (d) helps us to evaluate the different sampling misspecification effects. Our plots demonstrate that different coefficients show different types of effects. The plot for the variable number of children, for example, shows a common situation faced by data analysts. Note the coefficients are considered significant when the sampling design is not considered in (a). Moving from (a) to (d), sampling design features are gradually being considered, leading to the coefficient not being significant in (d). Plots for social class and gender show weighting and stratification effects in the standard-error estimation. The plot for age category illustrates the effects of weighting, and the possibility of bias, in the point estimates. Plots for time and health status show different patterns for the evaluated effects depending on which point estimator is being considered.

5. Simulation Study

As the results reported in Section 4 are subject to sampling error, we conducted a simulation study to evaluate the behaviour of the *meff* measures. Each of the $d = 1, \ldots, D$ replicate samples is based on the BHPS data subset described above, which is

Table 4. Estimates for the four-waves model with logarithm of the household income as the dependent variable and several covariates.

Covariates		$\hat{\beta}_k^u$	$se_n(\hat{\beta}_k^u)$	$se(\hat{\beta}_k^u)$	$meff_n$	$\hat{\beta}_k$	$se_n(\hat{\beta}_k)$	$se(\hat{\beta}_k)$	$meff_n$	$meff_g$
			(1)	*(2)*	$(2)^2/(1)^2$		*(3)*	*(4)*	$(4)^2/(3)^2$	$(4)^2/(1)^2$
Constant		3.66	0.029	0.028	0.947	3.64	0.031	0.031	0.970	1.138
Time		0.01	0.001	0.002	1.299	0.01	0.002	0.002	1.235	1.689
Gender	*g2*	−0.01	0.007	0.006	0.691	−0.02	0.012	0.010	0.680	2.085
Age category	*ac2*	−0.10	0.023	0.024	1.053	−0.08	0.025	0.028	1.172	1.433
	ac3	−0.09	0.022	0.023	1.095	−0.08	0.027	0.029	1.189	1.701
	ac4	−0.04	0.023	0.024	1.073	−0.03	0.025	0.026	1.145	1.310
	ac5	−0.10	0.022	0.023	1.076	−0.08	0.024	0.025	1.131	1.283
No. children		−0.01	0.004	0.005	1.179	−0.01	0.005	0.006	1.079	1.758
No. of hours	*nh2*	0.04	0.015	0.015	0.936	0.03	0.014	0.014	0.963	0.861
	nh3	0.06	0.012	0.011	0.915	0.06	0.013	0.013	0.938	1.126
	nh4	0.08	0.015	0.015	1.070	0.07	0.015	0.015	1.040	1.051
Education level	*el2*	−0.08	0.010	0.011	1.269	−0.08	0.011	0.012	1.056	1.371
	el3	−0.09	0.013	0.014	1.166	−0.09	0.015	0.016	1.067	1.359
	el4	−0.13	0.012	0.013	1.199	−0.13	0.014	0.014	1.102	1.378
	el5	−0.16	0.014	0.015	1.261	−0.15	0.015	0.016	1.208	1.396
Social class	*sc2*	−0.02	0.010	0.010	0.913	−0.02	0.015	0.014	0.958	1.920
	sc3	−0.05	0.011	0.011	1.073	−0.05	0.014	0.014	1.063	1.696
	sc4	−0.07	0.013	0.013	1.070	−0.07	0.016	0.017	1.134	1.681
Health status	*hs2*	−0.02	0.005	0.005	1.037	−0.01	0.004	0.004	1.117	0.536
	hs3	−0.03	0.006	0.006	1.216	−0.03	0.007	0.008	1.231	1.883
	hs4	−0.03	0.009	0.009	0.994	−0.03	0.011	0.012	1.104	1.806
Marital status	*ms2*	−0.07	0.011	0.011	0.967	−0.07	0.012	0.012	1.020	1.184

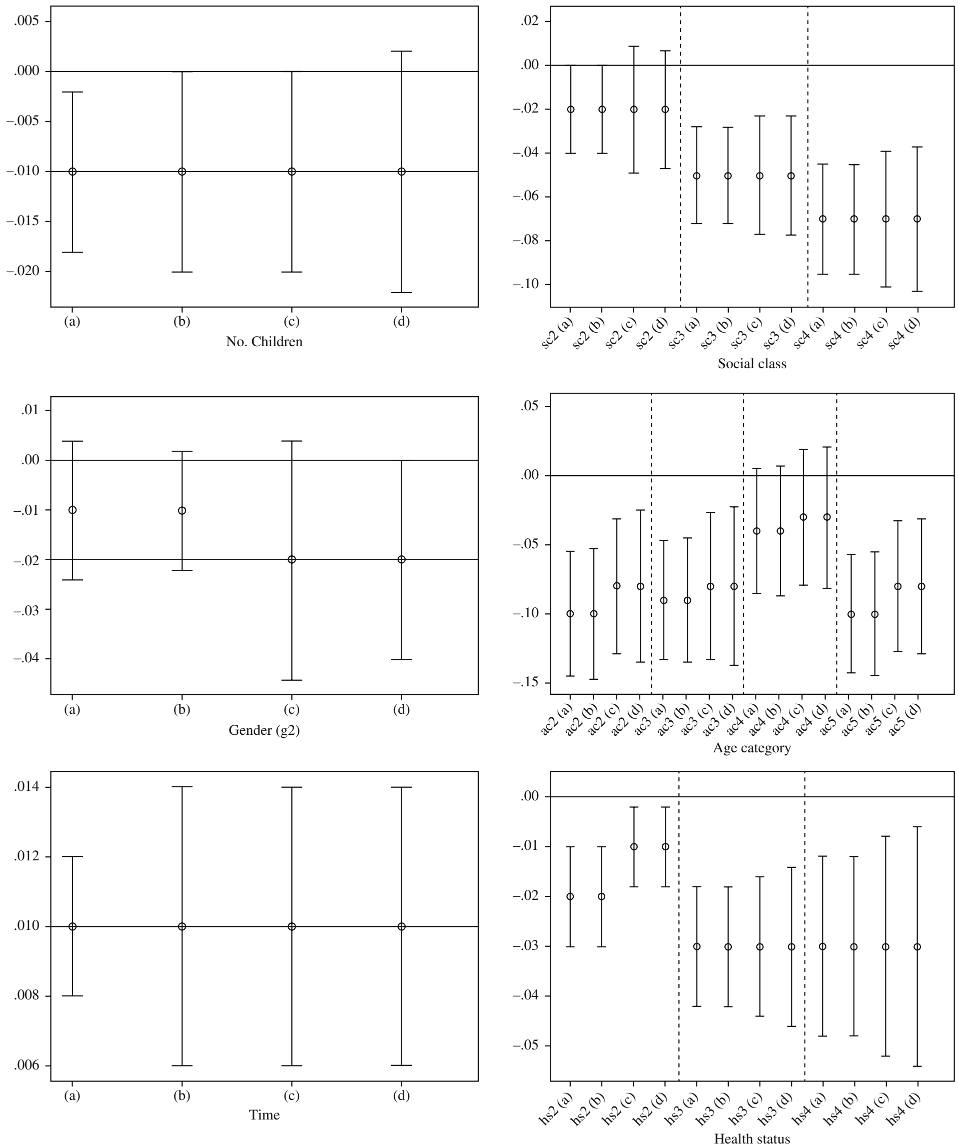


Fig. 1. Confidence intervals for coefficients of covariates with meff$_g$ > 1.5.

considered as the ‘target population’. We evaluated the properties of variance estimators for unweighted point estimators and assessed only the impacts of clustering. We studied the *meff* when the number of waves in the analysis is increased. Note that we do not assess the impact of stratification, unequal probability sampling, nonresponse, and attrition.

Let y_{iat} be the value for the study variable for unit $i = 1, 2, \ldots, n_a^{sim}$, in PSU $a = 1, \ldots, m^{sim}$, at Wave t of the survey, where n_a^{sim} and m^{sim} are the sample sizes and the number of PSUs for the replicate sample d. To generate the values of y_{iat} for the simulation study, we used the following uniform correlation model, which includes a clustering effect:

$$y_{iat} = \mathbf{x}_{iat}\boldsymbol{\beta} + \eta_a + u_{ia} + v_{iat}, \tag{3}$$

where η_a represents the PSU (postcode area) random effects, u_{ia} denotes individual-level random effects (or unobservable individual specific factors), and v_{iat} are residuals, with $\eta_a \sim N\left(0, \sigma_\eta^2\right), u_{ia} \sim N\left(0, \sigma_u^2\right)$, and $v_{iat} \sim N\left(0, \sigma_v^2\right)$. We consider the logarithm of the household income and a material satisfaction score as dependent variables and the remaining variables listed and described in Section 2 as covariates. We have held the values of the covariates fixed.

The values adopted for $\boldsymbol{\beta}$, σ_η, σ^u, and σ_v were based on maximum-likelihood estimates for the model fitted to the 'target population', which were 0.16 and 2.10 for σ_u, and 0.11 and 1.88 for σ_v, respectively, for the models with the logarithm of the household income and with a material satisfaction score as dependent variables. In order to evaluate the effects of different impacts of clustering on the variance estimation procedures considered, we used the following realistic choices for σ_η: (i) $\sigma_\eta = 0.06$ (actual value estimated from fitting Model (3) to the data), $\sigma_\eta = 0.12$ and $\sigma_\eta = 0.18$, for the model with the logarithm of the household income as dependent variable; and (ii) $\sigma_\eta = 0.35$ (actual value estimated from fitting (3)), $\sigma_\eta = 0.70$, and $\sigma_\eta = 1.05$ for the model with a material satisfaction score as the dependent variable.

Let

$$\hat{E}(\hat{meff}) = \frac{1}{D}\sum_{d=1}^{D} \hat{meff}^{(d)}$$

be the mean of our *meff* of interest estimated over repeated simulation,

$$var(\hat{meff}) = \frac{1}{D-1}\sum_{d=1}^{D}\left[\hat{meff}^{(d)} - \hat{E}(\hat{meff})\right]^2$$

be a simulation estimator of VAR($\hat{meff}$), the population variance of the misspecification-effect measure, and

$$se[\hat{E}(\hat{meff})] = \sqrt{var(\hat{meff})/D}$$

be the simulation standard error of $\hat{E}(\hat{meff})$.

We initially set $x_i = 1$ in the models fitted to each generated replicate sample and therefore studied the behaviour of the *meff* for longitudinal means. We set n_a^{sim} equal the sample size for PSU a in our BHPS subsample and $m^{sim} = 234$, which is equal to the number of PSUs in our BHPS subsample. Therefore, Table 5 presents simulation results for four scenarios, including one that considers $\sigma_\eta = 0.00$ (i.e., no clustering effect), when $D = 1{,}000$.

The simulation results also provide evidence that the *meffs* increase as the number of waves in the analysis increases, at least for longitudinal means. This increase seems to be stronger for larger intracluster correlation. We also observe an increase in the *meff* when the intracluster correlation increases, as expected from the survey-sampling literature (Kish and Frankel 1974; Holt and Scott 1981; Scott and Holt 1982; Skinner 1986; and Skinner 1989a).

Table 5. $\hat{E}(\widehat{meff})$ *and* $se[\hat{E}(\widehat{meff})]$ *(in brackets), for four scenarios (for longitudinal means).*

Dependent Variable	σ_η	Waves			
		12	12 and 13	12 to 14	12 to 15
Log of the household income	0.00	1.1448 (0.0035)	1.1561 (0.0035)	1.1589 (0.0036)	1.1597 (0.0036)
	0.06	1.1862 (0.0042)	1.2018 (0.0043)	1.2078 (0.0043)	1.2107 (0.0043)
	0.12	1.2697 (0.0053)	1.2940 (0.0055)	1.3019 (0.0056)	1.3073 (0.0056)
	0.18	1.3774 (0.0068)	1.4061 (0.0070)	1.4190 (0.0070)	1.4255 (0.0071)
Material satisfaction score	0.00	1.0826 (0.0033)	1.0986 (0.0033)	1.1017 (0.0033)	1.1030 (0.0033)
	0.35	1.0890 (0.0033)	1.1063 (0.0034)	1.1105 (0.0035)	1.1129 (0.0035)
	0.70	1.1086 (0.0035)	1.1363 (0.0037)	1.1428 (0.0037)	1.1462 (0.0037)
	1.05	1.1498 (0.0044)	1.1806 (0.0046)	1.1889 (0.0047)	1.1936 (0.0048)

We also notice that the *meffs* are greater than one even when $\sigma_\eta = 0.00$. We believe that this is due to the model that is being fitted (with no covariates), which is different from the true model (with several covariates) that was used to generate the data. Therefore, this is a good example of the use of the *meff* to measure the effects of incorrect specification of both the sampling scheme and the model considered.

Following the same strategy considered in Section 4, we have elaborated the analysis by including educational level as a covariate. Tables 6 and 7 present the results for the constant term and one of the contrasts (one category) of the educational level covariate, for the logarithm of the household income and material satisfaction models respectively, using the same four scenarios as before.

The simulation results with the logarithm of the household income as the dependent variable and the educational level as the covariate also generally show a tendency for the *meffs* to increase as the number of waves in the analysis increases, more clearly for the constant (domain mean) but also for the contrasts (including those contrasts that were not presented in Table 6). This increase seems, again, to be stronger for larger clustering impacts. Furthermore, we notice once again that the *meffs* are greater than one even when $\sigma_\eta = 0.00$, but not as much as we observed in Table 5, as the model that is now being fitted (with one covariate) is slightly closer to the true model.

The simulation results with the material satisfaction score as the dependent variable and the educational level as the covariate, presented in Table 7, lead to very similar conclusions to those drawn from Table 6. In fact the increase in the *meff* is now even clearer. Moreover, when comparing the results from Tables 6 and 7 to those presented in Table 5, we confirm our results from Section 4, and the observation of Kish and Frankel (1974) and Skinner and Vieira (2007) that *meffs* for regression coefficients tend not to be greater than *meffs* for the means of the dependent variable.

Table 6. Ê(mêff) and se[Ê(mêff)] (in brackets), considering four scenarios for the logarithm of the household-income model with one education covariate.

			Waves			
Dependent Variable	σ_η	Coefficient	12	12 and 13	12 to 14	12 to 15
Log of the household income	0.00	Constant	1.0454	1.0441	1.0427	1.0435
			(0.0043)	(0.0043)	(0.0043)	(0.0042)
		el5	1.0478	1.0473	1.0493	1.0473
			(0.0036)	(0.0036)	(0.0036)	(0.0037)
	0.06	Constant	1.0872	1.0908	1.0921	1.0933
			(0.0044)	(0.0044)	(0.0043)	(0.0043)
		el5	1.0864	1.0897	1.0881	1.0827
			(0.0037)	(0.0037)	(0.0038)	(0.0037)
	0.12	Constant	1.1671	1.1892	1.1920	1.1971
			(0.0055)	(0.0057)	(0.0056)	(0.0056)
		el5	1.1683	1.1835	1.1791	1.1709
			(0.0048)	(0.0049)	(0.0048)	(0.0048)
	0.18	Constant	1.2760	1.2950	1.2926	1.2976
			(0.0069)	(0.0068)	(0.0067)	(0.0067)
		el5	1.2644	1.2786	1.2607	1.2458
			(0.0057)	(0.0058)	(0.0055)	(0.0053)

We included the following additional covariates: time, gender, age category, marital status, number of children in the household, education level, social class, health status, and numbers of hours normally worked. As the simulation results presented in Tables 5, 6, and 7 suggested very similar conclusions drawn from the models with the two different

Table 7. Ê(mêff) and se[Ê(mêff)] (in brackets), considering four scenarios for the material satisfaction score model with one education covariate.

			Waves			
Dependent Variable	σ_η	Coefficient	12	12 and 13	12 to 14	12 to 15
Material satisfaction score	0.00	Constant	1.0604	1.0667	1.0721	1.0794
			(0.0043)	(0.0042)	(0.0041)	(0.0041)
		el5	1.0488	1.0513	1.0551	1.0570
			(0.0034)	(0.0034)	(0.0034)	(0.0034)
	0.35	Constant	1.0672	1.0786	1.0843	1.0897
			(0.0043)	(0.0043)	(0.0043)	(0.0043)
		el5	1.0503	1.0585	1.0644	1.0662
			(0.0037)	(0.0036)	(0.0035)	(0.0035)
	0.70	Constant	1.0886	1.0986	1.1075	1.1148
			(0.0043)	(0.0044)	(0.0045)	(0.0045)
		el5	1.0752	1.0837	1.0895	1.0913
			(0.0036)	(0.0037)	(0.0037)	(0.0037)
	1.05	Constant	1.1106	1.1300	1.1406	1.1507
			(0.0048)	(0.0047)	(0.0048)	(0.0047)
		el5	1.0924	1.1094	1.1151	1.1188
			(0.0038)	(0.0039)	(0.0040)	(0.0040)

Table 8. $\hat{E}(\widehat{meff})$ *and* $se[\hat{E}(\widehat{meff})]$ *(in brackets), considering four scenarios for the logarithm of the household-income model with several covariates.*

Dependent Variable	σ_η	Coefficient	Waves 12	Waves 12 and 13	Waves 12 to 14	Waves 12 to 15
Log of the household income	0.00	Constant	0.9855	0.9903	0.9925	0.9926
			(0.0038)	(0.0037)	(0.0037)	(0.0039)
		el5	0.9884	0.9914	0.9933	0.9955
			(0.0034)	(0.0033)	(0.0033)	(0.0033)
	0.06	Constant	0.9911	1.0105	1.0158	1.0193
			(0.0037)	(0.0037)	(0.0038)	(0.0040)
		el5	0.9834	1.0087	1.0196	1.0222
			(0.0033)	(0.0033)	(0.0033)	(0.0034)
	0.12	Constant	0.9938	1.0655	1.0879	1.0961
			(0.0039)	(0.0042)	(0.0047)	(0.0049)
		el5	0.9869	1.0572	1.0870	1.1003
			(0.0032)	(0.0037)	(0.0040)	(0.0041)
	0.18	Constant	0.9793	1.1109	1.1607	1.1766
			(0.0037)	(0.0050)	(0.0057)	(0.0059)
		el5	0.9877	1.1138	1.1626	1.1814
			(0.0033)	(0.0043)	(0.0049)	(0.0052)

dependent variables considered, we have chosen to present results for the logarithm of the household-income models for the more complex model with several covariates. Table 8 presents results for the constant term and for the same contrast that was included in Tables 6 and 7.

Table 8 also shows an increase in the *meff* as the number of waves in the analysis increases. We may draw very similar conclusions to those regarding Tables 6 and 7. Furthermore, we now notice that the *meffs* are much closer to one when $\sigma_\eta = 0.00$, especially for the situation where we consider four waves, as the model that is being fitted in that case (with several covariates) is the true model and no clustering effect is induced. We believe these *meff* results are not significantly different to one as their 95% simulation confidence intervals include one for the four-waves model for most of the estimated coefficients.

6. Discussion

We have presented evidence that the impact of clustering may be stronger for longitudinal studies than for cross-sectional studies, and that *meffs* for the regression coefficients increase with the number of waves considered in the analysis, which confirms previous theoretical results by Skinner and Vieira (2007; Expression (11)). Longitudinal household surveys tend to have a long life in most countries (e.g., Panel Study of Income Dynamics in the United States; German Social Economic Panel in Germany) and therefore a large number of waves, and in such cases our conclusions are particularly relevant. Moreover, we have also observed that *meffs* for regression coefficients tend not to be greater than *meffs* for the means of the dependent variable. In fact, lower *meffs* are expected for models with increasing complexity (with more covariates) or for models that are closer to the true population model, which has been observed in our results. However, as previously stated, official statistical agencies often wish to estimate domain means, which correspond to simple models with, for example, a single covariate, and again in such cases our conclusions are particularly relevant.

Furthermore, our application results suggest that stratification effects remain constant with increases in the number of waves. This conclusion does not seem to be dependent upon the complexity of the model (i.e., number of covariates) that is being considered.

The main implication of our findings is that standard errors estimated in the analysis of panel data may be misleading if the initial sample was clustered and if this clustering is ignored in the analysis, more strongly so in situations where descriptive statistics (such as means) are being estimated or when the model that is being fitted is not well specified. Our results also suggest that longitudinal weighting has implications on both point and standard-error estimation. The analysis of longitudinal data collected by surveys that adopt unequal probability selection procedures, unit-nonresponse weighting adjustments for protection against attrition, and other weighting adjustments requires allowances for such features. Therefore, the types of misspecification that investigators need to protect against are those related to clustering and weighting. We believe that by taking our findings into account, analysts of longitudinal data will be able to produce better inferential results for panel surveys.

Possible future work could include investigating the impacts of various sampling design features in the analysis of panel data based on estimating marginal models for a binary response, such as the ones considered by Roberts et al. (2009). Moreover, the effects of item nonresponse and of the use of imputation in variance estimation in the longitudinal data context, which has not been dealt with here, could also be investigated in future work.

7. References

Binder, D.A. 1983. “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” *International Statistical Review* 51: 279–292. Doi: http://dx.doi.org/10.2307/1402588.

Binder, D.A. and G.R. Roberts. 2003. “Design-Based and Model-Based Methods for Estimating Model Parameters.” In *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner. 29–48. Chichester: Wiley.

Chambers, R.L. and C.J. Skinner. 2003. *Analysis of Survey Data*. Chichester: Wiley.

Diggle, P.J., P. Heagerty, K. Liang, and S.L. Zeger. 2002. *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press.

Feder, M., “Fitting Regression Models to Complex Survey Data – Gelman’s Estimator Revisited.” International Statistical Institute, Proceedings of the 58th World Statistical Congress, Session CPS028, 21–26 August 2011, Dublin, Ireland. Available at: http://2011.isiproceedings.org/ (accessed January 2016).

Feder, M., G. Nathan, and D. Pfeffermann. 2000. “Multilevel Modelling of Complex Survey Longitudinal Data with Time Varying Random Effects.” *Survey Methodology* 26: 53–65.

Fitzmaurice, G.M., M.N. Laird, and J.H. Ware. 2004. *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley.

Holt, D. and A.J. Scott. 1981. "Regression Analysis Using Survey Data." *The Statistician* 30: 169–178. Doi: http://dx.doi.org/10.2307/2988047.

Kish, L. and M.R. Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society, Series B* 36: 1–37.

Liang, K. and S.L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73: 13–22. Doi: http://dx.doi.org/10.1093/biomet/73.1.13.

Lynn, P. and D. Lievesley. 1991. *Drawing General Population Samples in Great Britain*. London: Social and Community Planning Research.

Nathan, G. and D. Holt. 1980. "The Effect of Survey Design on Regression Analysis." *Journal of the Royal Statistical Society, Series B* 42: 377–386.

Pfeffermann, D. 2011. "Modelling of Complex Survey Data: Why is it a Problem? How Should We Approach It?" *Survey Methodology* 37: 115–136.

Roberts, G., Q. Ren, and J.N.K. Rao. 2009. "Using Marginal Mean Models for Data from Longitudinal Surveys with a Complex Design: Some Advances in Methods." In *Methodology of Longitudinal Surveys*, edited by P. Lynn. 351–366. Chichester: Wiley.

Salgueiro, M.F., P.W.F. Smith, and M.D.T. Vieira. 2013. "A Multi-Process Second-Order Latent Growth Curve Model for Subjective Well-Being." *Quality and Quantity* 47: 735–752. Doi: http://dx.doi.org/10.1007/s11135-011-9541-y.

Scott, A.J. and D. Holt. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." *Journal of the American Statistical Association* 77: 848–854. Doi: http://dx.doi.org/10.1080/01621459.1982.10477897.

Skinner, C.J. 1986. "Design Effects of Two-Stage Sampling." *Journal of the Royal Statististical Society, Series B* 48: 89–99.

Skinner, C.J. 1989a. "Introduction to Part A." In *Analysis of Complex Surveys*, edited by C.J. Skinner, D. Holt, and T.M.F. Smith. 23–58. Chichester: Wiley.

Skinner, C.J. 1989b. "Domain Means, Regression and Multivariate Analysis." In *Analysis of Complex Surveys*, edited by C.J. Skinner, D. Holt, and T.M.F. Smith. 59–87. Chichester: Wiley.

Skinner, C.J. and D. Holmes. 2003. "Random Effects Models for Longitudinal Survey Data." In *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner. 205–219. Chichester: Wiley.

Skinner, C.J., D. Holt, and T.M.F. Smith. 1989. *Analysis of Complex Surveys*. Chichester: Wiley.

Skinner, C. and M.D.T. Vieira. 2007. "Variance Estimation in the Analysis of Clustered Longitudinal Survey Data." *Survey Methodology* 33: 3–12.

Smith, P., P. Lynn, and D. Elliot. 2009. "Sample Design for Longitudinal Surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn. 21–33. Chichester: Wiley.

Sutradhar, B.C. and M. Kovacevic. 2000. "Analysing Ordinal Longitudinal Survey Data: Generalised Estimating Equations Approach." *Biometrik* 87: 837–848. Doi: http://dx.doi.org/10.1093/biomet/87.4.837.

Taylor, M.F. (ed.), J. Brice, N. Buck, and E. Prentice-Lane. 2010. *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex.

UK Data Service. *UK Data Service*. Available at: http://ukdataservice.ac.uk (accessed January 2016).

Vieira, M.D.T. 2009. *Analysis of Longitudinal Survey Data*, 1st ed. Saarbrücken: VDM Verlag.

Vieira, M.D.T. and C.J. Skinner. 2008. "Estimating Models for Panel Survey Data under Complex Sampling." *Journal of Official Statistics* 24: 343–364.

Received July 2014
Revised December 2015
Accepted January 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 507–539, http://dx.doi.org/10.1515/JOS-2016-0026

Weight Smoothing for Generalized Linear Models Using a Laplace Prior

Xi Xia[1] *and Michael R. Elliott*[1,2]

When analyzing data sampled with unequal inclusion probabilities, correlations between the probability of selection and the sampled data can induce bias if the inclusion probabilities are ignored in the analysis. Weights equal to the inverse of the probability of inclusion are commonly used to correct possible bias. When weights are uncorrelated with the descriptive or model estimators of interest, highly disproportional sample designs resulting in large weights can introduce unnecessary variability, leading to an overall larger mean square error compared to unweighted methods.

We describe an approach we term ‘weight smoothing’ that models the interactions between the weights and the estimators as random effects, reducing the root mean square error (RMSE) by shrinking interactions toward zero when such shrinkage is allowed by the data. This article adapts a flexible Laplace prior distribution for the hierarchical Bayesian model to gain a more robust bias-variance tradeoff than previous approaches using normal priors. Simulation and application suggest that under a linear model setting, weight-smoothing models with Laplace priors yield robust results when weighting is necessary, and provide considerable reduction in RMSE otherwise. In logistic regression models, estimates using weight-smoothing models with Laplace priors are robust, but with less gain in efficiency than in linear regression settings.

Key words: Weight trimming; winsorization; Bayesian finite population inference; Hierarchical models.

1. Introduction

Studies based on data sampled with unequal inclusion probabilities typically apply case weights equal to the inverse of the probability of inclusion to reduce or remove bias in estimators of descriptive population quantities, such as means or totals (Horvitz and Thompson 1952). This “fully weighted” approach can be extended to estimate analytical quantities that focus on association between risk factors and outcomes, such as population slopes in linear and generalized linear models, by applying sampling weights to score equations, and solving for the resulting “pseudo-maximum likelihood” estimators (PMLEs) (Binder 1983; Pfeffermann 1993). Unweighted and weighted estimators generally correspond when the underlying model (either implicit or explicit) is correctly specified and the sampling scheme is noninformative. When the model is misspecified or the sampling

[1] Dept. of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48109, USA. Email: xiaxi@umich.edu.
[2] Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48109, USA. Email: mrelliot@umich.edu.

Acknowledgments: We thank the Editor, the Associate Editor, and three reviewers for their careful review that helped to improve the quality of the article. This work was supported in part by Grant Number R01CA129101 from the National Cancer Institute.

scheme is informative, weighted estimators typically reduce bias, often (although not always) at the cost of increased variance. As model assumptions improve and/or sampling better approximates noninformativeness, the increase in variance from weighted analysis could overwhelm the reduction in bias, leading to an overall larger mean square error (MSE) than would be the case if the weights were ignored or at least controlled in some fashion.

In many, if not most cases, fully weighted estimators are used without concerns about such tradeoffs. When variability in weights is of concern, weight trimming, or "winsorization," is used to control the variation in weights by capping the weights at some value w_0, and redistributing the values above w_0 among the rest (Alexander et al. 1997; Kish 1992; Potter 1990). Various criteria have been used to determine the cap value based on data. Some examples include the National Assessment of Education Progress (NAEP) method by Potter (1988), which set the cutoff point equal to $\sqrt{c\sum_{i\in s} w_i^2/n}$, where c was chosen in an ad-hoc manner. Cox and McGrath (1981) approached it by estimating the cutoff point value which optimizes the empirical MSE estimated by $\widehat{MSE}(\hat{\theta}_t) = (\hat{\theta}_t - \hat{\theta}_w)^2 - \widehat{Var}(\hat{\theta}_t) + 2\sqrt{\widehat{Var}(\hat{\theta}_t)\widehat{Var}(\hat{\theta}_w)}$, where $\hat{\theta}_w$ is the fully weighted estimator, and $\hat{\theta}_t$, $t = 1, \ldots, T$, is the weight-trimmed estimator, with t denoting various trimming levels ordered from lowest to highest. These levels are ad hoc, except for $w_0 = \bar{w}$ for $t = 1$, which fixes all weights to their mean value and yields the unweighted estimator, and $w_0 = \max w_i$ for $t = T$, which yields the fully weighted (untrimmed) estimator. Chowdhury et al. (2007) suggested treating the weights as coming from a skewed cumulative distribution (e.g., an exponential distribution), and using the upper one percent of the fitted distribution as a cut point for weight trimming. Beaumont (2008) proposed a generalized design-based method, replacing the actual weights with weights predicted using a function of response and design variables. Details of these design-based approaches are summarized in Henry and Valliant's (2012) review.

An alternative to standard design-based weighted estimation is a model-based approach that accommodates disproportional probability-of-selection design in a finite population Bayesian inference setting. By creating dummy variables stratified by equal or approximately equal case weights, a fully weighted data analysis is obtained by building a model with indicators for the weight strata together with interaction terms between the weight stratum indicators and model parameters of interest, then obtaining inference about the population quantity of interest from its posterior predictive distribution. Elliott and Little (2000) established two model-based approaches for weight-trimming: model averaging, or "weight pooling", and hierarchical modeling, or "weight smoothing". A weight pooling model collapses strata with similar weights together with their associated interaction terms, mimicking a data-driven weight-trimming process. Weight smoothing treats the underlying weight strata as random effects, and achieves a balance between fully weighted and unweighted estimates using a shrinkage estimator: thus the weight strata are smoothed if data provide little evidence of difference between strata, and are separated if data suggest that interactions with strata are present. Under a Bayesian framework, a two-level model is implemented, assigning a multivariate normal prior for the random effects, with inference obtained from the posterior predictive distribution of the population parameter of interest. Elliott (2007) extended the application of weight-smoothing models to linear and generalized linear models, and discussed different structures for the random-effect priors, namely exchangeable, autoregressive, linear and nonparametric random slopes. Both of these papers found that there was an efficient/robustness tradeoff with respect to structure in the prior mean and variance, with simple exchangeable models providing highly efficient estimates when the weights provided little bias correction, but being susceptible to "oversmoothing" and yielding biased estimators when weights really were necessary to provide substantial bias correction. Highly structured models such as smoothing splines provided robust bias correction, but had less dramatic gains in efficiency and were more complex to implement. Hence we are motivated to find alternative models that will induce weight smoothing under simple mean and covariance matrix settings, while improving the bias-variance tradeoff. A logical choice would be a Laplace prior, which can be viewed as providing a Bayesian version of a LASSO regression model (Park and Casella 2008). This heavier-tailed prior might be expected to provide little or no shrinkage when bias correction is required, but still allow approximation to an unweighted estimator when the data suggest weak relationships with probability of selection.

In this article we extend the weight-smoothing approach by use of Laplace priors for the random-effect weight strata and interaction terms instead of multivariate normal priors, in order to achieve more robustness against "oversmoothing" in settings where weights are required to accommodate model misspecification or nonignorable sampling. We evaluate the performance of our proposed model in a simulation study, under both model misspecification and informative sampling, for both continuous and dichotomous outcomes, and compare it with competing methods. The article is organized as follows. In Section 2 we review the theory of model smoothing together with recently proposed model-assisted methods, and develop our model with Laplace priors. Section 3 provides simulation studies, and compares bias, coverage and MSE of the proposed method with competing methods. Section 4 demonstrates the method's performance for both linear and logistic scenarios by applications to dioxin data from the National Health and Nutrition Examination Survey (NHANES) and Partners for Child Passenger Safety dataset. Section 5 provides a summary discussion.

2. Weight-Smoothing Methodology

2.1. Finite Bayesian Population Inference

For finite Bayesian population inference, we model the population data Y: $Y \sim f(Y|\theta,Z)$, where Z are the variables associated with the sample design (probabilities of selection, cluster indicators, stratum variables). Note that the parametric model f can either be highly parametric with a low dimension θ (e.g., a normal model with common mean and variance), or have a more semiparametric or nonparametric flavor with a high-dimension θ (such as a spline or Dirichlet process model). Inference about some population quantity of interest $Q(Y)$ is based on the posterior predictive distribution of

$$p(Y_{nob}|Y_{obs},I,Z)=\frac{\int\int p(I|Y,Z,\theta,\phi)p(Y_{nob}|Y_{obs},Z,\theta,\phi)p(Y_{obs}|Z,\theta)p(\theta,\phi)d\theta d\phi}{\int\int\int p(I|Y,Z,\theta,\phi)p(Y_{nob}|Y_{obs},Z,\theta,\phi)p(Y_{obs}|Z,\theta)p(\theta,\phi)d\theta d\phi dY_{nobs}} \quad (1)$$

where Y_{nob} consists of the $N-n$ unobserved cases in the population, θ models Y (possibly conditional on Z) and ϕ models the inclusion indicator I (equal to 1 if the unit is sampled and observed and 0 otherwise). Thus $p(I\,|\,Y, Z, \theta, \phi)$ refers to the distribution of the sample mechanism given the design variables and the data of interest, $p(Y_{nob}\,|\,Y_{obs}, Z, \theta, \phi)$ gives the distribution of the unobserved elements in the sample given the observed elements in the sample and the (fully observed) design variables, and $p(Y_{obs}\,|\,Z, \theta)$ models the observed data given any design variables. Assuming that ϕ and θ have independent priors, the sampling mechanism is said to be "noninformative" if the distribution of I is independent of $Y|Z$, or "ignorable" if the distribution of I only depends on $Y_{obs}|Z$. When the sampling design is ignorable, $p(I\,|\,Y, Z, \theta, \phi) = p(I\,|\,Y_{obs}, Z, \phi)$, and thus (1) reduces to

$$p(Y_{nob}|Y_{obs},Z)=\frac{\int p(Y_{nob}|Y_{obs},Z,\theta)p(Y_{obs}|Z,\theta)p(\theta)d\theta}{\int\int p(Y_{nob}|Y_{obs},Z,\theta)p(Y_{obs}|Z,\theta)p(\theta)d\theta dY_{nobs}},$$

allowing inference about $Q(Y)$ to be made without explicitly modeling the sampling inclusion parameter I (Ericson 1969; Holt and Smith 1979; Little 1993; Rubin 1987; Skinner et al. 1989). Note that if inference about quantities $Q(Y|X)$ involving covariates X is desired (e.g., regression slopes), noninformative or ignorable sample designs can be relaxed to have the distribution of I depend on X.

2.2. *Weight Modeling*

Beaumont (2008) proposed an alternative model-assisted method, tamping down the extreme values in weights by replacing weights with values from a prediction model of weights regressed on response and design variables. Denote $I = (I_1, \ldots, I_N)^T$ as the vector of sample inclusion indicators, that is, $I_i = 1$ as ith unit sampled and $I_i = 0$ otherwise, $Y = (Y_1, \ldots, Y_N)^T$ the vector of survey-response variables, and $Z = (Z_1, \ldots, Z_N)^T$ the vector of design variables. Assuming a noninformative sampling design, thus $P(I|Z, Y) = P(I|Z)$, the predicted weights are obtained by $\tilde{w}_i = E_M(w_i|I_i = 1, y_i)$, where M refers to the expectation of w_i under a given model. Beaumont discussed two models, the linear form $E_M(w_i|I, Y) = H_i^T\beta + v_i^{1/2}\epsilon_i$, and the exponential form, $E_M(w_i|I, Y) = 1 + exp\left(H_i^T\beta + v_i^{1/2}\epsilon_i\right)$, where H_i and $v_i > 0$ are known functions of y_i. (The exponential form prevents the predicted weights from being negative.) He presented two examples of $H_i^T\beta$, one-degree polynomial and five-degree polynomial of y_i. The predicted weights are obtained by fitting the (unweighted) model on the sampled data, then the reweighted estimator of the survey-response variable of interest is obtained using the predicted weights. Extensions to regression settings can consider models of the form $E_M(w_i|I, Y, X) = J_i^T\beta + u_i^{1/2}\epsilon_i$, where J_i and u_i are functions of Y_i and X_i (possibly including interactions).

2.3. *Weight Smoothing*

In general, weight smoothing stratifies the data by inclusion probability, and applies a hierarchical model treating strata means as random effects, thus achieving trimming via shrinkage. Considering the population mean as the quantity of interest, a general weight-smoothing model is as follows:

$$Y_{hi} \overset{iid}{\sim} N(\mu_h, \sigma^2)$$

$$\mu \sim N_H(\phi, G)$$

where $\mu = (\mu_1, \ldots, \mu_H)$, $\phi = (\phi_1, \ldots, \phi_H)$, and $h = 1, \ldots, H$ indexes different "weight strata" defined, for example, by same or similar inclusion probabilities. In the case of stratified or poststratified sample designs, h indexes the actual strata. In more general designs, subjects can be formed into strata with equal or similar weights. We assume ϕ, G, and σ^2 all have weak or noninformative priors. Typically these strata are ordered from the smallest weight (highest probability of selection) to the largest weight (lowest probability of selection), but this is not required if a more natural ordering is available, for example, if the weight strata represent a disproportionately stratified sample by age. Based on this model, the posterior mean of the population mean is derived as:

$$E(\bar{Y}|y) = \sum_{h=1}^{H} [n_h \bar{y}_h + (N_h - n_h)\hat{\mu}_h]/N$$

where N_h and n_h are the population and sample sizes in stratum h, respectively, and $\hat{\mu}_h = E(\mu_h|y)$. Various assumptions can be made for the prior distribution of μ, such as

Exchangeable random effect (XRE): $\phi_h = \phi_0$ for all h, $G = \tau^2 I_H$
Autoregressive (AR1): $\phi_h = \phi_0$ for all h, $G = \tau^2 A$, $A_{jk} = \rho^{|j-k|}$, $j,k = 1, \ldots, H$
Linear (LIN): $\phi_h = \phi_0 + \phi' * h$, $G = \tau^2 I_H$
Nonparametric (NPAR): $\phi_h = g(h)$, $G = 0$ where g is an unspecified, twice-differentiable function.
See Elliott and Little (2000) for a detailed review.

The weight-smoothing mechanism can be easily intuited in the simplest case of the exchangeable random-effect (XRE) model (Holt and Smith 1979; Ghosh and Meeden 1986; Little 1991; Lazzaroni and Little 1998), where $\phi_h = \mu$ for all h, and $G = \tau^2 I_H$. The estimation of $\hat{\mu}_h$ is now a shrinkage estimator as $\hat{\mu}_h = w_h \bar{y}_h + (1 - w_h)\tilde{y}$, for $w_h = \tau^2 n_h / (\tau^2 n_h + \sigma^2)$ and $\tilde{y} = \left(\sum_h n_h/(n_h\tau^2 + \sigma^2)\right)^{-1} \sum_h n_h/(n_h\tau^2 + \sigma^2)\bar{y}_h$. As $\tau^2 \to \infty$, $w_h \to 1$, and $E(\bar{Y}|y) = \sum_{h=1}^{H} [n_h \bar{y}_h + (N_h - n_h)\bar{y}_h]/N = \sum_{h=1}^{H} (N_h/N)\bar{y}_h$, the fully weighted estimator. On the other hand, as $\tau^2 \to 0$, $w_h \to 0$, and the estimation shrinks toward the unweighted mean: since $\tilde{y} = \frac{\sum_h n_h \bar{y}_h/\sigma^2}{\sum_h n_h/\sigma^2} = \bar{y}$ if $\tau^2 = 0$, $E(\bar{Y}|y) = \sum_{h=1}^{H} [n_h \bar{y}_h + (N_h - n_h)\bar{y}]/N = (n/N)\bar{y} + \bar{y}(1 - n/N) = \bar{y}$ if $\tau^2 = 0$. Since τ^2 is itself estimated from the data, and is a measure of the information available to distinguish how the population means within a weight strata differ, the weight-smoothing model achieves a "data-driven" compromise between the weighted estimator, which is design consistent but may be highly inefficient, and the unweighted estimator, which is fully efficient when the assumption of independence between inclusion probability and mean of Y holds, but is likely biased otherwise.

2.4. *Weight Smoothing for Linear and Generalized Linear Regression Models*

Generalized linear regression models (McCullagh and Nelder 1989) postulate a likelihood for y_i of the form

$$f(y_i|\theta_i, \sigma) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\sigma)} + c(y_i, \sigma)\right]$$

where $a_i(\sigma)$ is a known function of (nuisance) scale parameter σ, and the mean of y_i given by $\mu_i = b'(\theta_i)$ is based on a linear combination of fixed covariates x_i through some link function g such that $E(y_i|\theta_i) = \mu_i$, and $g(\mu_i) = g(b'(\theta_i)) = \eta_i = x_i^T\beta$. In the meantime,

$Var(y_i|\theta_i) = a_i(\sigma)V(\mu_i)$, where $V(\mu_i) = b''(\theta_i)$; thus the variance is usually a function of the mean, with the exception of normal distribution, for which $b''(\theta_i) = 1$. The link is considered canonical if $\theta_i = \eta_i$, with the simplifying result that $V(\mu_i) = 1/g'(\mu_i)$. Some examples include Gaussian (linear) regression, where $a_i(\sigma) = \sigma^2$ and the canonical link is $g(\mu_i) = \mu_i$; logistic regression, where $a_i(\sigma) = n_i^{-1}$ and the canonical link is $g(\mu_i) = log(\mu_i/(1 - \mu_i))$, and Poisson regression, where $a_i(\sigma) = 1$ and the canonical link is $g(\mu_i) = log(\mu_i)$.

When considering weighted estimators, we index by the inclusion stratum h, thus $g(E[y_{hi}|\beta_h]) = x_{hi}^T\beta_h$. For weight-smoothing models, the hierarchical structure is considered as

$$\left(\beta_1^T, \ldots, \beta_H^T\right)^T | \beta^*, G \sim N_{HP}(\beta^*, G)$$

where β^* is an unknown vector of mean values for the regression coefficients and G is an unknown covariance matrix. Our interest is to estimate the target-population quantity $B = (B_1, \ldots, B_p)^T$, which is the slope that solves the population score equation $U_N(B) = 0$ where

$$U_N(\beta) = \sum_{i=1}^{N} \frac{\partial}{\partial \beta} log f(y_i; \beta) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{y_{hi} - g^{-1}(\mu_i(\beta))x_{hi}}{V(\mu_{hi}(\beta))g'(\mu_{hi}(\beta))}$$

Notice that the quantity B that satisfies $U(B) = 0$ is always a meaningful population quantity even if the model is misspecified, since it is a linear approximation of x_i to η_i. A first-order approximation of $E(B|y, X)$ is given based on $\hat{B}$ where

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} \frac{(\hat{y}_{hi} - g^{-1}(\mu_i(\hat{B})))x_{hi}}{V(\mu_{hi}(\hat{B}))g'(\mu_{hi}(\hat{B}))} = 0$$

where $W_h = N_h/n_h$, $\hat{y}_{hi} = g^{-1}\left(x_{hi}^T\hat{\beta}_h\right)$, and $\hat{\beta}_h = E(\beta_h|y, X)$. For linear regression, where $V(\mu_i) = \sigma^2$ and $g'(\mu_i) = 1$,

$$\hat{B} = E(B|y, X) = \left[\sum_h W_h \sum_{i=1}^{n_h} x_{hi}x_{hi}^{'}\right]^{-1} \left[\sum_h W_h \left(\sum_{i=1}^{n_h} x_{hi}x_{hi}^{'}\right)\hat{\beta}_h\right].$$

In case of logistic regression, $V(\mu_i) = \mu_i(1 - \mu_i)$ and $g'(\mu_i) = \mu_i^{-1}(1 - \mu_i)^{-1}$, and $E(B|y, X)$ is obtained by solving the weighted score equation for population regression parameter B

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hi}\left(expit\left(x_{hi}^{'}B\right) - expit\left(x_{hi}^{'}\hat{\beta}_h\right)\right) = 0$$

where $expit(.) = exp(.)/(1 + exp(.))$.

2.5. Laplace Prior for Weight Smoothing

Instead of using a multivariate normal distribution as the prior of βs, we propose using a multivariate Laplace distribution. Unlike the normal distribution prior which restricts the variation between random-effect term and prior mean in an L^2-norm manner, Laplace measures by the L^1 distance. This should allow more severe downweighting of interactions between the regression parameters and the probability of selection for which there is only weak evidence in the data, while preserving those that we need for bias correction.

The general form of multivariate Laplace distribution is given by Eltoft et al. (2006):

$$p_Y(y) = \frac{1}{(2\pi)^{d/2}} \frac{2}{\lambda} \frac{K_{(d/2)-1}\left(\sqrt{\frac{2}{\lambda}q(y)}\right)}{\left(\sqrt{\frac{2}{\lambda}q(y)}\right)^{(d/2)-1}}$$

where y is a d-dimensional random variable $y = (y_1, \ldots, y_d)$; $K_m(x)$ denotes the modified Bessel function of the second kind and order m, evaluated at x; $q(y) = (y - \mu)^t \Gamma^{-1} (y - \mu)$; $\Gamma = \{\gamma_{jk}\}$, $j, k = 1, \ldots, d$ is a $d \times d$ matrix defining the internal covariance structure of the variable Y, $\mu = (\mu_1, \ldots, \mu_d)$ is the vector of means, and λ an overall scale parameter. However, this format is inconvenient for application. The alternative approach is to represent the Laplace distribution as a scale mixture of normals with an exponential mixing density:

$$\beta_h | \beta_h^*, D_\tau, \sigma^2 \sim MVN\left(\beta_h^*, \sigma^2 D_{\tau h}\right)$$

$$\beta_h^* | \sigma_0^2 \sim MVN\left(0, \sigma_0^2 I_p\right)$$

$$D_{\tau h} = diag\left(\tau_{h1}^2, \ldots, \tau_{hp}^2\right)$$

$$\sigma^2, \tau_1^2, \ldots, \tau_{Hp}^2 \sim 1/\sigma^2 \prod_{j=1}^{Hp} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2 / 2}$$

$$\lambda^2 \sim Gamma\,(\gamma, \delta)$$

where γ and δ are known constants. The first level of the model depends on the distribution assumption of the generalized linear model used. In this article, we take linear regression and logistic regression as examples, and provide the full hierarchical Bayesian model and related Gibbs Sampler algorithm.

For linear regression, Y conditional on all other parameters follows a normal distribution. Assuming that the residual variance σ^2 is independent from the latent mixing variables τ_i, the hierarchical model is as follows:

$$y_{hi} | x_{hi}, \beta_h, \sigma^2 \overset{ind}{\sim} N\left(x_{hi}^T \beta_h, \sigma^2\right)$$

$$\beta_h | \beta_h^*, D_\tau, \sigma^2 \overset{ind}{\sim} MVN\left(\beta_h^*, \sigma^2 D_{\tau h}\right)$$

$$\beta_h^* | \sigma_0^2 \overset{ind}{\sim} MVN\left(0, \sigma_0^2 I_p\right)$$

$$D_{\tau h} = diag\left(\tau_{h1}^2, \ldots, \tau_{hp}^2\right)$$

$$\sigma^2, \tau_1^2, \ldots, \tau_{Hp}^2 \sim 1/\sigma^2 \prod_{j=1}^{Hp} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2 / 2}$$

$$\lambda^2 \sim Gamma(\gamma, \delta)$$

where again γ and δ are known constants. Following the deduction in Park and Casella (2008), the analytical forms of all fully conditional distributions of β, σ^2 and so on exist, and the posterior predictive distribution could be obtained through a Gibbs Sampler as below. A detailed derivation is available online at www.doi.org/10.1515/jos-2016-0026, Appendix A.

$$\beta_h | rest \sim MVN\left(A^{-1}\left(X_h^T Y_h + D_{\tau h}^{-1}\beta_h^*\right), \sigma^2 A^{-1}\right), A = X_h^T X_h + D_{\tau h}^{-1}$$

$$\beta_h^* | rest \sim MVN\left((\sigma^2 D_{\tau h})^{-1}\left((\sigma^2 D_{\tau h})^{-1} + \left(\sigma_0^2 I\right)^{-1}\right)^{-1}\beta_h, \left((\sigma^2 D_{\tau h})^{-1} + \left(\sigma_0^2 I\right)^{-1}\right)^{-1}\right)$$

$$\sigma^2 | rest \sim InvGamma\left((n + Hp)/2, \frac{1}{2}\left[\sum_{h=1}^{H}(Y_h - X_h\beta_h)^T(Y_h - X_h\beta_h)\right.\right.$$

$$\left.\left. + \sum_{h=1}^{H}\left(\beta_h - \beta_h^*\right)^T(D_{\tau h})^{-1}\left(\beta_h - \beta_h^*\right)\right]\right)$$

$$1/\tau_{hi}^2 | rest \sim InvGaussian\left(\sqrt{\frac{\lambda^2\sigma^2}{\left(\beta_h - \beta_h^*\right)^2}}, \lambda^2\right)$$

$$\lambda^2 \sim Gamma\left(Hp + \gamma, \frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p}\tau_{hi}^2 + \delta\right)$$

At each step in the Gibbs Sampler chain, a draw of the linear regression population slope B is obtained from the draw of β_h as $B = \left[\sum_h W_h \sum_{i=1}^{n_h} x_{hi} x_{hi}^{'}\right]^{-1} \left[\sum_h W_h \left(\sum_{i=1}^{n_h} x_{hi} x_{hi}^{'}\right)\beta_h\right]$.

For logistic regression, the model is similar to that for linear regression, except that Y follows a binomial distribution, and estimation of σ^2 is no longer necessary:

$$y_{hi} | x_{hi}, \beta_h, \sim \prod_{h=1}^{H}\prod_{i=1}^{n_h}\left(\frac{exp(x_{hi}\beta_h)}{1 + exp(x_{hi}\beta_h)}\right)^{y_{hi}}\left(\frac{1}{1 + exp(x_{hi}\beta_h)}\right)^{1-y_{hi}}$$

$$\beta_h | \beta_h^*, D_\tau, \sim MVN(\beta_h^*, D_{\tau h})$$

$$\beta_h^* | \sigma_0^2 \sim MVN(0, \sigma_0^2 I_p)$$

$$D_{\tau h} = diag(\tau_{h1}^2, \ldots, \tau_{hp}^2)$$

$$\tau_1^2, \ldots, \tau_{Hp}^2 \sim \prod_{j=1}^{Hp}\frac{\lambda^2}{2}e^{-\lambda^2\tau_j^2/2}$$

$$\lambda^2 \sim Gamma(\gamma, \delta)$$

where γ and δ are known constants. When the first level is not normally distributed, the fully conditional distribution of β does not belong to any known distribution, and thus direct sampling is impossible. Instead we apply a Metropolis method, and the proposed β_h is drawn from $N_p(\beta_h^{'}, c_\beta D_\beta)$, for $D_\beta = \left(V_{\beta h}^{-1} + D_{\tau h}^{-1}\right)^{-1}$, where $\beta_h^{'}$ is the ML estimate of the logistic regression of y on z from strata h, and $V_\beta h$ the associated covariance matrix obtained from the expected information matrix evaluated at β_h. The proposed β_h is accepted with probability $r = max[1,\{f_\beta(\beta_{prop})\}/\{f_\beta(\beta)\}]$, where f_β is the posterior distribution of β proportional to $p(\beta_h)\prod_{i=1}^{n_h} f(y_{hi}|\beta_h)$. All other parameters follow the Gibbs Sampler algorithm, and are directly drawn from their fully conditional distributions as below (full derivation is available online at www.doi.org/10.1515/jos-2016-0026, Appendix B):

$$\beta_h^*|rest \sim MVN\left((D_{\tau h})^{-1}\left((D_{\tau h})^{-1} + \left(\sigma_0^2 I\right)^{-1}\right)^{-1}\beta_h, \left((D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1}\right)^{-1}\right)$$

$$1/\tau_{hi}^2|rest \sim InvGaussian\left(\sqrt{\frac{\lambda^2}{\left(\beta_h - \beta_h^*\right)^2}}, \lambda^2\right)$$

$$\lambda^2 \sim Gamma\left(Hp + \gamma, \frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p}\tau_{hi}^2 + \delta\right)$$

At each step in the Gibbs Sampler chain, a draw of the logistic regression population slope B is obtained from the draw of β_h by solving the weighted score equation $\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hi}\left(expit\left(x_{hi}^{'}B\right) - expit\left(x_{hi}^{'}\beta_h\right)\right) = 0$. In practice, B can be obtained by replacing the observed y_{hi} with the predicted values $g\left(x_{hi}^{'}\beta_h\right)$ for each draw of $\hat{\beta}_h$ and obtaining the weighted pseudo-MLE for the logistic regression model.

3. Simulation Study

To evaluate the performance of weight-smoothing models using Laplace priors, we create several scenarios for ordinary linear regression and logistic regression, generating separate populations with normally distributed outcomes and dichotomized outcomes accordingly. We also consider scenarios where heteroscedasticity or multiple covariates occur. The target of interest is the population slope in a regression model. In addition to our Laplace prior estimator, we include an unweighted estimator, a fully weighted estimator, a normal-prior (exchangable) estimator (Elliott and Little 2000; Elliott 2007), and several variations of the estimator proposed by Beaumont (2008) for comparison. For each scenario and estimator, we compute bias, square root of mean square error (RMSE) and coverage of 95% confidence or credible intervals as follows:

$$\text{bias} = S^{-1}\sum_{s=1}^{s}(\hat{B}s - B)$$

$$\text{RMSE} = \sqrt{S^{-1}\sum_{s=1}^{s}(\hat{B}s - B)^2}$$

$$\text{coverage} = S^{-1}\sum_{s=1}^{s} I\left(\hat{B}_s^L \leq B \leq \hat{B}_s^U\right)$$

where s indexes the independent samples drawn for each simulation, $\hat{B}_s$ is the point estimator of the regression coefficient of interest, $\hat{B}_s^L$ and $\hat{B}_s^U$ correspond to the lower and upper bounds of the 95% confidence or credible interval, and B to the regression coefficient computed using the population data (i.e., the inference target of interest).

Relative root mean square error (RRMSE) is reported as the ratio of the estimator's RMSE to the fully weighted estimator's RMSE.

3.1. Hierarchical Weight Smoothing Model for Ordinary Linear Regression

We first generate a population of size $N = 20{,}000$. The predictor X is uniformly distributed between zero and ten, and is equally divided into 20 strata with intervals of 0.5 each. The response variable Y is generated as a spline function of X noted below, with knots located between strata. Three sets of coefficients of the spline are applied separately to represent the various patterns of $Y|X$ from straight slope to accelerating and decelerating curves.

$$Y_i|X_i, \beta, \sigma^2 \sim N\left(\beta_0 + \sum_{h=1}^{20} \beta_h (x_i - h/2)_+, \sigma^2\right)$$

$$X_i \sim UNI(0, 10), i = 1, \ldots, N = 20{,}000$$

We assume $\beta_0 = 0$, and consider three sets of the spline coefficients

$$\beta_a = (2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\beta_b = (0, 0, 0, 0, 0, 0, .25, .25, .25, .25, .5, .5, .5, .5, 1, 1, 1, 1, 1, 1)$$

$$\beta_c = (11, -1, -1, -1, -1, -1, -1, -0.5, -0.5, -0.5, -0.5, -0.25,$$
$$-0.25, -0.25, -0.25, 0, 0, 0, 0, 0).$$

We let the population variance σ^2 vary across $10^{1.5}$, $10^{3.5}$ and $10^{5.5}$ to create varying levels of background noise compared to changes in slope. Figure 1 shows each of the $3 \times 3 = 9$ populations generated for the linear regression.

From the population, samples of size $n = 1{,}000$ are repeatedly drawn without replacement, according to inclusion probabilities proportional to $\pi_h = (1 + h/30) * h$ for the hth stratum, which results in a ratio of 35 times between the maximum and minimum probabilities. We also ensure that the sample size of each stratum is greater than three for computation convenience. Z is created as $Z = I \otimes X$, where $I = c(I_1, \ldots, I_h)$ is an indicator vector marking that the current observation belongs to the ith stratum. It is also centered within each column with respect to each stratum, and used as predictor in the simulations. Thus β_a corresponds to a linear model (no model misspecification); β_b to a setting where the nonlinearity is greatest where the probability of selection is the highest, and β_c to a setting where the nonlinearity is greatest where the probability of selection is the lowest.

Our inferential target is $B = \left(\sum_{i=1}^{N} \tilde{X}_i \tilde{X}_i^{'}\right)^{-1} \sum_{i=1}^{N} \tilde{X}_i Y_i$ for $\tilde{X}_i = (1 \;\; X_i)^{'}$, the least-squares linear approximation of Y to X. Under β_b and β_c, weights correct bias from model misspecification. Under β_a, the model is correctly specified, suggesting that the unweighted estimator is most efficient. Also note that under β_b, the curvature is largest where the data are most densely sampled, while the reverse is true under β_c, suggesting that varying degrees of trimming will be required to optimize the bias-variance tradeoff.

For the hyperprior parameters, σ_0^2 is arbitrarily defined as 1,000 to approximate a noninformative prior; the prior for λ follows a gamma hyperprior with parameter $\gamma = 1$ and $\delta = 1.78$, as suggested by Park and Casella (2008). All other parameters in simulation are initialized at zero, except for the variance estimator σ^2, which is initialized at one.

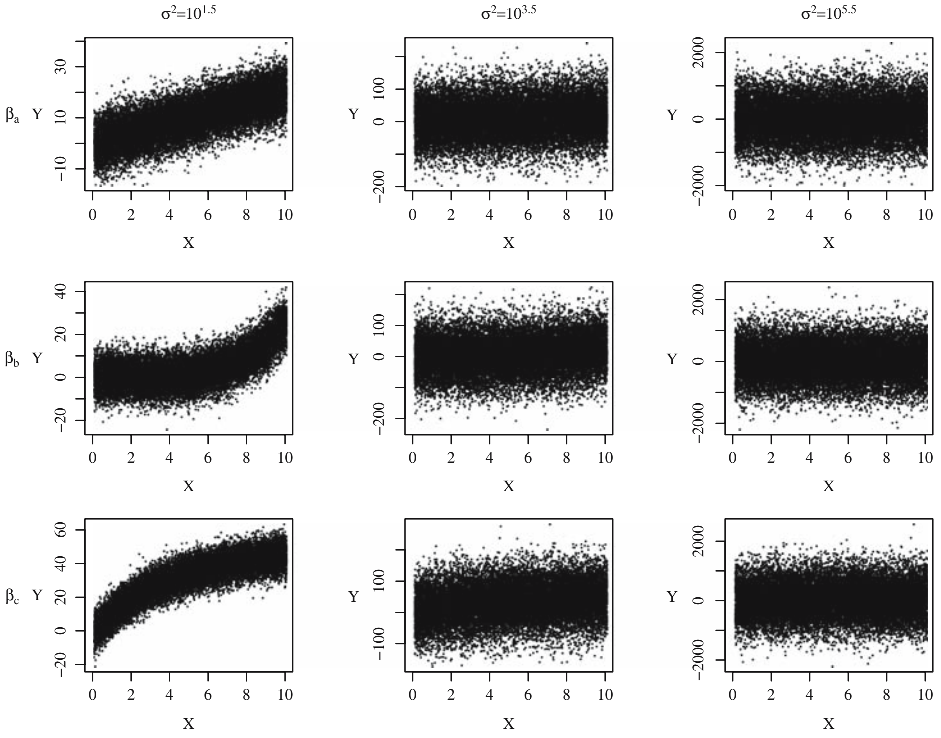


Fig. 1. Scatter plot of population for the linear regression simulation

A Gibbs Sampler method is applied, that is, all parameters are sequentially drawn from the full conditional distribution for each iteration. Then, to obtain the estimate from the posterior predictive distribution, the unobserved Y are generated based on sampled parameters from each iteration, and the target population slope B is obtained by fully weighted regression on observed and predicted Y. The process iterates 10,000 times, with a burn-in of 2,000. Diagnostic plots are generated to ensure the algorithm's convergence via visual inspection. Overall, 200 samples are generated from each population to provide the empirical distribution for the repeated measures properties.

We compare the properties of our Laplace model (HWS) with major competitors, including the unweighted model (UNWT), fully weighted model (FWT), weight-smoothing model with normal prior and exchangeable random-slope assumption (XRS), and two variations of the estimators proposed by Beaumont (2008): predicted weights on y and x (PREDYX) and predicted weights on degree 5 polynomial of y, together with x (PREDYX5). Bias and nominal 95% coverage are recorded directly, while RMSE is rescaled relative to the fully weighted estimator. Results are provided in Tables 1, 2, and 3.

Under β_a, where the model is correctly specified, all methods yield unbiased results, and the unweighted estimator maintains the best efficiency, with an approximate 30% decrease in RMSE compared to the fully weighted estimator. The original weight-smoothing method under XRS tends to provide unstable results, inflating the variance when the population signal is strong, but achieving similar RMSE as the unweighted estimator when the population signal is weak relative to the noise. Our model, under the same XRS

Table 1. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the linear regression coefficient B under β_a using populations with different residual variances corresponding to various models under consideration

	$\sigma^2 = 10^{1.5}$			$\sigma^2 = 10^{3.5}$			$\sigma^2 = 10^{5.5}$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
UNWT	−0.022	0.684	0.95	−0.079	0.637	0.97	0.156	0.644	0.95
FWT	−0.011	1	0.94	0.093	1	0.95	−0.222	1	0.93
HWS	−0.011	1.025	0.96	0.082	0.859	0.98	−0.906	0.741	0.99
XRS	−0.012	1.278	0.99	−0.062	0.678	0.96	−0.166	0.651	0.96
PREDYX	−0.011	0.791	0.97	0.083	0.783	0.95	−0.375	0.756	0.94
PREDYX5	−0.011	0.952	0.94	0.078	0.944	0.94	−1.089	0.960	0.94

Table 2. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the linear regression coefficient B under β_b using populations with different residual variances corresponding to various models under consideration

	$\sigma^2 = 10^{1.5}$			$\sigma^2 = 10^{3.5}$			$\sigma^2 = 10^{5.5}$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
UNWT	0.939	5.316	0	0.876	0.807	0.80	2.481	0.666	0.96
FWT	−0.010	1	0.96	0.093	1	0.95	2.736	1	0.93
HWS	−0.011	0.887	0.97	0.050	0.864	0.98	1.047	0.747	1
XRS	−0.063	1.012	0.95	0.869	0.873	0.94	1.134	0.729	0.94
PREDYX	0.114	1.375	0.52	0.385	0.735	0.91	0.746	0.726	0.93
PREDYX5	0.025	1.000	1	0.022	0.955	0.95	1.515	0.974	0.95

Table 3. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the linear regression coefficient B under β_c using populations with different residual variances corresponding to various models under consideration

	$\sigma^2 = 10^{1.5}$			$\sigma^2 = 10^{3.5}$			$\sigma^2 = 10^{5.5}$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
UNWT	−0.925	4.517	0	−0.976	0.848	0.76	−0.738	0.643	0.95
FWT	−0.011	1	1	0.094	1	0.95	−0.225	1	0.93
HWS	−0.010	0.779	0.95	0.171	0.875	0.98	−0.876	0.743	0.99
XRS	−0.264	1.527	0.26	−0.953	1.053	0.64	−0.729	0.662	0.93
PREDYX	−0.236	1.301	0.65	−0.382	0.991	0.85	−0.616	0.759	0.92
PREDYX5	−0.010	1.018	1	−0.043	0.965	0.96	−0.552	0.974	0.95

assumption but with a Laplace prior, gives more stable results that resemble the fully weighted estimator when variance is low, but increase in efficiency as the population variance increases. Both the XRS and HWS estimators have correct to somewhat conservative coverage when the linear model is correctly specified. The Beaumont estimator PREDYX has improved RMSE compared to the fully weighted estimator, since it models the weight as a linear function of X; PREDYX5 yields estimates similar to the fully weighted estimator, as variance reduction was minimal due to the increased variability in the model weights.

For scenarios under β_b and β_c, the unweighted estimator of B is biased, and the fully weighted estimator strongly prevails over the unweighted estimator with respect to both RMSE and true coverage for small to moderate levels of residual variances. The weight-smoothing method under XRS remains biased at moderate levels of variance for β_b and β_c, and also at small levels of variance for β_c, raising RMSE relative to FWT and destroying nominal coverage, suggesting that the exchangeable random-slope structure is not able to capture the relation in mean and variance among strata. The weight-smoothing estimator with Laplace prior has limited bias similar to that of the fully weighted estimator, but very substantially reduced RMSE, with correct to conservative coverage. The PREDYX estimator is insufficiently structured to reduce bias in the small-to-medium residual-variance settings; PREDYX5 mimics the fully weighted estimator and thus has little savings in relative RMSE under any of the scenarios.

3.2. Hierarchical Weight Smoothing Model for Logistic Regression

Following Elliott (2007), we set up populations in two settings: model misspecification and informative sampling. For model misspecification, the population is equally divided into 20 strata, and the predictor X is uniformly distributed within each stratum on an interval ranging from $0.5(h-1)$ to $0.5h$. The binary response variable is generated as follows:

$$P(Y_i = 1|X_i) \sim BER(expit(1.5 - .75X_i + C{*}X_i^2)),$$

$$X_{hi} \sim UNI(0.5{*}(h-1), 0.5{*}h), h = 1, \ldots, 20, i = 1, \ldots, 1000$$

Our inferential target is $B = (B_0\ B_1)'$, the value of $\beta = (\beta_0\ \beta_1)'$ that solves the score equation $U(\beta) = \sum_{i=1}^{N} \tilde{X}_i(Y_i - expit(\tilde{X}_i' \beta))$, corresponding to the best linear approximation to X_i and $log\left(\frac{E(Y_i|X_i)}{1-E(Y_i|X_i)}\right)$. For C, we consider values of 0, .027, .045, .061, .080, corresponding to no model misspecification at $c = 0$ to increasing levels of model misspecification. The selection probability for each observation remains the same within each stratum, and increases linearly along strata, with a ratio between maximum and minimum probabilities equal to 20.

For the informative sampling setting, we follow the same formula of

$$P(Y_i = 1|X_i) \sim BER(expit(1.5 - .75X_i + C{*}X_i^2)),$$

$$X_{hi} \sim UNI(0.5{*}(h-1), 0.5{*}h), h = 1, \ldots, 20, i = 1, \ldots, 1000$$

but fix $C = 0$, so the model is correctly specified. We also create a vector of binary value Z_i^* such that $Cor(Y_i, Z_i^*) = r$, and let r range from 0.05 to 0.95 to represent different levels of correlation with Y. Then we let $Z_i = Z_i^* U_i + (1 - Z_i^*) X_i$, where $U_i \sim U(0, 10)$ independent of X_i, and the selection probability is proportional to Z_i. Thus whether the selection probability is related to X or not is determined by the value of Z^*, which is correlated with Y to some level. The process results in a ratio of roughly 30 between the maximum weight and minimum weight, with the correlation between selection probability and Y varying from 0 to 30% as the correlation between Z^* and Y increases from .05 to .95. Twenty strata of equal size are created by pooling observations with similar selection probabilities together.

From this population, samples with $n = 1{,}000$ are selected without replacement, with the selection probability stated above. We create weight strata using the values of h. A total of 200 samples are generated to create the empirical distribution for inference. A single MCMC chain is built for each data set, and for each iteration in the algorithm, all parameters are sequentially drawn from the full conditional distribution, except for β, which is drawn via a Metropolis step (proposed from a normal distribution centered at MLE with inverse expected information as covariance matrix, and accepted according to likelihood ratio times prior distribution). Then the predicted Y is calculated based on drawn parameters, and the target population slope is obtained by fully weighted logistic regression. The initial values of parameters are assigned the same as linear regression setting, and the process iterates 10,000 times, with a burn-in of 2,000.

We compare the properties of our Laplace model (HWS) with the same major competitors as in the linear regression setting. Bias and nominal 95% coverage are recorded directly, while RMSE is rescaled according to the fully weighted estimator. Results are provided in Table 4 and 5.

While comparing different models under the model misspecification settings, the unweighted model has increased bias as the population model is less correctly specified, resulting in a change from an efficient estimate to a poor estimate (RMSE ratio from 69.7% to 281.9% of FWTs as C increases) and poor coverage as misspecification increases. The exchangeable random-slope model estimator is not robust, with bias similar to the unweighted model, and larger RMSE than the fully weighted estimator, although coverage is conservative. The hierarchical weight smoothing model with Laplace prior provides a more robust estimator, with minimal bias, and RMSE reduced by up to 14% compared to the FWT estimator, with true coverage similar to that of the fully weighted estimator. The weight-prediction model PREDYX performs similarly to the unweighted estimator, gaining efficiency when the model is correctly specified, and suffering as misspecification increases. PREDYX5, which predicts weights with a degree-five polynomial of x, essentially mimics the fully weighted estimator.

Under informative sampling, the unweighted estimator has only slightly larger RMSE than the fully weighted estimator, but is substantially biased with poor coverage. The exchangeable random-effect model has a similar degree of bias compared to the unweighted estimator, but has increased variability that, while providing conservative coverage, yields substantially increased RMSE over the fully weighted estimator. The hierarchical weight-smoothing model with Laplace prior again provides a more robust estimator, with minimal bias, and RMSE reduced by up to twelve percent compared to the FWT estimator, although true coverage suffers to a moderate degree except when the

Table 4. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the logistic regression coefficient using populations with different degrees of model misspecification corresponding to various models under consideration

	$c = 0$			$c = .45$			$c = .80$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
UNWT	0.014	0.697	0.96	0.063	1.151	0.47	0.132	2.819	0
FWT	0.006	1	0.84	−0.015	1	0.82	−0.014	1	0.94
HWS	0.011	0.915	0.84	−0.014	0.909	0.82	−0.013	0.860	0.88
XRS	0.038	1.125	1	0.078	1.696	0.94	0.042	1.644	0.98
PREDYX	0.003	0.791	0.96	0.021	0.903	0.92	0.038	1.123	0.79
PREDYX5	−0.004	0.962	0.93	−0.005	0.965	0.94	−0.001	0.967	0.98

Table 5. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the logistic regression coefficient using populations with different degrees of informative sampling corresponding to various models under consideration

	$r = .05$			$r = .50$			$r = .95$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
UNWT	0.057	0.990	0.76	0.069	1.155	0.52	0.053	0.914	0.64
FWT	0.023	1	0.94	0.009	1	0.88	0.001	1	1
HWS	0.022	0.906	0.82	0.009	0.914	0.84	0.001	0.875	0.96
XRS	0.067	1.417	1	0.071	1.463	0.98	0.059	1.272	1
PREDYX	0.021	0.832	0.94	0.023	0.859	0.91	0.031	0.880	0.93
PREDYX5	0.004	0.977	0.94	0.002	0.969	0.95	0.007	0.997	0.97

sampling is highly informative. PREDYX improves RMSE by up to 17% while having only slight undercoverage. PREDYX5 again mimics the fully weighted model.

3.3. Hierarchical Weight-Smoothing Model for Heteroscedasticity Scenario

In this section we evaluate the performance of the hierarchical weight-smoothing model compared to unweighted and fully weighted models under the heteroscedasticity setting. Here the main purpose of weighting is not to correct bias, but to adjust for the violation in the homoscedasticity assumption, and to yield correct inference on target quantities. We expect the data-driven method to capture the heteroscedasticity pattern in data well, and lead to proper inference.

First, we create a population of size $N = 20{,}000$. The interval between zero and ten is evenly divided into ten strata with a length of one, and the predictor X is uniformly distributed within each stratum. The response variable Y is then generated from a normal distribution, with mean equal to twice of X, and variance as an increasing function of X:

$$X_i \sim UNI(0, 10), i = 1, \ldots, N = 20{,}000$$

$$P_i = (1 + \lceil X_i \rceil/30) * \lceil X_i \rceil/2$$

$$Y_i | X_i, \sigma^2 \sim N(2 * X_i, P_i * \sigma^2)$$

where the population variance σ^2 is set to 10^1, 10^3 and 10^5 to adjust for different scales of population variance.

We repeatedly draw samples from the population without replacement by inclusion probabilities proportional to P, to link the corresponding weights to the heteroscedasticity pattern. The inferential target remains the population slope; all other settings for parameters and simulations remain the same. Altogether 200 samples are drawn, and the HWT model is fit using 10,000 iterations, with 2,000 as burn-in. To evaluate the results, we compare bias, relative RMSE and true coverage of the nominal 95% confidence interval or credible interval across the unweighted model, fully weighted model, and hierarchical weight-smoothing model in Table 6.

The results suggest that violation in homoscedasticity undermines the performance of the unweighted estimator when the population variance is small. But as the population variance increases, this effect is quickly overwhelmed by concerns about efficiency, where the unweighted estimator has about a 35% reduction in RMSE compared to the fully weighted estimator. The hierarchical weight-smoothing model performs well in the heteroscedastic setting, where it correctly retains the weight interactions at low variances, yet "tunes them out" when efficiency is the dominating component.

3.4. Hierarchical Weight-Smoothing Model with Multiple Covariates

In the last simulation study we focus on the hierarchical weight-smoothing model's performance when multiple covariates exist in the model. The challenge lies in that some covariates may be related to the sampling scheme, and thus could benefit from weighting, but some may be independently distributed, and could lose efficiency in weighting. Simply applying the fully weighted or the unweighted model will sacrifice either variance for

Table 6. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators

	$\sigma^2 = 10$			$\sigma^2 = 10^3$			$\sigma^2 = 10^5$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
UNWT	−0.030	1.432	0.95	0.009	0.608	0.95	−0.766	0.679	0.96
FWT	−0.002	1	0.97	−0.012	1	0.93	−1.253	1	0.96
HWS	−0.004	1	1	−0.020	0.896	0.96	−1.112	0.653	0.99

covariate associations unrelated to the sampling scheme, or bias for covariate associations related to the sampling scheme; we expect our data-driven method could reach a balance in between these options.

To replicate such a situation, we look into three different scenarios: independent covariates X_1 and X_2 with model misspecification on X_1, and weight related to X_1; independent covariates X_1 and X_2 with model misspecification on X_1, and weight related to X_2; and correlated covariates X_1 and X_2, with model misspecification on X_1, and weight related to X_1.

For the first scenario, we generate a population of size $N = 20{,}000$, with two covariates X_1 and X_2 independently drawn from uniform distributions on an interval between zero and ten. The outcome Y is generated from a normal distribution with mean equal to a spline function of X_1 plus X_2, thus our linear approximation to X_1 leads to model misspecification.

$$Y_i | X_{1i}, X_{2i}, \beta, \sigma^2 \sim N(\beta_0 + x_{2i} + \sum_{h=1}^{10} \beta_h (x_{1i} - h)_+, \sigma^2)$$

$$X_{1i} \sim UNI(0,10), i = 1, \ldots, N = 20{,}000$$

$$X_{2i} \sim UNI(0,10), i = 1, \ldots, N = 20{,}000$$

$$\beta = (0, 0, 0.5, 0.5, 0.5, 0.5, 1, 1, 1, 2)$$

where the variance σ^2 is set to 10^1, 10^3, and 10^5 for variation in background noise.

The inclusion probabilities are proportional to $\pi_i = (1 + \lceil X_{1i} \rceil / 30) * \lceil X_{1i} \rceil / 2$, and samples of size $n = 1{,}000$ are repeatedly drawn. Weights equal to inverse inclusion probabilities naturally create ten strata. For computation convenience, we also create Z as $Z = I \otimes (X_1, X_2)$, where $I = c(I_1, \ldots, I_h)$ indicates to which stratum an observation belongs. We assess the model performance through its inference on the population slopes of Y to X_1 and X_2: $\tilde{B} = (B_0, B_1, B_2) = \left(\sum_{i=1}^{N} \tilde{X}_i \tilde{X}_i^{`} \right)^{-1} \sum_{i=1}^{N} \tilde{X}_i Y_i$ for $\tilde{X}_i = (1 \ \ X_{1i} \ \ X_{2i})'$.

When initializing the process, we retain all previous hyperprior parameters settings. The same Gibbs Sampler method is applied, and the predicted Y is generated to yield inference on target population slopes B_1 and B_2. The process iterates 10,000 times, with a burn-in of 2,000. We conduct 200 simulations to provide the empirical distribution to estimate repeated measurement properties. Table 7 compares bias, relative RMSE and true coverage of the nominal 95% confidence interval or credible interval across the hierarchical weight-smoothing method, unweighted model and fully weighted model.

The results in Table 7 suggest that when the population variance is small, the unweighted model result suffers from model misspecification and has a substantially larger bias compared to the fully weighted estimator. Yet for the covariate on which the model is correctly specified, the unweighted estimator gains efficiency and yields an RMSE about half the size of the RMSE of the weighted estimator. As background variance increases, the influence of bias decreases and the unweighted estimator prevails over the fully weighted estimator in estimating both X_1 and X_2. The hierarchical weight-smoothing model closely resembles the fully weighted estimator with minor improvements. That is, it

Table 7. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the multivariate linear regression coefficients B_1, B_2 when weight is related to X_1 using populations with different residual variances corresponding to various models under consideration

	$\sigma^2 = 10$			$\sigma^2 = 10^3$			$\sigma^2 = 10^5$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
B_1									
UNWT	0.976	6.592	0	0.947	1.293	0.55	2.747	0.697	0.99
FWT	−0.074	1	0.93	−0.385	1	0.97	1.519	1	0.93
HWS	−0.076	1.007	0.84	−0.377	0.999	0.96	1.442	0.992	0.99
B_2									
UNWT	−0.013	0.440	0.94	0.075	0.392	0.97	−1.754	0.490	0.93
FWT	−0.012	1	0.89	0.153	1	0.95	−0.656	1	0.89
HWS	−0.014	0.991	0.88	0.160	0.993	0.98	−0.841	0.990	0.91

correctly applies weight when model misspecification occurs, but fails to tune the result when efficiency is more important.

For the second scenario, we follow the same setting as the previous scenario, except that the inclusion probabilities are now proportional to $\pi_i = (1 + \lceil X_{2i} \rceil / 30) * \lceil X_{2i} \rceil / 2$. Since the weights are related to the covariate that is correctly specified in the model, and independent from the covariate that requires adjustment, we expect no bias correction from the fully weighted estimator, and better efficiency from the unweighted estimator.

We keep the same parameter initialization and simulation setting. The simulation consists of 200 samples, 10,000 iterations within each sample, including 2,000 burn-in. The bias, relative RMSE and true coverage of the nominal 95% confidence interval or credible interval are reported in Table 8.

As expected, the result from the unweighted model is more efficient than the fully weighted model, leading to a 30% to 70% reduction in RMSE. Since weighting is unnecessary for either X_1 or X_2 according to the population setup, the weight-smoothing model is able to limit the side effect of weighting, and achieves on average a 30% reduction in RMSE compared to the weighted model. Also note that the weight-smoothing model result suffers a moderate drop in the true coverage of a nominal 95% credible interval when the population variance is small relative to the misspecification.

For the last scenario, we study the model behavior when X_1 and X_2 are correlated. For this purpose, we create X_1 by the former approach, that is, from a uniform distribution on interval between zero and ten, but define X_2 as having a uniform distribution centered at X_1 to yield a correlation of about 0.45 between X_1 and X_2. The rest of the settings stay the same:

$$Y_i | X_{1i}, X_{2i}, \beta, \sigma^2 \sim N\left(\beta_0 + x_{2i} + \sum_{h=1}^{10} \beta_h (x_{1i} - h)_+, \sigma^2\right)$$

$$X_{1i} \sim UNI(0, 10), i = 1, \ldots, N = 20{,}000$$

$$X_{2i} \sim UNI(0, 10) + X_{1i}, i = 1, \ldots, N = 20{,}000$$

$$\beta = (0, 0, 0.5, 0.5, 0.5, 0.5, 1, 1, 1, 2)$$

Similarly, three settings of variance are considered, and the inclusion probabilities are related to X_1 according to the formula $\pi_i = (1 + \lceil X_{1i} \rceil / 30) * \lceil X_{1i} \rceil / 2$. The simulation again consists of 200 samples and 10,000 iterations within each sample, including 2,000 burn-in. The results are presented in Table 9.

Although X_1 and X_2 are correlated in this scenario, the results are very close to the first scenario. The weighted method is useful when model misspecification exists and the population variance is small, but loses to the unweighted method due to lack of efficiency when the model is correctly specified, or the variance is large compared to potential biasedness. The hierarchical weight-smoothing model fails to balance the two situations, closely resembling the fully weighted method.

Combining all three scenarios, we conclude that the weight-smoothing model with Laplace prior has large gains in efficiencies when weighting is not necessary for any of the covariates, at the cost of a moderate drop in the true coverage rate when the residual variance is small. In other settings, its performance is similar to the fully weighted estimator.

Table 8. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the multivariate linear regression coefficients B_1, B_2 when weight is related to X_2 using populations with different residual variances corresponding to various models under consideration

	$\sigma^2 = 10$			$\sigma^2 = 10^3$			$\sigma^2 = 10^5$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
B_1									
UNWT	−0.009	0.478	0.95	0.005	0.286	1	0.205	0.420	1
FWT	0.018	1	0.86	0.096	1	0.92	1.856	1	0.92
HWS	0.021	0.912	0.77	0.067	0.565	0.96	1.741	0.794	0.95
B_2									
UNWT	0.022	0.533	0.98	−0.254	0.629	0.97	0.833	0.748	0.95
FWT	0.044	1	0.95	0.045	1	0.95	−1.517	1	0.98
HWS	0.029	0.683	0.74	−0.091	0.615	0.91	1.230	0.698	0.97

Table 9. Comparisons of bias, relative root mean square error (RRMSE) and coverage of nominal 95% confidence interval/credible interval (CI) of various estimators for the multivariate linear regression coefficients B_1, B_2 when weight is related to X_1 and X_1 and X_2 are correlated, using populations with different residual variances corresponding to various models under consideration

	$\sigma^2 = 10$			$\sigma^2 = 10^3$			$\sigma^2 = 10^5$		
Estimator	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage	Bias	RRMSE	95% CI coverage
B_1									
UNWT	0.942	6.934	0	1.005	1.024	0.57	0.749	0.527	0.94
FWT	−0.071	1	0.96	0.067	1	0.96	−0.638	1	0.88
HWS	−0.099	1.118	0.85	0.033	0.999	0.95	−0.756	0.995	0.96
B_2									
UNWT	0.015	0.393	0.93	0.033	0.499	1	0.982	0.533	1
FWT	0.006	1	0.95	0.031	1	0.95	1.695	1	0.97
HWS	0.023	1.016	0.77	0.050	1.005	0.92	1.478	0.966	0.99

4. Application

4.1. Application on Dioxin data from NHANES

To demonstrate the performance of our method in the linear regression setting, we consider its application on the dioxin dataset from the National Health and Nutrition Examination Survey (NHANES). During the 2003–2004 survey, 1,250 representative adult subjects were selected in a probability sample of the US, and had their blood biomarkers measured, including 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), a compound usually formed through incomplete combustion such as incineration, paper and plastics manufacturing, and smoking. Other demographic variables including age and gender are available from the survey. The sampled data are stratified into 25 strata, with each consisting of two Masked Variance Units (MVU's) to account for geographic clustering in the sample design without compromising confidentiality. Survey weights are provided as well. Due to technical limits, 674 readings are below limit of detection, and are imputed through multiple imputation using the model described in Chen et al. (2010), resulting in five replicate data sets. Both survey structure and imputation are incorporated in analysis using a jackknife method and Rubin's formula (Rubin 1987).

To determine the connection between log of TCDD level and individual demographic information, four linear regression models are fitted as log TCDD on age, log TCDD on gender, log TCDD on age and gender, and log TCDD on age, gender, and interaction. The hierarchical model is built as described before, with same initial value of parameters as those in the simulation. For each model setting, the unweighted (UNWT), fully weighted (FWT), and the hierarchical weight-smoothing (HWS) estimators are obtained (the exchangeable random-slope model fails to converge and is removed from the result). To estimate mean square error, the fully weighted version is treated as unbiased. Note that the fully weighted estimator is unbiased only in expectation, leading to an unbiased estimated square bias of regression coefficient $\hat{\beta}$ given by $max(0,(\hat{\beta}-\hat{\beta}_w)^2-\hat{V}_{01})$, where $\hat{V}_{01}=\widehat{Var}(\hat{\beta})+\widehat{Var}(\hat{\beta}_w)-2\widehat{Cov}(\hat{\beta},\hat{\beta}_w)$ (Little et al. 1997). To fully account for the design features, all variance/covariance estimates are calculated via jackknife as $\widehat{Var}(\hat{\beta}_w)=\sum_h\frac{k_h-1}{k_h}\sum_{i=1}^{k_h}\left(\hat{\beta}_{w_{(hi)}}-\hat{\beta}_w\right)^2$, $\hat{\beta}_{w_{(hi)}}=(X'W_{(hi)}X)^{-1}XW_{(hi)}y$, where $\hat{\beta}_{w_{(hi)}}$ denotes the weighted β estimator from sample excluding i_{th} MVU in h_{th} stratum, and $W_{(hi)}$ is a diagonal matrix consisting of case weight w_j for all elements $j\notin h,j\notin i$, $\frac{k_h}{k_h-1}w_j$ for all elements $j\in h,j\notin i$, and 0 for elements $j\in h,j\in i$. $\widehat{Var}(\hat{\beta})$ and $\widehat{Cov}(\hat{\beta}_w,\hat{\beta})$ are calculated accordingly, and estimates from five imputed replicate datasets are combined using Rubin's formula (Rubin 1987). All Gibbs Sampler estimates are based on 10,000 iterations after discarding 2,000 draws as burn-in. The resulting bias and RMSE estimates are summarized in Tables 10 through 13.

For the first two models of log TCDD on age and gender separately, the estimation of the single predictor using an unweighted model appears to be biased compared to the fully weighted model, resulting in an estimated bias of about 40% and 70% of RMSE. However, the weighted model also fails to provide an efficient estimate for effect on age, supported by a RMSE of 3.888, larger than the RMSE of 3.265 from the unweighted model. Meanwhile, the hierarchical weight-smoothing model shows its ability to improve efficiency, both reducing the bias comparing to the unweighted model, and maintaining a

Table 10. Regression of log TCDD on Age. Bias and RMSE for linear slope estimated for age: unweighted, fully weighted and hierarchical weight smoothing

Estimator	$\text{Est.}_{95\%\text{CI}}$	Bias(10^{-3})	RMSE(10^{-3})
UNWT	$.0331_{.0284,.0378}$	-1.262	3.265
FWT	$.0343_{.0266,.0420}$	0	3.888
HWS	$.0343_{.0319,.0367}$	-0.086	1.214

RMSE similar to or smaller than the fully weighted model depending on the severity of variance inflation.

As more predictors enter the model, the estimated bias rapidly decreases in scale, leading to a scenario in which both bias and inflation in variance could dominate the overall RMSE, and neither the unweighted model nor the fully weighted model prevails in estimating all predictors. Hence the hierarchical weight-smoothing model cannot reduce bias further, yet it succeeds in reducing variance, resulting in overall smaller RMSE comparing to both the unweighted and fully weighted estimator (although the narrowness of the interval suggests that its coverage may be compromised to some degree, as in Subsection 3.4 of the simulation study).

4.2. *Application on Partners for Child Passenger Safety Data*

In this section, we use a Partners for Child Passenger Safety dataset to demonstrate our method's performance under a logistic regression setting. Unit observations in the dataset are damaged vehicles disproportionally sampled from State Farm Insurance claims records between December 1998 and December 2005, when at least one child occupant less than 16 years of age was a passenger in a model year 1990 or newer State Farm-insured vehicle with a damage claims report. The focus of the study is children's consequential injuries, defined by either facial lacerations or other injuries rated two or more on the Abbreviated Injury Scale (AIS) (Association for the Advancement of Automotive Medicine 1990). Due to the rare occurrence of the injury among all claims, to improve accuracy of the corresponding estimation of this rare outcome, the overall population was divided into three strata based on injury status – vehicles with at least one child occupant screened positive for injury at the time of the crash, vehicles with all child occupants reported receiving medical treatment but screened negative for injury, and vehicles with no occupants receiving medical treatment – and crossed with two strata defined by whether the vehicle was driveable or not. Since the stratification was associated

Table 11. Regression of log TCDD on Gender. Bias and RMSE for linear slope estimated for gender: unweighted, fully weighted and hierarchical weight smoothing

Estimator	$\text{Est.}_{95\%\text{CI}}$	Bias(10^{-2})	RMSE(10^{-1})
UNWT	$.154_{.002,.306}$	-8.219	1.248
FWT	$.236_{.110,.362}$	0	0.637
HWS	$.242_{.122,.362}$	0.589	0.607

Table 12. Regression of log TCDD on age and gender. Bias and RMSE for linear slope estimated for age and gender: unweighted, fully weighted and hierarchical weight smoothing

	Age			Gender		
Estimator	Est.$_{95\%\text{CI}}$	Bias(10^{-4})	RMSE (10^{-3})	Est.$_{95\%\text{CI}}$	Bias(10^{-2})	RMSE (10^{-2})
UNWT	$.0336_{.0287,.0385}$	-9.067	3.296	$.256_{.108,.404}$	-0.159	9.017
FWT	$.0345_{.0268,.0422}$	0	3.895	$.254_{.132,.377}$	0	6.161
HWS	$.0344_{.0320,.0368}$	-0.841	1.227	$.265_{.153,.377}$	1.058	5.659

with risk of injury (ascertained by follow-up survey), and cannot be fully explained by other auxiliary variables, the sampling design is informative, with weights varying from one to 50, and nine percent of weights lying outside three times their standard deviation.

As determined by Winston et al. (2002), children rear seated in compacted extended cab pickups are at greater risk of consequential injuries than children rear seated in other vehicles. To strengthen the conclusion, two models are applied, the unadjusted logistic model of injury status on car type (compacted extended cab pickups or others), and adjusted logistic model adapting control variables including child age (years), use of restraint (Y/N), intrusion into the passenger cabin in accident (Y/N), tow away after accident (Y/N), direction of impact (front/side/rear/other), and weight of the vehicle (pounds). The logistic hierarchical weight-smoothing model is set up as stated in the previous section, then the Gibbs sampler is executed for 10,000 iterations with 2,000 burn-in, and odds ratios are compared with the unweighted and fully weighted model. As this is a disproportionally stratifed sample design, standard variance estimators are used for the unweighted and fully weighted estimators, while the posterior predictive distribution of the HWT model is used to compute point estimates and 95% credible intervals for the HWT estimator.

The estimated odds ratios for compacted extended cab pickups indicator did not vary much from the unadjusted model to the fully adjusted model, while unweighted regression and fully weighted regression lead to quite different results, from an OR of 3.534 to 11.317 for the unadjusted model, and from 3.448 to 13.890 when all other control variables are included (see Table 14). The hierarchical weight-smoothing model estimates lie in between the unweighted and weighted estimates, although much closer to the fully weighted model. It is also worth noting that with similar point estimates, the HWS model provides a considerable reduction in estimated standard deviation, leading to a narrower 95% confidence interval compared to the fully weighted model, a characteristic also presented in the previous simulation study.

5. Discussion

Model-based approaches to "trimming" survey weights attempt to formally balance bias and variance, resulting in an estimate usually lying between those from the unweighted model and fully weighted model. The weight-smoothing model using a Laplace prior shows the potential to provide a more efficient estimate than either the unweighted model

Table 13. Regression of log TCDD on age and gender, and interaction between age and gender. Bias and RMSE for linear slope estimated for age, gender and interaction: unweighted, fully weighted and hierarchical weight smoothing

	Age			Gender			Interaction		
Estimator	Est.$_{95\%CI}$	Bias(10^{-4})	RMSE(10^{-3})	Est.$_{95\%CI}$	Bias(10^{-2})	RMSE(10^{-1})	Est.$_{95\%CI}$	Bias(10^{-3})	RMSE(10^{-3})
UNWT	$.0284_{.0227,.0341}$	−5.063	3.758	$-.195_{-.508,.118}$	2.882	1.591	$.0096_{.0031,.0161}$	−0.880	3.285
FWT	$.0289_{.0236,.0342}$	0	2.661	$-.223_{-.865,.419}$	0	3.259	$.0105_{-.0040,.0250}$	0	7.335
HWS	$.0280_{.0241,.0319}$	−9.142	2.048	$-.289_{-.541,\ -\ .037}$	−6.530	1.282	$.0122_{.0069,.0174}$	1.646	2.667

Table 14. Odds ratio and relevant 95% confidence interval for estimated effect on injury from compacted extended cab pickups: unweighted, fully weighted, hierarchical weight smoothing, and exchangeable random effect

	Odds Ratio (OR)	
Estimator	Unadjusted	Adjusted
UNWT	$3.534_{(2.003,6.234)}$	$3.448_{(1.850,6.430)}$
FWT	$11.317_{(2.737,46.784)}$	$13.890_{(3.176,60.760)}$
HWS	$10.559_{(3.731,29.876)}$	$13.268_{(7.919,22.232)}$

or the fully weighted model, using an approach that is nearly as simple to implement in a regression setting as an exchangeable model, with equivalent or improved increases in efficiency but better robustness properties. Large increases in efficiency occur when bias is present due to model misspecification, and population variance is small so the weight-smoothing model is able to model the underlying data structure precisely, yielding an estimate with greatly reduced mean square error. However, this aggressive estimation comes at some cost of robustness, that is, the reduced variance could lead to lower than nominal coverage rates. As presented in the simulation, the HWS model suffers a moderate drop in the coverage rate when population variance is small, although it is usually competitive with the coverage of the fully weighted estimator. In future, it would be worth exploring the model's mechanism in reducing the overall RMSE, and the limit of the scenarios under which it still maintains reasonable coverage.

The distribution of weights with a high degree of variability can itself vary considerably, from relatively uniform or heavy-tailed distributions, as we have seen in the simulations or the NHANES examples, to a small number of extreme outlying weights, to intermediate cases, as was the case in the Partners for Passenger Safety example. Since the method works by smoothing a given weight-stratum estimate inversely to its stability relative to other weight-stratum estimates, we would anticipate that there would not be consistent differences between these different types of weight distributions. If either the heavy-tailed or outlier weight strata are well estimated and sufficiently different from other weight strata, this interaction will be preserved and the fully weighted estimator will be approximated; otherwise the interaction will be shrunk and the estimator will move away from the fully weighted estimator.

We also note that the Laplace weight-smoothing model is largely agnostic to the construction of the weights. Instead, it focuses on whether there are enough data to support main effects and interactions between the weights and the parameter of interest (note that the main effects themselves can be viewed as interactions with the intercept in the case of estimating a population mean). While simple to implement, there may be settings where one wants to smooth some components of the weights (e.g., selection and nonresponse) while retaining others (e.g., post stratification or calibration). Such "partial smoothing" model-based approaches remain a topic for future research.

Comparing the results of the Laplace prior weight-smoothing models with the model-based estimators of Beaumont (2008), we find that the Laplace estimators offer the promise of relatively simple estimators that can approximate fully weighted estimators when weights are required for bias correction, but improve over weighted estimators in terms of variability while maintaining an approximately correct nominal coverage of credible intervals. In contrast, in some settings the Beaumont estimators can "over smooth" weights when bias correction is needed and yield unstable estimators when the weight prediction is weak. The predicted weights in the weight-modeling approach of Beaumont incorporate information from design variables, thus yielding better predictions for weighted mean and population total estimates than unweighted estimators. However, in some settings even a degree-five polynomial may fail to correctly approximate the relationship between the inverse of the probability of selection and the sample statistic of interest. Perhaps even more importantly, highly structured models for weight prediction such as high-degree polynomials may result in unstable estimates of weights, adding unnecessary variance rather than dampening it, although model-selection methods may reduce such impacts. Methods such as those proposed by Pfeffermann (2011) and Kim and Skinner (2013), who proposed a form of "stabilized" weight models as $\tilde{w}_i = E_M(w_i|I, Y, X)/E_M(w_i|I, Y)$, may be of use in informative sampling settings (in noninformative sampling settings, $\tilde{w}_i = w_i$ if the model is correctly specified). Ultimately we find attempts to model weights rather than data misguided, as this focuses on design factors on which we should be conditioning, rather than assessing uncertainties in the data that may be fertile ground for mean square error reduction while preserving approximate nominal coverage: that is, calibrated Bayes estimators (Little 2011).

As a final note, there is an issue of whether a census estimate of a parameter of a misspecified model is a sensible inferential target. Our perspective is that statistical models are rarely perfect, and that complex sample designs can sometimes magnify the degree of these failures. We recognize, however, that there is controversy in this area. For example, Rothman et al. (2013) make a case that truly scientific endeavors attempt to make causal statements that should be independent of sample selection. Keiding and Louis (2015) replied to this with an argument perhaps close to the one we make here, which is that the transportability (in the formal sense of Pearl and Bareinboim 2014) of model results may still require attention to the effects of sample design.

6. References

Alexander, C.H., S. Dahl, and L. Weidman. 1997. "Making Estimates from the American Community Survey." Paper presented at the 1997 Joint Statistical Meetings, August 10–14, 1997, Anaheim, CA. Available at: https://www.census.gov/content/dam/Census/library/working-papers/1997/acs/1997_Alexander_01.pdf (accessed March 2016).

Association for the Advancement of Automotive Medicine. 1990. *The Abbreviated Injury Scale*, 1990 Revision. Des Plaines, IL: Association for the Advancement of Automotive Medicine.

Beaumont, J.P. 2008. "A New Approach to Weighting and Inference in Sample Surveys." *Biometrika* 95: 539–553. Doi: http://dx.doi.org/10.1093/biomet/asn028.

Binder, D.A. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51: 279–292. Doi: http://dx.doi.org/10.2307/1402588.

Chen, Q., D.H. Garabrant, E. Hedgeman, R.J.A. Little, M.R. Elliott, B. Gillespie, B. Hong, S.Y. Lee, J.M. Lepkowski, A. Franzblau, P. Adriaens, A.H. Demond, and D.G. Patterson. 2010. "Estimation of Background Serum 2,3,7,8-TCDD Concentrations Using Quantile Regression in the UMDES and NHANES Populations." *Epidemiology* 21: S51–S57. Doi: http://dx.doi.org/10.1097/EDE.0b013e3181ce9550.

Chowdhury, S., M. Khare, and K. Wolter. 2007. "Weight Trimming in the National Immunization Survey." In Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association, July 29-August 2, 2007, 2651–2658. Available at: http://www.amstat.org/sections/SRMS/Proceedings/y2007/Files/JSM2007-000077.pdf (accessed March 2016).

Cox, B.G. and D.S. McGrath. 1981. "An Examination of the Effect of Sample Weight Truncation on the Mean Square Error of survey Estimates." Paper presented at the 1981 Biometric Society ENAR meeting, Richmond, VA.

Elliott, M.R. 2007. "Bayesian Weight Trimming for Generalized Linear Regression Models." *Survey Methodology* 33: 23–34.

Elliott, M.R. and R.J.A. Little. 2000. "Model-Based Approaches to Weight Trimming." *Journal of Official Statistics* 16: 191–210.

Eltoft, T., T. Kim, and T.W. Lee. 2006. "On the Multivariate Laplace Distribution." *Signal Processing Letters, IEEE* 13: 300–303. Doi: http://dx.doi.org/10.1109/LSP.2006.870353.

Ericson, W.A. 1969. "Subjective Bayesian Modeling in Sampling Finite Populations." *Journal of the Royal Statistical Society Series B* 31: 195–234. Available at: http://www.jstor.org/stable/2984206.

Ghosh, M. and G. Meeden. 1986. "Empirical Bayes Estimation of Means from Stratified Samples." *Journal of the American Statistical Association* 81: 1058–1062.

Henry, K. and R.V. Valliant. 2012. "Methods for Adjusting Survey Weights when Estimating a Total." In Proceedings of the 2012 Federal Committee on Statistical Methodology Research Conference. Available at: http://fcsm.sites.usa.gov/files/2014/05/Henry_2012FCSM_V-A.pdf (accessed March 2016).

Holt, D. and T.M.F. Smith. 1979. "Post Stratification." *Journal of the Royal Statistical Society Series A* 142: 33–46. Doi: http://dx.doi.org/10.2307/2344652.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association* 47: 663–685. Doi: http://dx.doi.org/10.1080/01621459.1952.10483446.

Keiding, N. and T.A. Louis. 2015. "Perils and Potentials of Self-Selected Entry to Epidemiological Studies and Surveys." *Journal of the Royal Statistical Society, Series A* 179: 319–376. Doi: http://dx.doi.org/10.1111/rssa.12136.

Kim, J.K. and C.J. Skinner. 2013. "Weighting in Survey Analysis under Informative Sampling." *Biometrika* 100: 385–398. Doi: http://dx.doi.org/10.1093/biomet/ass085.

Kish, L. 1992. "Weighting for Unequal Pi." *Journal of Official Statistics* 8: 183–200.

Lazzeroni, L.C. and R.J.A. Little. 1998. "Random-Effects Models for Smoothing Post-Stratification Weights." *Journal of Official Statistics* 14: 61–78.

Little, R.J.A. 1991. "Inference with Survey Weights." *Journal of Official Statistics* 7: 405–424.

Little, R.J.A. 1993. "Poststratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88: 1001–1012. Doi: http://dx.doi.org/10.1080/01621459.1993.10476368.

Little, R.J.A. 2011. "Calibrated Bayes, for Statistics in General, and Missing Data in Particular." *Statistical Science* 26: 162–174. Doi: http://dx.doi.org/10.1214/10-STS318.

Little, R.J.A., S. Lewitzky, S. Heeringa, J. Lepkowski, and R.C. Kessler. 1997. "Assessment of Weighting Methodology for the National Comorbidity Survey." *American Journal of Epidemiology* 146: 439–449.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. Boca Raton, FL: CRC Press.

Park, T. and G. Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103: 681–686. Doi: http://dx.doi.org/10.1198/016214508000000337.

Pearl, J. and E. Bareinboim. 2014. "External Validity: from Do-calculus to Transportability Across Populations." *Statistical Science* 29: 579–595. Doi: http://dx.doi.org/10.1214/14-STS486.

Pfeffermann, D. 1993. "The Role of Sampling Weights when Modeling Survey Data." *International Statistical Review* 61: 317–337. Doi: http://dx.doi.org/10.2307/1403631.

Pfeffermann, D. 2011. "Modelling of Complex Survey Data: Why Model? Why is it a Problem? How Can We Approach It." *Survey Methodology* 37: 115–136.

Potter, F.A. 1988. "Survey of Procedures to Control Extreme Sampling Weights." In Proceedings of the Section on Survey Research Methods, American Statistical Association. 453–458. Available at: http://www.amstat.org/sections/srms/Proceedings/papers/1988_083.pdf (accessed March 2016).

Potter, F. 1990. "A Study of Procedures to Identify and Trim Extreme Sample Weights." In Proceedings of the Survey Research Methods Section, American Statistical Association. 225–230. Available at: http://www.amstat.org/sections/srms/Proceedings/papers/1990_034.pdf (accessed March 2016).

Rothman, K.J., E.E. Gallacher, and E.E. Hatch. 2013. "Why Representativeness Should be Avoided." *International Journal of Epidemiology* 42: 1012–1014. Doi: http://dx.doi.org/10.1093/ije/dys223.

Rubin, D.B. 1987. *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.

Skinner, C.J., D. Holt, and T.M.F. Smith. 1989. *Analysis of Complex Surveys*. New York: Wiley.

Winston, F.K., M.K. Kallan, M.R. Elliott, R.A. Menon, and D.R. Durbin. 2002. "Risk of Injury to Child Passengers in Compact Extended Pickup Trucks." *Journal of the American Medical Association* 287: 1147–1152. Doi: http://dx.doi.org/10.1001/jama.287.9.1147.

Received June 2014
Revised September 2015
Accepted November 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 541–544, http://dx.doi.org/10.1515/JOS-2016-0027

Book Review

Polly A. Phipps[1] *and Daniell Toth*[1]

Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D.K. *Designing and Conducting Business Surveys*. 2013. Hoboken, NJ: John Wiley and Sons. ISBN: 978-0-470-90304-9, 640 pp., £58.95.

The six-hundred plus page book *Designing and Conducting Business Surveys* is an ambitious effort between the four listed authors (and an additional four contributing authors) with vast experience and diverse expertise to provide a complete guide on the entire process of administering business surveys. The resulting text should prove to be a valuable reference for both researchers and practitioners of business survey theory and methods. Covering the entire process, from the question and sample design to estimation and dissemination of results, the book contains a discussion of most survey methods and sample theory necessary for conducting surveys, focusing on the issues especially relevant or unique to business surveys.

Given the multiple authors, there are different styles throughout, but each step of the complex process of conducting a business survey is covered by an author with the appropriate experience. The level of detail is greater in some chapters than in others, however, a starting place for additional research and information is provided for all topics, including an extensive set of references. For these reasons, the volume seems most appropriate as a comprehensive reference book on business surveys.

Chapter 1 provides an overview of the book, briefly introducing definitions, frameworks, and concepts that are important in understanding business statistics and surveys. Information is provided on business data and its key association with economic performance and policy making, as well as the statistical and survey production processes that underlie economic statistics. The authors point out unique features of business surveys beginning a discussion that continues throughout the book on the business survey characteristics that drive survey practice. Given the number and range of topics discussed, this chapter can be difficult to follow; however, the authors expand upon the topics in subsequent chapters.

Chapter 2 sets out the attributes and behaviors of businesses as organizations and economic actors, with a key focus on implications for survey response. One of the strengths of this chapter is the discussion drawn from the rich literature on the business survey response process model. This includes steps in the decision to participate in and complete a survey that are linked with the business environment, organizational and social characteristics associated with the business and the respondent, as well as respondent

[1] U.S. Bureau of Labor Statistics – Office of Survey Methods Research 2, Massachusetts Ave., NE Ste. 1950 Washington District of Columbia 20212, U.S.A. Email: phipps.polly@bls.gov and toth.daniell@bls.gov

cognitive processes. Readers with an interest in administrative science as well as survey methodology will be well served by the detail in this chapter.

Chapter 3 introduces a quality framework with multiple dimensions that has been developed and refined over time to assess the "fitness of use" of statistical products for clients and users; the framework has been adopted by many national statistical institutes and statistical governing bodies. The author also expands upon the total survey error (TSE) framework mentioned in Chapter 1, a framework originating from the statistical literature that corresponds to the "accuracy" dimension of survey quality in "fitness for use" models. The TSE framework provides the structure for the rest of the chapter, with a detailed discussion of sampling and measurement components of survey error tied to the unique features and complexity of business surveys.

In Chapter 4, the authors start the "how to" conversation that is continued in the next eight chapters of the book. They include project management basics with references and the planning for steps of a survey, from the specification of survey objectives, to designing, testing and building data collection and processing systems, through data dissemination. The authors provide a strong discussion of the planning process and how it integrates into existing processes in organizations that already have a survey infrastructure and process in place. Another strength of the chapter is the conversation about risk management and how to plan what can go wrong in the survey process and how to moderate it.

Chapter 5, the most mathematical of all the chapters, provides an overview of sampling theory with emphasis on techniques and challenges most commonly associated with sampling businesses (a very skewed population). The chapter goes into substantial detail on the construction and maintenance of a business frame or registry from which to draw the sample and the concomitant challenges. The chapter then discusses a number of sample plans (sample designs with their associated estimators) commonly used in business or establishment surveys. Fittingly, considerable attention is given to designing stratified samples and appropriate estimators. The chapter also touches on more complex estimators that use auxiliary information, including model-assisted estimators and estimators for small areas, but the treatment is necessarily brief, giving just an idea of the issues involved as well as pointing the reader to external references for more details.

The authors focus on respondent burden in Chapter 6, beginning with an interesting discussion on the costs and benefits of surveys from both a business and political perspective. The chapter provides considerable detail on how to measure and calculate actual and perceived burden, a research topic with numerous studies conducted across Europe, due to national policies and European Union goals to reduce the administrative reporting of businesses. Strategies to reduce survey burden are described, covering possible reductions tied to survey sampling and coordination, communication with respondents, and questionnaire improvement. The chapter ends on an important note, discussing the need for further research evaluating specific burden measures and measurements to better understand the effectiveness of burden reduction interventions.

Chapter 7 focuses on data collection instrument content and testing, with the interest of assuring quality responses while controlling or reducing burden. This is an area with a considerable literature to draw from. The chapter highlights the uniqueness of business respondents and their use of records to complete surveys. This, the author notes, makes it difficult to observe how respondents process and complete survey instruments in real time.

Developing content, measurements, and questions for a survey can be a particularly challenging task, and the author provides a strong focus on theory, and moving from concepts to operationalization. Methods and best practices for testing and evaluating questions, instructions, and instruments are well covered and summarized for different stages in the data collection process, including a section on web data collection that readers will find very useful.

Chapter 8 pulls together an enormous amount of information on questionnaire design, referred to as "questionnaire communication." The first part of the chapter concentrates on the wording, structure, and visual design of self-administered questionnaires and associated quality issues. The author weaves together recommendations from past and current research and practice to provide comprehensive instructions on designing a business questionnaire. Examples focus on visual principles with discussions of cues to attract a person's attention in tandem with integrating consistency into questionnaire features. Another section of the chapter covers how to build a coherent message to potential respondents through introductory materials, the design of web portals, and questionnaire layout for surveys that require multiple respondents within an organization. Final sections of this ambitious chapter include designing the questionnaire to communicate intent, definitions and tasks, and minimize response burden.

In Chapter 9, the authors extend "communication" to the data collection process. This chapter is full of new theory and practice on how to build a strong communication strategy to assure the questionnaire gets into the correct hands and quality responses are received back from businesses, and all the steps in between. First, the authors outline communication objectives for prefielding, fielding, and postfielding data collection phases, providing process models with activities and actions linked to the survey organizations and businesses. Then the authors turn their focus to tailoring communication, taking into account both external and internal businesses characteristics (e.g., economic sector/size and organizational policies, respectively). The authors devote a final section to planning and testing a comprehensive communications strategy, as well as detailing guidelines and practices, with the aim of increasing survey participation.

Chapter 10 is no exception to the rule – the authors are very diligent about continuity and integration in the book—so in this chapter and others they summarize and build on earlier chapters, setting an effective stage for new material. This chapter moves to implementation and active data collection management in the field and associated paradata and quality indicator measurement and monitoring, using the TSE framework. The authors incorporate responsive design planning in their discussion, i.e., designing data collection in such a way that it can be changed in real time. Since responsive design is a promising but relatively untested method, particularly for business surveys, this is an area very much in need of research. The authors set out good principles and very specific and all-inclusive measures/metrics relevant to business surveys to monitor quality and the production process of the survey organization.

Chapter 11 provides an overview of the iterative process of data editing and cleaning, covering data capture, coding and cleaning, both during and after survey collection. The section on data capture includes a major focus on specific data collection modes, and the coding section covers major business classification systems. For all processes, the authors set out detail on minimizing, measuring, and monitoring errors, and document implications for instrument design. For editing and imputation, types, identification and treatment of micro errors are delineated (e.g., missing values, systematic, random and influential errors), as well as errors at the macroediting stage.

Chapter 12 discusses the process of making inferences from the collected data and how to balance the need to protect respondents' proprietary data with the desire to make as much information as possible available to the public. The author provides an excellent overview, particularly given the scope and complexity of business surveys, covering analytical techniques, disclosure control methods, dissemination, and archiving for economic statistical products.

Given this is a lengthy book, we hesitate to make suggestions for more detail and additional topics, but there are a few areas that are candidates for further development, when the authors begin a second edition (!). The book primarily focuses on surveys that produce official economic statistics; we think that the book could benefit from more detail and examples on business surveys that cover other topics, for example, workplace conditions, health and education, energy, and agriculture, as well as non-governmental surveys. Self-administered questionnaires are commonly used in business surveys, but we are aware of many agencies and organizations that conduct a great deal of telephone interviews, and to a lesser extent, personal interviews with business respondents. We would welcome detail on these additional modes of data collection. Given the skewed nature of employment, the book correctly spends considerable time on large, complex businesses. Small businesses have different issues, and their participation is often of concern, a topic that seems ripe for greater detail.

Finally, the authors are to be commended for putting together this comprehensive book and crediting the international workshops and conferences that contributed to building the networks and inspiration to undertake this major endeavor. As they note, the business survey methodology literature is dispersed across disciplines, and in our experience, is often found in workshop and conference papers, and to a lesser extent, in books, and peer reviewed journals. To have synthesized and provided new direction to this large body of literature is a major accomplishment.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 545–547, http://dx.doi.org/10.1515/JOS-2016-0028

Book Review

Paul C. Beatty[1]

Giampietro Gobo and Sergio Mauceri. *Constructing Survey Data: An Interactional Approach.* 2014. Thousand Oaks, CA: Sage. ISBN: 9781849201773, 392pp., £25.99.

A key premise of *Constructing Survey Data* is that survey research, as generally practiced today, relies on oversimplified questioning approaches and untenably rigid interactions with respondents, thereby straying far from the fundamental underpinnings that made it such a powerful tool for social inquiry. While today's data collection practices apparently offer standardized measurement, the authors argue that the benefits are illusory, and in reality gravely threaten the validity of survey data.

In one sense, Gobo and Mauceri's proposal to "rescue the survey from the surveyists" places them within a long tradition of critics of survey empiricism. Some of their criticisms are reminiscent of C. Wright Mills' (1959) opposition to "the ascendency of research teams of technicians"; their concerns about the failures of survey interviews to establish mutually understood meanings echo the critiques of Suchman and Jordan (1990); in short, few of the concerns they raise about validity are really new. Yet Gobo and Mauceri are distinct from their predecessors in several important regards. First, whereas many critics argue for alternatives to surveys as generally practiced, what Gobo and Mauceri propose is closer to a fix within the bounds of survey methods—at least, survey methods as originally envisioned by pioneers such as Paul Lazarsfeld. Second, although the most vocal survey critics generally cite external perspectives (e.g., ethnography and qualitative sociology), Gobo and Mauceri have an exceptional command of the survey field's interdisciplinary methodological literature.

Their synthesis of this literature—encompassing questionnaire design, the social and cognitive aspects of survey response, interviewing, and interviewer-respondent interaction—is very impressive, and is a key reason that this book will be of great interest to many survey statisticians. Also impressive is their integration of some classic methodological works by Lazarsfeld, Merton and others along with the literature of the last few decades, during which survey methodology has matured as a more cohesive discipline. This alone is a terrific accomplishment, a concise and informative synthesis that is highly recommended to researchers and students interested in questionnaire development and data collection methods.

However, the larger arc of the book—arguing how current survey practice fails to capitalize on what we (should) have learned, and proposing an alternative approach—is not always as successful, on both logical and practical grounds. For example, the first

[1] U.S. Census Bureau, Center for Survey Measurement, 5K011 Suitland Federal Center, 4600 Silver Hill Road, Washington District of Columbia 20233, U.S.A. Email: paul.c.beatty@census.gov

chapters of the book trace the evolution of the survey from its earliest form as a mixed quantitative/qualitative tool for social investigation into a more recognizable entity using probability samples and standardized data collection. In Gabo and Mauceri's account, researchers were seduced by simpler questions that promised simpler answers, such as percentages. In reality, this shift was largely driven by an increasing need for timely and objective statistical data, which was incompatible with labor intensive procedures requiring extensive investigator judgments about what to ask and who to include. Furthermore, although they characterize this as a move from *measuring* to *counting,* Gabo and Mauceri fail to acknowledge that measurement was actually a critical concern of the transitional generation of social scientists. In reality, these social scientists considered the known perils of uncontrolled variation to be more problematic than the potential—but harder to measure—loss of validity from simpler and more standardized questions (an arc described in more detail in Beatty, 1995). The resulting methodology, in their judgment, minimized the total error of statistics given realistic constraints.

Yet Gabo and Mauceri are quite correct that standardized approaches put a significant burden on survey questions and can threaten validity if not implemented with great care. This sets the stage for the most successful chapters of the book, which consider the contributions of researcher, questionnaire, interviewer, and respondent to the production of survey data. Clearly, literature on questionnaire design and the psychology of response emerging from the "CASM" (Cognitive Aspects of Survey Methodology) movement beginning in the 1980s is central to such a discussion. But whereas some excellent volumes have integrated this work and presented it within the context of larger psychological science (e.g., Tourangeau et al. 2000, Sudman et al. 1996), Gabo and Mauceri fold this literature into classic methodological work (largely sociological) on social science data collection. The results reveal a longer arc of theory and practice than many may appreciate, and are quite intellectually engaging—reclaiming the contributions of Goffman's social constructionism, Cicourel's cognitive sociology, and Cantril's classics on polling, among many others, into the science of data collection. They also yield insights into how the dynamics of survey interviewing lead to construction of meaning, with numerous practical implications (e.g., choice of open and closed questions, selection of response categories, acceptance of don't know responses). Other volumes have explored the specific topics here in greater depth, but arguably not with greater breadth. Virtually any survey practitioner will learn from what they present.

These chapters are highly informative in their own right, but Gabo and Mauceri use them to set up the final sections of the book, outlining an "interactional" alternative to current standardized data collection that is still largely recognizable as a survey. The payoff of these sections is likely to vary depending upon the reader. Social scientists will continue to find a great deal of engaging material here about the challenges of truly standardizing meaning and defining survey quality; the discussions of pretesting and deviant case analysis are as comprehensive and pragmatic as earlier sections. Gabo and Mauceri make a largely convincing case that what they propose has great promise for maximizing validity, particularly regarding the study of complex attitudes and behaviors. However, statisticians may deliberately relinquish some validity to maximize transparency and efficiency—and furthermore, may be more concerned with tracking differences over time and across groups than the true validity of a particular estimate.

In that light, statisticians may consider what Gabo and Mauceri propose to be impractical or even counterproductive. They may also ask, from a statistical perspective, how much more accurate would data actually be if the new procedures and additional costs involved were accepted? The answer, of course, is not readily forthcoming.

In spite of these caveats, Gabor and Mauceri have produced a very valuable book, one that at the very least challenges us to reconsider how well current practices align with our data needs. As methodologists themselves, their understanding of actual practices and how they evolved leads to a wealth of insights applicable to questionnaire design and data collection. Even if one does not fully accept their premise or conclusions, their constructive and insightful observations will be of great help to current practitioners, and meaningfully contribute to the discussion of where to go from here.

References

Beatty, P. 1995. "Understanding the Standardized/Non-Standardized Interviewing Controversy." *Journal of Official Statistics* 11: 147–160.

Mills, C.W. 1959. *The Sociological Imagination*. New York: Oxford University Press.

Suchman, L. and B. Jordan. 1990. "Interactional Troubles in Face-to-Face Survey Interviews." *Journal of the American Statistical Association* 85: 232–241. Doi: http://dx.doi.org/10.2307/2289550.

Sudman, S., N.M. Bradburn, and N. Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

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